Understanding Gender Discrimination by Managers

Christina Brown (University of Chicago)

July 25, 2023
NBER SI Gender

Support from DFID, JPAL, the Weiss Fund, CEGA, the Strandberg Fund, NAEd/Spencer Foundation, DERF and IRLE
Women in the Workforce

- Extensive evidence showing bias in the evaluation of women’s on-the-job productivity (hiring, wages, promotions)

Ratio of female to male labor force participation (2019)

Figure: Our World in Data; Data: ILO
Women in the Workforce

- Extensive evidence showing bias in the evaluation of women’s on-the-job productivity (hiring, wages, promotions)

- South Asia: FLFP is 1/4 of MLFP

Figure: Our World in Data; Data: ILO
Women in the Workforce

- Extensive evidence showing bias in the evaluation of women’s on-the-job productivity (hiring, wages, promotions)
- South Asia: FLFP is 1/4 of MLFP
- This paper →
  - How do employment policies affect bias?
  - What does that tell us about the underlying features of the bias?

Figure: Our World in Data; Data: ILO
Overview - Design

(1) Research questions

a. **Policy:** How do personnel policies affect the extent of gender bias in performance evaluations?

b. **Theoretical:** What are the underlying mechanisms for gender bias in performance evaluations: taste-based, statistical, other?
Overview - Design

(1) Research questions
   a. **Policy:** How do personnel policies affect the extent of gender bias in performance evaluations?
   b. **Theoretical:** What are the underlying mechanisms for gender bias in performance evaluations: taste-based, statistical, other?

(2) Design: RCT with 5,000 teachers and 200 managers in Pakistan
   a. Vary features of performance evaluation process
      - **Financial stakes:** Whether manager’s evaluation determines employee’s raise
      - **Observation:** Vary frequency of classroom observations done by managers
Overview - Design

(1) Research questions
   a. **Policy**: How do personnel policies affect the extent of gender bias in performance evaluations?
   b. **Theoretical**: What are the underlying mechanisms for gender bias in performance evaluations: taste-based, statistical, other?

(2) Design: RCT with 5,000 teachers and 200 managers in Pakistan
   a. Vary features of performance evaluation process
      - **Financial stakes**: Whether manager’s evaluation determines employee’s raise
      - **Observation**: Vary frequency of classroom observations done by managers
   b. Measure:
      - **Employee effort and productivity**: value-added, clock in times, videos of classes
      - **Performance evaluation**: Evaluation criteria and scores
      - **Manager beliefs**: Employee effort, gender bias, preferences
   c. Follow up vignette survey to test mechanisms
Overview - Results

(1) Explicit Bias: Minimal explicit bias on ability as measured by manager and teacher surveys
Overview - Results

(1) Explicit Bias: Minimal explicit bias on ability as measured by manager and teacher surveys

(2) Evaluation Scores (controlling for productivity)
Overview - Results

(1) Explicit Bias: Minimal explicit bias on ability as measured by manager and teacher surveys

(2) Evaluation Scores (controlling for productivity)
   - No financial stakes $\rightarrow$ No gender bias
Overview - Results

(1) Explicit Bias: Minimal explicit bias on ability as measured by manager and teacher surveys

(2) Evaluation Scores (controlling for productivity)

- No financial stakes $\rightarrow$ *No gender bias*
- Adding financial stakes $\rightarrow$ *Women receive 10% lower raise*
Overview - Results

(1) Explicit Bias: Minimal explicit bias on ability as measured by manager and teacher surveys

(2) Evaluation Scores (controlling for productivity)
   - No financial stakes → No gender bias
   - Adding financial stakes → Women receive 10% lower raise
   - Increasing monitoring → Reduces effect of financial stakes by 66%
Overview - Results

(1) Explicit Bias: Minimal explicit bias on ability as measured by manager and teacher surveys

(2) Evaluation Scores (controlling for productivity)
- No financial stakes → No gender bias
- Adding financial stakes → Women receive 10% lower raise
- Increasing monitoring → Reduces effect of financial stakes by 66%

(3) Discrimination Framework
Overview - Results

(1) Explicit Bias: Minimal explicit bias on ability as measured by manager and teacher surveys

(2) Evaluation Scores (controlling for productivity)
   - No financial stakes → No gender bias
   - Adding financial stakes → Women receive 10% lower raise
   - Increasing monitoring → Reduces effect of financial stakes by 66%

(3) Discrimination Framework
   - No differences in mean, variance or noisiness of productivity
Overview - Results

(1) Explicit Bias: Minimal explicit bias on ability as measured by manager and teacher surveys

(2) Evaluation Scores (controlling for productivity)
   - No financial stakes $\rightarrow$ No gender bias
   - Adding financial stakes $\rightarrow$ Women receive 10% lower raise
   - Increasing monitoring $\rightarrow$ Reduces effect of financial stakes by 66%

(3) Discrimination Framework
   - No differences in mean, variance or noisiness of productivity
   - Effort is imperfectly observable + differences in dis-utility of low raise by gender
      $\rightarrow$ Disparate outcomes
Overview - Results

(1) Explicit Bias: Minimal explicit bias on ability as measured by manager and teacher surveys

(2) Evaluation Scores (controlling for productivity)
   - No financial stakes → *No gender bias*
   - Adding financial stakes → *Women receive 10% lower raise*
   - Increasing monitoring → *Reduces effect of financial stakes by 66%*

(3) Discrimination Framework
   - No differences in mean, variance or noisiness of productivity
   - Effort is imperfectly observable + differences in dis-utility of low raise by gender
     → *Disparate outcomes*

(4) Mechanisms (source of differential dis-utility):
Overview - Results

(1) Explicit Bias: Minimal explicit bias on ability as measured by manager and teacher surveys

(2) Evaluation Scores (controlling for productivity)
   - No financial stakes → No gender bias
   - Adding financial stakes → Women receive 10% lower raise
   - Increasing monitoring → Reduces effect of financial stakes by 66%

(3) Discrimination Framework
   - No differences in mean, variance or noisiness of productivity
   - Effort is imperfectly observable + differences in dis-utility of low raise by gender
     → Disparate outcomes

(4) Mechanisms (source of differential dis-utility):
   - Differences in perceived “deservedness” (household income) ✔
Overview - Results

(1) Explicit Bias: Minimal explicit bias on ability as measured by manager and teacher surveys

(2) Evaluation Scores (controlling for productivity)
   - No financial stakes → No gender bias
   - Adding financial stakes → Women receive 10% lower raise
   - Increasing monitoring → Reduces effect of financial stakes by 66%

(3) Discrimination Framework
   - No differences in mean, variance or noisiness of productivity
   - Effort is imperfectly observable + differences in dis-utility of low raise by gender
     → Disparate outcomes

(4) Mechanisms (source of differential dis-utility):
   - Differences in perceived “deservedness” (household income) ✓
   - Differences in response (turnover, complaining) ✗
Contribution

- Interaction between bias and HR policies: Better information and less financial stakes decreases bias
  Biasi and Sarsons, 2022; Beg, Fitzpatrick and Lucas, 2021; Blau and Kahn, 2017

- Information and Discrimination: Better information about worker productivity lowers bias
  Laouénan and Rathelot, 2022; Bohren, Imas and Rosenberg, 2019; Sarsons, 2017; Bordalo et al, 2017

- Financial discrimination: Disparate employment outcomes without discrimination on productivity
Setting

Experiment

- Large private school network operating hundreds of schools across urban Pakistan
- Grades 4-13 in English, Urdu, math and science
- Managers are principal or vice principals
Setting

Experiment
- Large private school network operating hundreds of schools across urban Pakistan
- Grades 4-13 in English, Urdu, math and science
- Managers are principal or vice principals

Relevance
- **Useful for personnel econ:** Multiple, hard to measure outcomes
- **Very relevant for FLFP:** 51% of women in labor force with HS degree are teachers (8% of entire female labor force)
1). Taste-based

- Minimal stated/perceived gender bias on ability

Managers are very gender progressive

When jobs are scarce, men should have more right to a job than women

On average, teachers do not think there is bias in favor or against female teachers in evaluations.
Conceptual Framework - Context

1). Taste-based
   - Minimal stated/perceived gender bias on ability

2). Statistical
   - Same productivity (mean & sd)
Conceptual Framework - Context

1). Taste-based
   - Minimal stated/perceived gender bias on ability

2). Statistical
   - Same productivity (mean & sd)
   - Same production function noise
Conceptual Framework - Context

1). Taste-based
   - Minimal stated/perceived gender bias on ability

2). Statistical
   - Same productivity (mean & sd)
   - Same production function noise

3). This paper –
   Disparate outcomes arise from:
   - Noisy production function
   - Dis-utility to manager of giving low wages varies by gender
Conceptual Framework - Set-up

The performance evaluation system takes place in three stages

1). Production: Employees work and produce output which is a noisy function of their effort

2). Evaluation: Manager provides evaluation score
The performance evaluation system takes place in three stages

1). Production: Employees work and produce output which is a noisy function of their effort

2). Evaluation: Manager provides evaluation score

3). Post-Evaluation:
   a. Dis-utility from inaccurate scores
      - e.g. system legitimacy, psychic cost of lying, punishment
      - Decreasing in production function noisiness
Conceptual Framework - Set-up

The performance evaluation system takes place in three stages

1). Production: Employees work and produce output which is a noisy function of their effort

2). Evaluation: Manager provides evaluation score

3). Post-Evaluation:
   a. Dis-utility from inaccurate scores
      - e.g. system legitimacy, psychic cost of lying, punishment
      - *Decreasing in production function noisiness*
   b. Dis-utility from low wages
      - e.g. complaints, guilt, turnover
      - *Increasing in financial stakes of evaluation*
Conceptual Framework - Set-up

The performance evaluation system takes place in three stages

1). Production: Employees work and produce output which is a noisy function of their effort

2). Evaluation: Manager provides evaluation score

3). Post-Evaluation:
   a. Dis-utility from inaccurate scores
      - e.g. system legitimacy, psychic cost of lying, punishment
      - Decreasing in production function noisiness
   b. Dis-utility from low wages
      - e.g. complaints, guilt, turnover
      - Increasing in financial stakes of evaluation
      - Allow this to vary by worker type
Conceptual Framework - Set-up

The performance evaluation system takes place in three stages

1). Production: Employees work and produce output which is a noisy function of their effort

2). Evaluation: Manager provides evaluation score

3). Post-Evaluation:
   a. Dis-utility from inaccurate scores
      - e.g. system legitimacy, psychic cost of lying, punishment
      - Decreasing in production function noisiness
   b. Dis-utility from low wages
      - e.g. complaints, guilt, turnover
      - Increasing in financial stakes of evaluation
      - Allow this to vary by worker type

Manager chooses evaluation score given post-evaluation costs and benefits.
Conceptual Framework - Set-up

The performance evaluation system takes place in three stages

1). Production: Employees work and produce output which is a noisy function of their effort

2). Evaluation: Manager provides evaluation score

3). Post-Evaluation:
   a. Dis-utility from inaccurate scores
      - e.g. system legitimacy, psychic cost of lying, punishment
      - *Decreasing in production function noisiness*
   b. Dis-utility from low wages
      - e.g. complaints, guilt, turnover
      - *Increasing in financial stakes of evaluation*
      - Allow this to vary by worker type

Manager chooses evaluation score given post-evaluation costs and benefits.
Design - Treatments

- **All teachers**: Managers set performance evaluation criteria and rate teachers
Design - Treatments

- **All teachers:** Managers set performance evaluation criteria and rate teachers

- **Treatment 1:** Financial Stakes of Evaluation (randomized at school level)
  - **Control:** Employee’s end of year raise is determined by:
    - *Flat Raise:* Employees receive a raise of 5% of their base salary
    - *Objective Raise:* Teachers receive a raise from 0-10% based on student test scores

- **Treatment 2:** Classroom Observations (randomized at teacher level)
  - Manager told to conduct monthly, unannounced 20 minute observations for 4 months before evaluation (relative to status quo)
  - Treatment increases number of observations received by 50%
  - Baseline balance: 2 of 27 coefficients are stat. sig.
Design - Treatments

- **All teachers:** Managers set performance evaluation criteria and rate teachers

- **Treatment 1:** Financial Stakes of Evaluation (randomized at school level)
  - **Control:** Employee’s end of year raise is determined by:
    - *Flat Raise:* Employees receive a raise of 5% of their base salary
    - *Objective Raise:* Teachers receive a raise from 0-10% based on student test scores
  - **Treatment:** *Subjective Raise:* Teachers receive a raise from 0-10% based on their performance evaluation score

- **Treatment 2:** Classroom Observations (randomized at teacher level)
  - Manager told to conduct monthly, unannounced 20 minute observations for 4 months before evaluation (relative to status quo)
  - Treatment increases number of observations received by 50%
  - Baseline balance: 2 of 27 coefficients are stat. sig.

*Attrition: Administrative data available for all; 12% attrition for endline teacher survey*
Design - Treatments

- **All teachers**: Managers set performance evaluation criteria and rate teachers

- **Treatment 1**: Financial Stakes of Evaluation (randomized at school level)
  - **Control**: Employee’s end of year raise is determined by:
    - *Flat Raise*: Employees receive a raise of 5% of their base salary
    - *Objective Raise*: Teachers receive a raise from 0-10% based on student test scores
  - **Treatment**: *Subjective Raise*: Teachers receive a raise from 0-10% based on their performance evaluation score

- **Treatment 2**: Classroom Observations (randomized at teacher level)
  - Manager told to conduct monthly, unannounced 20 minute observations for 4 months before evaluation (relative to status quo)
  → *Treatment increases number of observations received by 50%*

Baseline balance: 2 of 27 coefficients are stat. sig. Table
Attrition: Administrative data available for all; 12% attrition for endline teacher survey
Design - Timeline

- Baseline survey
  - Informed about incentive schemes
  - Measure teacher characteristics

Example criteria 10/23
Design - Timeline

- Research team meets in person with managers
- School system HR does in person presentation at each school
- Email information
- Displayed on teacher’s dashboard

Managers set criteria for ALL teachers (not new to experiment)

Baseline survey

Example criteria 10/23
Design - Timeline

- **Baseline survey**
  - Oct 2018

- **Performance Eval Info (Treatment 1)**
  - Jan 2018

- **Clock in/out data**
  - Apr 2018

- **Classroom video-taping**
  - Jul 2018

- **Record 5 hours of class time for 1500 teachers**

- **Tripod in back of classroom to minimize interference**

Legend:
- **Blue** = Data collection activity
- **Yellow** = Treatment implementation
- **Green** = Both

Example criteria 10/23
Design - Timeline

- Baseline survey
- Performance Eval Info (Treatment 1)
- Clock in/out data
- Classroom video-taping
- Manager Observation (Treatment 2)
- Managers score teachers on criteria
- Managers asked to observe random set of teachers

Example criteria

- Data collection activity
- Treatment implementation
- Both
Design - Timeline

- Baseline survey
- Performance Eval Info (Treatment 1)
- Clock in/out data
- Classroom video-taping
- Manager Observation (Treatment 2)
- Teacher/Manager Survey (Endline)
- Performance Eval
- Raises set
- Student tests

Example criteria

- Data collection activity
- Treatment implementation
- Both
## Design - Data

<table>
<thead>
<tr>
<th>Type</th>
<th>N</th>
<th>Source</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teachers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Ground truth”</td>
<td>1,500</td>
<td>Class video</td>
<td>Rubric covering 20 aspects of pedagogy (Araujo et al, 2016)</td>
</tr>
<tr>
<td></td>
<td>3,600</td>
<td>Admin data</td>
<td>Value-added (From 5 years of student test scores)</td>
</tr>
<tr>
<td></td>
<td>9,100</td>
<td>Admin data</td>
<td>Daily clock in and out time</td>
</tr>
</tbody>
</table>
### Design - Data

<table>
<thead>
<tr>
<th>Type</th>
<th>N</th>
<th>Source</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Teachers</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Ground truth”</td>
<td>1,500</td>
<td>Class video</td>
<td>Rubric covering 20 aspects of pedagogy (Araujo et al, 2016)</td>
</tr>
<tr>
<td></td>
<td>3,600</td>
<td>Admin data</td>
<td>Value-added (From 5 years of student test scores)</td>
</tr>
<tr>
<td></td>
<td>9,100</td>
<td>Admin data</td>
<td>Daily clock in and out time</td>
</tr>
<tr>
<td><strong>Managers</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beliefs</td>
<td>189</td>
<td>Survey</td>
<td>Rate teachers on several criteria</td>
</tr>
<tr>
<td></td>
<td>189</td>
<td>Admin data</td>
<td>Rate teachers on several criteria (after observation)</td>
</tr>
<tr>
<td>Preferences</td>
<td>189</td>
<td>Survey</td>
<td>Vignettes (rating hypothetical teachers)</td>
</tr>
<tr>
<td></td>
<td>189</td>
<td>Survey</td>
<td>Rank importance of teacher behaviors</td>
</tr>
<tr>
<td></td>
<td>189</td>
<td>Evaluation</td>
<td>Points allocated to criteria</td>
</tr>
</tbody>
</table>
Design - Data

<table>
<thead>
<tr>
<th>Type</th>
<th>N</th>
<th>Source</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teachers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Ground truth”</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Ground truth”</td>
<td>1,500</td>
<td>Class video</td>
<td>Rubric covering 20 aspects of pedagogy (Araujo et al, 2016)</td>
</tr>
<tr>
<td></td>
<td>3,600</td>
<td>Admin data</td>
<td>Value-added (From 5 years of student test scores)</td>
</tr>
<tr>
<td></td>
<td>9,100</td>
<td>Admin data</td>
<td>Daily clock in and out time</td>
</tr>
<tr>
<td>Managers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beliefs</td>
<td>189</td>
<td>Survey</td>
<td>Rate teachers on several criteria</td>
</tr>
<tr>
<td></td>
<td>189</td>
<td>Admin data</td>
<td>Rate teachers on several criteria (after observation)</td>
</tr>
<tr>
<td>Preferences</td>
<td>189</td>
<td>Survey</td>
<td>Vignettes (rating hypothetical teachers)</td>
</tr>
<tr>
<td></td>
<td>189</td>
<td>Survey</td>
<td>Rank importance of teacher behaviors</td>
</tr>
<tr>
<td></td>
<td>189</td>
<td>Evaluation</td>
<td>Points allocated to criteria</td>
</tr>
<tr>
<td>Evaluation</td>
<td>189</td>
<td>Evaluation</td>
<td>Total score and criteria-level score</td>
</tr>
<tr>
<td>Bias</td>
<td>189</td>
<td>Survey</td>
<td>World Values Survey questions</td>
</tr>
<tr>
<td></td>
<td>189</td>
<td>Survey</td>
<td>Teacher’s rating of manager’s bias</td>
</tr>
<tr>
<td></td>
<td>189</td>
<td>Survey</td>
<td>Varying gender of name in vignette</td>
</tr>
</tbody>
</table>
### Panel A. Teacher Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Study Sample</th>
<th>US Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>St. Dev.</td>
</tr>
<tr>
<td>Age</td>
<td>35.0</td>
<td>8.9</td>
</tr>
<tr>
<td>Female</td>
<td>0.80</td>
<td>0.40</td>
</tr>
<tr>
<td>Years of experience</td>
<td>5.1</td>
<td>5.2</td>
</tr>
<tr>
<td>Has Post BA Education</td>
<td>0.68</td>
<td>0.47</td>
</tr>
<tr>
<td>Salary, USD</td>
<td>4,000</td>
<td>1,700</td>
</tr>
</tbody>
</table>

### Panel B. Manager Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Study Sample</th>
<th>US Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>St. Dev.</td>
</tr>
<tr>
<td>Age</td>
<td>44.9</td>
<td>9.2</td>
</tr>
<tr>
<td>Female</td>
<td>0.61</td>
<td>0.49</td>
</tr>
<tr>
<td>Years of experience</td>
<td>9.6</td>
<td>7.9</td>
</tr>
</tbody>
</table>

### Panel C. Manager Time Use

<table>
<thead>
<tr>
<th></th>
<th>Study Sample</th>
<th>US Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>St. Dev.</td>
</tr>
<tr>
<td>Total hours worked</td>
<td>47.2</td>
<td>16.3</td>
</tr>
</tbody>
</table>

#### Hours spent on:
- Administrative tasks
  - Study Sample: 18.5, St. Dev.: 10.3
  - US Sample: 18.2, St. Dev.: 2.3
- Teacher management and teaching
  - Study Sample: 17.5, St. Dev.: 8.2
  - US Sample: 15.1, St. Dev.: 2.0
- Student and parent interactions
  - Study Sample: 6.3, St. Dev.: 4.4
  - US Sample: 20.2, St. Dev.: 2.7
- Other tasks
  - Study Sample: 6.9, St. Dev.: 12.3
  - US Sample: 4.0, St. Dev.: 2.6

Col. 3 and 4 Source: School and Staff Survey (National Center for Education Statistics)
Treatment “First Stage”

Treatments effect teacher and manager behavior:

- Financial treatment: Teachers work harder and this effects student outcomes

![Graph showing treatment effects](image-url)
Treatment “First Stage”

Treatments effect teacher and manager behavior:

- Financial treatment: Teachers work harder and this effects student outcomes
- Observation treatment: Accuracy of managers’ beliefs about teacher effort improves
Effect of Financial Stakes on Bias

- No difference in evaluation score when no financial stakes
  Control for value-added, attendance, time use, video scores
Effect of Financial Stakes on Bias

- No difference in evaluation score when no financial stakes
  Control for value-added, attendance, time use, video scores

- 10% lower raise for women when there are financial stakes of evaluation
Effect of Observation on Bias

- On average, 12% lower evaluations for women (controlling for productivity)
Effect of Observation on Bias

**Table**

- On average, 12% lower evaluations for women (controlling for productivity)
- Gender gap disappears with better monitoring
Effect of Both Treatments on Bias

- More monitoring reduces the negative effect of financial stakes on gender bias by 2/3
Effect of Both Treatments on Bias

- More monitoring reduces the negative effect of financial stakes on gender bias by 2/3
- No stat. sig. difference of treatments by manager gender, experience and baseline bias
Mechanisms

- No detectable effect of financial or observation treatment on:
  - “Care” spent on evaluation scores (use of round numbers, variance across sub-criteria)
  - Evaluation criteria selected
  - No heterogeneous teacher effort response by gender (consistent with literature, Bandiera et al, 2021)
Mechanisms

Potential mechanisms from focus groups with teachers and managers
Mechanisms

Potential mechanisms from focus groups with teachers and managers

1). Differential response to raise by gender which creates dis-utility for the manager. 
*E.g. More likely to complain, turnover, reduce public good provision, etc.*
Mechanisms

Potential mechanisms from focus groups with teachers and managers

1). Differential **response to raise** by gender which creates dis-utility for the manager
   *E.g. More likely to complain, turnover, reduce public good provision, etc.*

2). Differential beliefs about **“deservedness”** of financial rewards by gender
   *E.g. Differences in total household income, breadwinner norms, etc.*
Mechanisms

Potential mechanisms from focus groups with teachers and managers

1). Differential response to raise by gender which creates dis-utility for the manager
   E.g. More likely to complain, turnover, reduce public good provision, etc.

2). Differential beliefs about “deservedness” of financial rewards by gender
   E.g. Differences in total household income, breadwinner norms, etc.

Test mechanisms: Vary teacher attributes and evaluation score
Mechanisms

Potential mechanisms from focus groups with teachers and managers

1). Differential **response to raise** by gender which creates dis-utility for the manager

   *E.g. More likely to complain, turnover, reduce public good provision, etc.*

2). Differential beliefs about **“deservedness”** of financial rewards by gender

   *E.g. Differences in total household income, breadwinner norms, etc.*

Test mechanisms: Vary teacher attributes and evaluation score

   → Vignette survey experiment
Mechanisms: Vignette Survey

- Separate manager sample
Mechanisms: Vignette Survey

- Separate manager sample
- Vary teacher attributes

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

- Separate manager sample
- Vary teacher attributes
- Vary evaluation features (across subject)

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”

Privacy of Decision

- Private: “The teacher will not learn who gave them this appraisal score so they will not know you made the decision”
- Public: “You would need to tell the teacher what appraisal score you gave them”
Mechanisms: Vignette Survey

- Separate manager sample
- Vary teacher attributes
- Vary evaluation features (across subject)
- Manager ranks teachers

<table>
<thead>
<tr>
<th>Category/Name</th>
<th>Ahmad</th>
<th>Zainab</th>
<th>Iqbal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top category</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle category</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bottom category</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Please select which of the following teachers should receive each appraisal category
Mechanisms: Vignette Survey

- Separate manager sample
- Vary teacher attributes
- Vary evaluation features (across subject)
- Manager ranks teachers
- Managers predict response to hypothetical scores
- Provide attributes for a teacher and a hypothetical score
- Rate whether you agree or disagree with response
  - Teacher would:
    - Complain
    - Feel happy
    - Look for another job
    - Be less willing to help with extra tasks
    - Feel disappointed
    - Volunteer for extra duties
    - Suffer financial hardship
  - I (the manager) would:
    - Feel bad for the teacher
    - Feel good about the decision
# Mechanisms: Vignette Survey

- Separate manager sample
- Vary teacher attributes
- Vary evaluation features (across subject)
- Manager ranks teachers
- Managers predict response to hypothetical scores
- Usual concerns
- Checks for inattention (response time, internal consistency)
- Consistent with World Values Survey bias responses
- Conservative test of mechanisms

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

20/23
Mechanisms: Response to Raise

- Minimal differences in predicted response
Mechanisms: Response to Raise

- Minimal differences in predicted response
- No differential turnover after low raise (in cross section)

Teacher would:

I would feel:
Mechanisms: Deservedness

- Large differences in HH income by employee gender
Mechanisms: Deservedness

- Large differences in HH income by employee gender
- Teachers rated higher when their spouse does not work under financial stakes
Mechanisms: Deservedness

- Large differences in HH income by employee gender
- Teachers rated higher when their spouse does not work under financial stakes
- Effects are larger when decision is public
Mechanisms: Deservedness

- Large differences in HH income by employee gender
- Teachers rated higher when their spouse does not work *under financial stakes*
- Effects are larger when decision is public
- Effects of financial treatment (RCT) are smaller for young teachers

---

### Graphs

#### ≥ 30 years old

- **Male**
- **Female**

#### < 30

- **Male**
- **Female**
- Women receive lower evaluation scores (controlling for productivity) only when the evaluation affects wages

- Gender bias decreases when managers have better information about employee effort
Conclusion

- Women receive lower evaluation scores (controlling for productivity) only when the evaluation affects wages
- Gender bias decreases when managers have better information about employee effort
- Suggests trade-off between manager’s desire for accuracy and dis-utility from low wages
Conclusion

- Women receive lower evaluation scores (controlling for productivity) only when the evaluation affects wages
- Gender bias decreases when managers have better information about employee effort
- Suggests trade-off between manager’s desire for accuracy and dis-utility from low wages
- Evidence that differential household income by gender contributes to effects
Conclusion

- Women receive lower evaluation scores (controlling for productivity) only when the evaluation affects wages
- Gender bias decreases when managers have better information about employee effort
- Suggests trade-off between manager’s desire for accuracy and dis-utility from low wages
- Evidence that differential household income by gender contributes to effects
- Understand whether household income could be important omitted variable
Thank you!

Contact: christinabrown@uchicago.edu
christinalbrown.com

Research Team-Centre of Economic Research in Pakistan (CERP)

Haya Mubasher
Anam Tariq
Attefaq Ahmed
Mujahid Murtaza
Maheen Rashid
Zahra Niazi
Zohaib Hassan
Valuation of Teacher Attributes

vig_partner_doesnt_work=1
Treatment: Evaluation affects teacher’s pay=1
vig_partner_doesnt_work=1 # Treatment: Evaluation affects teacher’s pay=1
Class 1 Urdu
Class 5 Math
Class 5 Urdu
average
below average
vig_classroom_env
vig_observation
10 days, more than average
7 days, about average
vig_job_status
vig_experience
_cons

-1 0 1 2
Conceptual Framework

1. **Production**: Employee $i$ produces output $y_i$, the sum of their true ability/effort, $\theta_i$, and noise, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. 

\[
y_i = \theta_i + \epsilon_i
\]
Appendix

Conceptual Framework

1. **Production:** Employee $i$ produces output $y_i$, the sum of their true ability/effort, $\theta_i$ and noise, $\epsilon_i \sim \mathcal{N}(0, \sigma^2_\epsilon)$.

   $$y_i = \theta_i + \epsilon_i$$

2. **Evaluation:** Managers observe $y_i$ but not $\theta_i$. Evaluation score is a function of output and a discretionary component, $d_i$:

   $$s_i = y_i + d_i = \theta + \epsilon + d_i$$
Conceptual Framework

1. Production: Employee $i$ produces output $y_i$, the sum of their true ability/effort, $\theta_i$ and noise, $\epsilon_i \sim \mathcal{N}(0, \sigma_{\epsilon}^2)$.

   $$y_i = \theta_i + \epsilon_i$$

2. Evaluation: Managers observe $y_i$ but not $\theta_i$. Evaluation score is a function of output and a discretionary component, $d_i$:

   $$s_i = y_i + d_i = \theta + \epsilon_i + d_i$$

3. Post-Evaluation Manager Utility
   i. Dis-utility from inaccurate scores
      (system legitimacy, psychic cost of lying, punishment)

      $$E[P_i] = p(s_i - \theta_i)^2$$
      $$= \frac{1}{\sigma_{\epsilon}^2} (s_i - \theta_i)^2$$

      $p$ selected to maximize benefit (punishment for inaccurate scores) - cost (unnecessarily punishing for noisy production functions)
**Conceptual Framework**

1. **Production:** Employee $i$ produces output $y_i$, the sum of their true ability/effort, $\theta_i$ and noise, $\epsilon_i \sim \mathcal{N}(0, \sigma^2_\epsilon)$.

   $$y_i = \theta_i + \epsilon_i$$

2. **Evaluation:** Managers observe $y_i$ but not $\theta_i$. Evaluation score is a function of output and a discretionary component, $d_i$:

   $$s_i = y_i + d_i = \theta + \epsilon_i + d_i$$

3. **Post-Evaluation Manager Utility**
   
   i. **Dis-utility from inaccurate scores**
      (system legitimacy, psychic cost of lying, punishment)
      
      $$E[P_i] = p(s_i - \theta_i)^2$$
      
      $$= \frac{1}{\sigma^2_\epsilon} (s_i - \theta_i)^2$$

      $p$ selected to maximize benefit (punishment for inaccurate scores) - cost (unnecessarily punishing for noisy production functions)

   ii. **Dis-utility of low eval scores**
      (complaints from employees, guilt, turnover)
      
      $$C_i = -c \rho_i s_i$$

      - $c$ is the unit-cost
      - $\rho_i$ is the dis-utility from a given employee, conditional on the score
Managers select the discretionary component of the salary to minimize the dis-utility from inaccuracy and low scores they expect to face in the next period:

\[ u(d_i) = \min_{d_i} E[-c \rho_i s_i + p(s_i - \theta_i)^2] \]

\[ = \min_{d_i} E[-c \rho_i (\theta_i + \epsilon_i + d_i) + \frac{1}{\sigma^2_{\epsilon}} (\epsilon_i + d_i)^2] \]

\[ \frac{\partial u_i}{\partial d_i} = E[-c \rho_i + 2 \frac{1}{\sigma^2_{\epsilon}} (\epsilon_i + d_i)] = 0 \]

\[ d_i^* = \frac{c \rho_i \sigma^2_{\epsilon}}{2} \]

Therefore an employee’s evaluation score will be \( s_i^* = y_i + \frac{c \rho_i \sigma^2_{\epsilon}}{2} \).
Conceptual Framework

Gender differences

- Same mean and variance:  $\theta_i^f, \theta_i^m \sim \mathcal{N}(\mu, \sigma^2_\theta)$
- Same noisiness:  $\epsilon_i^f, \epsilon_i^m \sim \mathcal{N}(0, \sigma^2_\epsilon)$
- Difference in dis-utility to manager from low evaluation (guilt, turnover, complaints):  $\rho_m > \rho_f$

The difference in expected scores, conditional on ability, by gender then is:

$$\frac{\partial s_i^*}{\partial \text{female}}|_{\theta_i} = (y_f - y_m)|_{\theta_i} + \frac{c\sigma^2_\epsilon}{2}(\rho_f - \rho_m)$$

$$= \frac{c\sigma^2_\epsilon}{2}(\rho_f - \rho_m) < 0$$
Appendix

Conceptual Framework

Effect of changes in:
- the magnitude of the inconvenience cost \((c)\)
- the accuracy of information managers have \((\sigma^2_\epsilon)\)

i. On Evaluation scores:
\[
\frac{\partial s^*_i}{\partial c} = \frac{\rho_i \sigma^2_\epsilon}{2} > 0 \quad \frac{\partial s^*_i}{\partial \sigma^2_\epsilon} = \frac{c \rho_i}{2} > 0
\]

ii. On Gender gap:

Prediction 1:
\[
\frac{\partial^2 s^*_i}{\partial c \partial \text{female}} \bigg|_{\theta_i} = \frac{\sigma^2_\epsilon}{2} (\rho_f - \rho_m) < 0
\]

Prediction 2:
\[
\frac{\partial^2 s^*_i}{\partial \sigma^2_\epsilon \partial \text{female}} \bigg|_{\theta_i} = \frac{c}{2} (\rho_f - \rho_m) < 0
\]

Prediction 3:
\[
\frac{\partial^3 s^*_i}{\partial c \partial \sigma^2_\epsilon \partial \text{female}} \bigg|_{\theta_i} = \frac{1}{2} (\rho_f - \rho_m) < 0
\]
## Heterogeneous Effects of Financial Treatment

### Predicted Raise Amount (USD)

<table>
<thead>
<tr>
<th></th>
<th>(1) Male</th>
<th>(2) Age</th>
<th>(3) Avg. Bias</th>
<th>(4) Math</th>
<th>(5) Jobs</th>
<th>(6) Family</th>
<th>(7) Teacher Age</th>
<th>(8) Young</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>-31.77*</td>
<td>-88.23</td>
<td>-44.40</td>
<td>-55.51</td>
<td>-38.07</td>
<td>-18.46</td>
<td>-6.121</td>
<td>-32.78*</td>
</tr>
<tr>
<td></td>
<td>(16.26)</td>
<td>(103.0)</td>
<td>(66.30)</td>
<td>(66.46)</td>
<td>(48.12)</td>
<td>(36.18)</td>
<td>(71.86)</td>
<td>(17.83)</td>
</tr>
<tr>
<td>Female</td>
<td>-84.65</td>
<td>-3.522*</td>
<td>-17.54</td>
<td>-15.80</td>
<td>-9.782</td>
<td>-4.119</td>
<td>8.982***</td>
<td>-137.7**</td>
</tr>
<tr>
<td></td>
<td>(64.42)</td>
<td>(1.967)</td>
<td>(43.21)</td>
<td>(41.70)</td>
<td>(25.62)</td>
<td>(29.20)</td>
<td>(2.061)</td>
<td>(58.94)</td>
</tr>
<tr>
<td>Interaction</td>
<td>22.93</td>
<td>-141.6</td>
<td>197.5</td>
<td>126.6</td>
<td>158.8</td>
<td>116.2</td>
<td>49.21</td>
<td>56.29</td>
</tr>
<tr>
<td></td>
<td>(39.60)</td>
<td>(198.7)</td>
<td>(149.3)</td>
<td>(125.2)</td>
<td>(109.2)</td>
<td>(120.2)</td>
<td>(100.0)</td>
<td>(48.31)</td>
</tr>
<tr>
<td>Financial Treatment</td>
<td>56.39</td>
<td>3.678</td>
<td>-76.17</td>
<td>-39.97</td>
<td>-67.75</td>
<td>-36.53</td>
<td>-0.252</td>
<td>-33.87</td>
</tr>
<tr>
<td></td>
<td>(76.74)</td>
<td>(3.890)</td>
<td>(59.53)</td>
<td>(51.85)</td>
<td>(45.28)</td>
<td>(39.77)</td>
<td>(2.914)</td>
<td>(80.30)</td>
</tr>
<tr>
<td>Interaction*Female</td>
<td>-1.904</td>
<td>-1.824</td>
<td>63.31</td>
<td>27.27</td>
<td>68.35*</td>
<td>30.47</td>
<td>-1.269</td>
<td>75.91</td>
</tr>
<tr>
<td></td>
<td>(64.59)</td>
<td>(3.277)</td>
<td>(40.08)</td>
<td>(37.85)</td>
<td>(36.45)</td>
<td>(25.17)</td>
<td>(2.825)</td>
<td>(71.88)</td>
</tr>
<tr>
<td>Financial Treatment*Female</td>
<td>47.03</td>
<td>1.355</td>
<td>8.217</td>
<td>14.63</td>
<td>6.094</td>
<td>-3.274</td>
<td>0.639</td>
<td>24.16</td>
</tr>
<tr>
<td></td>
<td>(49.76)</td>
<td>(2.082)</td>
<td>(27.58)</td>
<td>(29.44)</td>
<td>(20.89)</td>
<td>(13.30)</td>
<td>(1.941)</td>
<td>(51.83)</td>
</tr>
<tr>
<td>Interaction*Financial Treatment</td>
<td>415.7***</td>
<td>571.2***</td>
<td>444.9***</td>
<td>438.5***</td>
<td>425.6***</td>
<td>417.4***</td>
<td>66.57</td>
<td>430.8***</td>
</tr>
<tr>
<td></td>
<td>(25.15)</td>
<td>(103.2)</td>
<td>(107.1)</td>
<td>(91.62)</td>
<td>(63.23)</td>
<td>(86.96)</td>
<td>(69.55)</td>
<td>(28.35)</td>
</tr>
</tbody>
</table>

### Summary

- **Observations**: 3650
- **Clusters**: 208
- **Dep. Var. Mean**: 368.4
- **Dep. Var. SD**: 176.3
## Heterogeneous Effects of Observation Treatment

<table>
<thead>
<tr>
<th></th>
<th>(1) Male</th>
<th>(2) Age</th>
<th>(3) Avg. Bias</th>
<th>(4) Math</th>
<th>(5) Jobs</th>
<th>(6) Family</th>
<th>(7) Teacher Age</th>
<th>(8) Young</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Female</strong></td>
<td>-60.06***</td>
<td>-225.0</td>
<td>-91.40</td>
<td>-100.3*</td>
<td>-117.7**</td>
<td>-30.64</td>
<td>12.87</td>
<td>-71.06***</td>
</tr>
<tr>
<td></td>
<td>(19.47)</td>
<td>(191.0)</td>
<td>(58.34)</td>
<td>(54.16)</td>
<td>(46.15)</td>
<td>(39.70)</td>
<td>(93.04)</td>
<td>(23.10)</td>
</tr>
<tr>
<td></td>
<td>(54.62)</td>
<td>(4.113)</td>
<td>(29.40)</td>
<td>(29.00)</td>
<td>(21.10)</td>
<td>(19.39)</td>
<td>(2.500)</td>
<td>(68.29)</td>
</tr>
<tr>
<td><strong>Observation Treatment</strong></td>
<td>-12.47</td>
<td>-212.9</td>
<td>22.77</td>
<td>4.496</td>
<td>-31.04</td>
<td>30.11</td>
<td>42.34</td>
<td>-22.15</td>
</tr>
<tr>
<td></td>
<td>(27.35)</td>
<td>(194.5)</td>
<td>(87.14)</td>
<td>(65.33)</td>
<td>(60.49)</td>
<td>(73.69)</td>
<td>(140.8)</td>
<td>(35.58)</td>
</tr>
<tr>
<td><strong>Observation Treatment</strong>*Female</td>
<td>18.25</td>
<td>266.0</td>
<td>-21.73</td>
<td>16.57</td>
<td>60.82</td>
<td>-53.11</td>
<td>-78.88</td>
<td>21.19</td>
</tr>
<tr>
<td></td>
<td>(29.29)</td>
<td>(228.2)</td>
<td>(102.1)</td>
<td>(68.18)</td>
<td>(68.37)</td>
<td>(83.01)</td>
<td>(143.4)</td>
<td>(35.24)</td>
</tr>
<tr>
<td></td>
<td>(103.5)</td>
<td>(4.062)</td>
<td>(44.75)</td>
<td>(29.80)</td>
<td>(29.92)</td>
<td>(33.83)</td>
<td>(3.919)</td>
<td>(96.14)</td>
</tr>
<tr>
<td><strong>Interaction</strong>*Female**</td>
<td>80.48</td>
<td>3.609</td>
<td>16.45</td>
<td>21.53</td>
<td>31.23</td>
<td>-9.175</td>
<td>-1.872</td>
<td>52.71</td>
</tr>
<tr>
<td></td>
<td>(52.47)</td>
<td>(3.952)</td>
<td>(25.28)</td>
<td>(23.73)</td>
<td>(19.48)</td>
<td>(16.29)</td>
<td>(2.496)</td>
<td>(63.01)</td>
</tr>
<tr>
<td><strong>Interaction</strong><em>Observation Treatment</em>**Female</td>
<td>-40.40</td>
<td>-5.288</td>
<td>16.61</td>
<td>0.621</td>
<td>-22.24</td>
<td>25.05</td>
<td>2.360</td>
<td>-46.73</td>
</tr>
<tr>
<td></td>
<td>(112.3)</td>
<td>(4.670)</td>
<td>(47.72)</td>
<td>(30.75)</td>
<td>(32.75)</td>
<td>(34.35)</td>
<td>(3.927)</td>
<td>(95.41)</td>
</tr>
<tr>
<td><strong>Constant</strong></td>
<td>418.2***</td>
<td>674.0***</td>
<td>494.1***</td>
<td>480.7***</td>
<td>464.0***</td>
<td>435.7***</td>
<td>77.17</td>
<td>444.4***</td>
</tr>
<tr>
<td></td>
<td>(22.38)</td>
<td>(198.4)</td>
<td>(65.65)</td>
<td>(64.46)</td>
<td>(50.80)</td>
<td>(47.43)</td>
<td>(90.57)</td>
<td>(27.49)</td>
</tr>
</tbody>
</table>

| Observations | 2614 | 2614 | 2614 | 2614 | 2614 | 2614 | 2269 | 2269 |
| Clusters     | 147  | 147  | 147  | 147  | 147  | 147  | 135  | 135  |
| Dep. Var. Mean | 368.4 | 368.4 | 368.4 | 368.4 | 368.4 | 368.4 | 368.4 | 368.4 |
| Dep. Var. SD  | 176.3 | 176.3 | 176.3 | 176.3 | 176.3 | 176.3 | 176.3 | 176.3 |
## Heterogeneous Raise

### Predicted Raise Amount (USD)

<table>
<thead>
<tr>
<th></th>
<th>(1) Male</th>
<th>(2) Age</th>
<th>(3) Avg. Bias</th>
<th>(4) Math</th>
<th>(5) Jobs</th>
<th>(6) Family</th>
<th>(7) Teacher Age</th>
<th>(8) Young</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>16.41</td>
<td>-29.07</td>
<td>-2.460</td>
<td>-2.877</td>
<td>-13.83</td>
<td>26.73</td>
<td>-86.27</td>
<td>18.62</td>
</tr>
<tr>
<td></td>
<td>(12.73)</td>
<td>(97.02)</td>
<td>(38.67)</td>
<td>(30.58)</td>
<td>(28.62)</td>
<td>(28.11)</td>
<td>(53.38)</td>
<td>(15.21)</td>
</tr>
<tr>
<td></td>
<td>(41.23)</td>
<td>(1.870)</td>
<td>(13.57)</td>
<td>(11.24)</td>
<td>(11.13)</td>
<td>(8.248)</td>
<td>(1.580)</td>
<td>(35.30)</td>
</tr>
<tr>
<td>Interaction*Female</td>
<td>1.292</td>
<td>0.958</td>
<td>7.797</td>
<td>7.767</td>
<td>15.55</td>
<td>-4.596</td>
<td>2.821**</td>
<td>-2.957</td>
</tr>
<tr>
<td></td>
<td>(39.72)</td>
<td>(2.003)</td>
<td>(15.40)</td>
<td>(12.21)</td>
<td>(12.13)</td>
<td>(9.462)</td>
<td>(1.395)</td>
<td>(32.51)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Observations</th>
<th>Clusters</th>
<th>Dep. Var. Mean</th>
<th>Dep. Var. SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2018</td>
<td>170</td>
<td>368.4</td>
<td>176.3</td>
</tr>
<tr>
<td></td>
<td>2018</td>
<td>170</td>
<td>368.4</td>
<td>176.3</td>
</tr>
<tr>
<td></td>
<td>2018</td>
<td>170</td>
<td>368.4</td>
<td>176.3</td>
</tr>
<tr>
<td></td>
<td>2018</td>
<td>170</td>
<td>368.4</td>
<td>176.3</td>
</tr>
<tr>
<td></td>
<td>2018</td>
<td>170</td>
<td>368.4</td>
<td>176.3</td>
</tr>
<tr>
<td></td>
<td>1728</td>
<td>156</td>
<td>368.4</td>
<td>176.3</td>
</tr>
<tr>
<td></td>
<td>1728</td>
<td>156</td>
<td>368.4</td>
<td>176.3</td>
</tr>
<tr>
<td></td>
<td>1728</td>
<td>156</td>
<td>368.4</td>
<td>176.3</td>
</tr>
</tbody>
</table>
## Effect of Treatments on Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Predicted Raise Amount (USD)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td></td>
<td>-28.21*** (6.600)</td>
<td>-2.228 (12.59)</td>
<td>-41.90*** (13.15)</td>
<td>-18.43 (18.71)</td>
</tr>
<tr>
<td>Financial Treatment</td>
<td></td>
<td>23.36 (20.81)</td>
<td>36.88 (28.91)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Financial Treatment*Female</td>
<td></td>
<td>-34.33** (15.32)</td>
<td>-51.01** (25.65)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observation Treatment</td>
<td></td>
<td>-36.46** (15.93)</td>
<td>-15.51 (25.09)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observation Treatment*Female</td>
<td></td>
<td>41.55** (17.07)</td>
<td>31.46 (27.42)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Financial Treatment*Observation Treatment</td>
<td></td>
<td>-46.46 (31.71)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Financial Treatment<em>Observation Treatment</em>Female</td>
<td></td>
<td>37.52 (34.62)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Observations</th>
<th>Clusters</th>
<th>Dep. Var. Mean</th>
<th>Dep. Var. SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5051</td>
<td>4300</td>
<td>2626</td>
<td>2326</td>
</tr>
<tr>
<td></td>
<td>.</td>
<td>263</td>
<td>.</td>
<td>158</td>
</tr>
<tr>
<td></td>
<td>365.4</td>
<td>365.4</td>
<td>365.4</td>
<td>365.4</td>
</tr>
<tr>
<td></td>
<td>164.7</td>
<td>164.7</td>
<td>164.7</td>
<td>164.7</td>
</tr>
</tbody>
</table>
Percentile Value Added

• Construction of the value added percentile:
  • Within each grade/year/subject bin, calculate each student’s percentile rank.
  • For the following year’s score, construct the student’s percentile within the lagged percentile-grade-subject bin.
  • Compute the teacher’s percentile in a given year by taking the average across all students

• Reasons for using percentile measure
  • Barlevy and Neal (2016) show results are similar to other value added models
  • Only relies on ordinal information allowing for new tests each year (less susceptible to manipulation)
  • Muralidharan/Walters and Lucas/Neal use same approach in India and Uganda, respectively
Percentile Value Added

• Validating the Percentile Value Added
  • Year to year correlation
    • Standard models: 0.4
    • Our measure: 0.56
  • Increase in first 5 years of teaching
    • Standard models: 0.5
    • Our measure: 0.35

• Correlation with Other VA Models
  • Controlling for lagged score in the same subject: 0.44
  • CFR 2013: 0.25
## Balance in Baseline Covariates

### Panel A: Teacher Characteristics

<table>
<thead>
<tr>
<th>Variable</th>
<th>(1) N/ Mean/ SE</th>
<th>(2) N/ Mean/ SE</th>
<th>(3) N/ Mean/ SE</th>
<th>T-test Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Performance evaluation score</strong></td>
<td>656/3.360(0.030)</td>
<td>384/3.362(0.039)</td>
<td>3566/3.338(0.010)</td>
<td>-0.002 0.022 0.024</td>
</tr>
<tr>
<td><strong>Salary (USD)</strong></td>
<td>920/5417.984(313.504)</td>
<td>535/5125.462(295.013)</td>
<td>4928/5329.416(124.042)</td>
<td>292.523 88.569 -203.954</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td>921/36.591(0.738)</td>
<td>539/36.083(0.846)</td>
<td>4926/36.630(0.298)</td>
<td>0.507 -0.039 -0.546</td>
</tr>
<tr>
<td><strong>Years of experience</strong></td>
<td>918/5.505(0.277)</td>
<td>534/5.487(0.425)</td>
<td>4897/5.725(0.156)</td>
<td>0.019 -0.220 -0.238</td>
</tr>
</tbody>
</table>

### Panel B: Student Test Scores

<table>
<thead>
<tr>
<th>Variable</th>
<th>(1) N/ Mean/ SE</th>
<th>(2) N/ Mean/ SE</th>
<th>(3) N/ Mean/ SE</th>
<th>T-test Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Math Test Z-Score</strong></td>
<td>9959/0.071(0.070)</td>
<td>5292/-0.146(0.065)</td>
<td>51775/-0.014(0.026)</td>
<td>0.217** 0.085 -0.132*</td>
</tr>
<tr>
<td><strong>Urdu Test Z-Score</strong></td>
<td>9702/0.041(0.072)</td>
<td>5259/-0.048(0.063)</td>
<td>50915/-0.002(0.028)</td>
<td>0.089 0.043 -0.046</td>
</tr>
<tr>
<td><strong>English Test Z-Score</strong></td>
<td>9755/0.017(0.056)</td>
<td>5289/-0.049(0.050)</td>
<td>51356/0.002(0.032)</td>
<td>0.067 0.016 -0.051</td>
</tr>
<tr>
<td><strong>Social Studies Test Z-Score</strong></td>
<td>9171/0.041(0.046)</td>
<td>5030/-0.064(0.056)</td>
<td>49411/0.007(0.022)</td>
<td>0.105 0.033 -0.071</td>
</tr>
<tr>
<td><strong>Science Test Z-Score</strong></td>
<td>9636/-0.010(0.041)</td>
<td>5065/-0.064(0.042)</td>
<td>50268/0.001(0.024)</td>
<td>0.055 -0.011 -0.066</td>
</tr>
</tbody>
</table>
## Endline Student Survey

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

What we know

1. What we know about the ability for contracts to screen types
   - Lazear (other general ad sel lit)
2. Make clear tension between lit that suggests effects should be large vs. lit that predicts effects are zero and why this setting is different than Lazear 2000
   - Mention barbara, jesse and owen
3. Performance Pay literature: lots of great stuff but missing sorting
Distortion and Noise

For example, a school’s value function, $V$ may be that they value test scores and socio-emotional outcomes at a 2:1 ratio.
Distortion and Noise

Distortion is captures how aligned the incentive scheme is with the actions which produce $V$. 
Distortion and Noise

Distortion is captures how aligned the incentive scheme is with the actions which produce $V$. 

![Diagram showing the relationship between Test scores and Socio-emotional skills, with an angle $\theta$.]
Distortion and Noise

Distortion is captures how aligned the incentive scheme is with the actions which produce $V$.
Distortion and Noise

Noise determines how high-powered the incentives are and hence, how large the effort response is
Distortion and Noise

Noise determines how high-powered the incentives are and hence, how large the effort response is.
Distortion and Noise

For example, here is an incentive scheme which pays based on endline test scores.
Distortion and Noise

For example, here is an incentive scheme which pays based on endline test scores

\[ \text{Test scores} \]
\[ \text{Socio-emotional skills} \]
\[ \theta \]
\[ \text{Teacher effort response} \]
\[ V \]
\[ \text{Test scores} \]
Experimental Design

- **2017**
  - Oct: Teacher & Manager Baseline Survey
  - Dec: End Year Performance Eval & Goal Setting

- **2018**
  - Feb: Control randomize
  - Treatment Info Campaign
  - July: Treatment reminder & midterm info
  - Fall: Classroom observation
  - Dec: Manager Evaluation

- **2019**
  - Jan: Student Testing/Survey
  - March: Raises announced
  - May: Teacher & Manager Endline Survey
Teacher Evaluation Vignettes

Example vignette:

“Haya is in the bottom 10% of teachers in terms of students’ test score growth, in the middle 10% of teachers in terms of behavioral management, and is in the top 10% in terms of attendance and timeliness at work. If you had to give her a performance evaluation score, what score would you assign to her?”
# Student Outcomes - Test Scores

## Endline Test (z-score)

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

*Significance levels: *: p < 0.1, **: p < 0.05, ***: p < 0.01
### Student Outcomes - Socio-emotional

<table>
<thead>
<tr>
<th></th>
<th>All (1)</th>
<th>Remedial (2)</th>
<th>External (3)</th>
<th>Math/Science (4)</th>
<th>English/Urdu (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objective Treatment</strong></td>
<td>0.0918*</td>
<td>0.189***</td>
<td>0.119**</td>
<td>0.104*</td>
<td>0.0917</td>
</tr>
<tr>
<td></td>
<td>(0.0575)</td>
<td>(0.00518)</td>
<td>(0.0335)</td>
<td>(0.0668)</td>
<td>(0.166)</td>
</tr>
<tr>
<td></td>
<td>[0.0730]</td>
<td>[0.0260]</td>
<td>[0.0200]</td>
<td>[0.194]</td>
<td>[0.144]</td>
</tr>
<tr>
<td><strong>Subjective Treatment</strong></td>
<td>0.0859**</td>
<td>0.142**</td>
<td>0.0855*</td>
<td>0.0884*</td>
<td>0.0986**</td>
</tr>
<tr>
<td></td>
<td>(0.0220)</td>
<td>(0.0113)</td>
<td>(0.0601)</td>
<td>(0.0646)</td>
<td>(0.0267)</td>
</tr>
<tr>
<td></td>
<td>[0.0130]</td>
<td>[0.0240]</td>
<td>[0.0170]</td>
<td>[0.121]</td>
<td>[0.0260]</td>
</tr>
<tr>
<td><strong>F-test pval (subj=OBJ)</strong></td>
<td>0.89</td>
<td>0.38</td>
<td>0.43</td>
<td>0.77</td>
<td>0.90</td>
</tr>
<tr>
<td><strong>Randomiz infer pval (subj=OBJ)</strong></td>
<td>0.884</td>
<td>0.453</td>
<td>0.388</td>
<td>0.819</td>
<td>0.873</td>
</tr>
</tbody>
</table>

| Control Group Mean     | -0.04  | -0.09        | -0.05        | -0.04            | -0.04            |
| Clusters               | 234    | 204          | 225          | 223              | 225              |
| Observations           | 141566 | 31944        | 100318       | 72714            | 68852            |
## Teacher Effort - Teaching Practices

<table>
<thead>
<tr>
<th></th>
<th>Classroom Observation Rubric</th>
<th>Test Prep</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All (1)</td>
<td>Class Climate (2)</td>
</tr>
<tr>
<td>Objective Treatment</td>
<td>-0.0713</td>
<td>-0.0791*</td>
</tr>
<tr>
<td></td>
<td>(0.123)</td>
<td>(0.0788)</td>
</tr>
<tr>
<td></td>
<td>[0.171]</td>
<td>[0.101]</td>
</tr>
<tr>
<td>Subjective Treatment</td>
<td>-0.00206</td>
<td>-0.00704</td>
</tr>
<tr>
<td></td>
<td>(0.959)</td>
<td>(0.822)</td>
</tr>
<tr>
<td></td>
<td>[0.946]</td>
<td>[0.838]</td>
</tr>
<tr>
<td>F-test pval (subj=obj)</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>Randomiz infer pval</td>
<td>0.109</td>
<td>0.0830</td>
</tr>
<tr>
<td>(subj=obj)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control Group Mean</td>
<td>4.67</td>
<td>5.64</td>
</tr>
<tr>
<td>Clusters</td>
<td>142</td>
<td>142</td>
</tr>
<tr>
<td>Observations</td>
<td>6827</td>
<td>6827</td>
</tr>
</tbody>
</table>
## Design - Teacher Sample

### Panel A. Teacher Characteristics

<table>
<thead>
<tr>
<th></th>
<th>Study Sample Mean (1)</th>
<th>Study Sample St. Dev. (2)</th>
<th>US Sample Mean (3)</th>
<th>US Sample St. Dev. (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>35.0</td>
<td>8.9</td>
<td>41.8</td>
<td>7.5</td>
</tr>
<tr>
<td>Female</td>
<td>0.80</td>
<td>0.40</td>
<td>0.76</td>
<td>0.43</td>
</tr>
<tr>
<td>Years of experience</td>
<td>5.1</td>
<td>5.2</td>
<td>13.8</td>
<td>9.6</td>
</tr>
<tr>
<td>Has Post BA Education</td>
<td>0.68</td>
<td>0.47</td>
<td>0.54</td>
<td>0.50</td>
</tr>
<tr>
<td>Salary, USD(PPP)</td>
<td>17,160</td>
<td>5,700</td>
<td>52,400</td>
<td>18,400</td>
</tr>
</tbody>
</table>

### Panel B. Teacher Evaluation

<table>
<thead>
<tr>
<th></th>
<th>Study Sample Mean (1)</th>
<th>Study Sample St. Dev. (2)</th>
<th>US Sample Mean (3)</th>
<th>US Sample St. Dev. (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations per year</td>
<td>4.7</td>
<td>8.2</td>
<td>2.5</td>
<td>2.9</td>
</tr>
<tr>
<td>Use evaluation for compensation</td>
<td>-</td>
<td>-</td>
<td>0.12</td>
<td>0.32</td>
</tr>
<tr>
<td>Frequency of evaluation (months)</td>
<td>-</td>
<td>-</td>
<td>13.0</td>
<td>7.0</td>
</tr>
<tr>
<td>Performance metric used for evaluation:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Principal evaluation</td>
<td>-</td>
<td>-</td>
<td>0.90</td>
<td>0.30</td>
</tr>
<tr>
<td>Test scores</td>
<td>-</td>
<td>-</td>
<td>0.35</td>
<td>0.48</td>
</tr>
<tr>
<td>Peer evaluations</td>
<td>-</td>
<td>-</td>
<td>0.26</td>
<td>0.44</td>
</tr>
<tr>
<td>Student ratings</td>
<td>-</td>
<td>-</td>
<td>0.05</td>
<td>0.22</td>
</tr>
</tbody>
</table>
### Appendix

#### Design - Manager Sample

<table>
<thead>
<tr>
<th></th>
<th>Study Sample</th>
<th></th>
<th>US Sample</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (1)</td>
<td>St. Dev. (2)</td>
<td>Mean (3)</td>
<td>St. Dev. (4)</td>
</tr>
<tr>
<td>Age</td>
<td>44.9</td>
<td>9.2</td>
<td>48.8</td>
<td>9.7</td>
</tr>
<tr>
<td>Female</td>
<td>0.61</td>
<td>0.49</td>
<td>0.53</td>
<td>0.50</td>
</tr>
<tr>
<td>Years of experience</td>
<td>9.6</td>
<td>7.9</td>
<td>13.0</td>
<td>7.5</td>
</tr>
</tbody>
</table>

**Panel A. Manager Characteristics**

**Panel B. Manager Time Use**

- Total hours worked 47.2  16.3  57.0  13.2
- Hours spent on:
  - Administrative tasks 18.5  10.3  18.2  2.3
  - Teacher management and teaching 17.5  8.2  15.1  2.0
  - Student and parent interactions 6.3  4.4  20.2  2.7
  - Other tasks 6.9  12.3  4.0  2.6

**Panel C. Management Practice Rating**

- Overall Management Score (out of 5) 4.27  0.43  2.76  0.43
- People management (out of 5) 4.14  0.53  2.51  0.49
- Operations (out of 5) 4.32  0.61  2.89  0.49
- Performance monitoring (out of 5) 4.32  0.49  2.81  0.75
What do principals value?

We give principal short vignettes describing an example teacher and ask them to give a hypothetical evaluation score. 

Vignette text
Appendix

What do principals value?

We give principal short vignettes describing an example teacher and ask them to give a hypothetical evaluation score.

<table>
<thead>
<tr>
<th>Vignette text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female name</td>
</tr>
<tr>
<td>Teacher percentile (0-1) in:</td>
</tr>
<tr>
<td>- Value-added</td>
</tr>
<tr>
<td>- Behavioral Management</td>
</tr>
<tr>
<td>- Attendance</td>
</tr>
</tbody>
</table>

Principal rating of teacher in vignette (SD)

Principals appear to value teacher test score value-added about twice as much as attendance or behavioral management.
### Heterogeneous Effects by Manager Characteristics

Subjective incentives appear to be effective for all but the bottom quintile of managers

<table>
<thead>
<tr>
<th>Endline Test Scores</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjective Treatment</td>
<td>-0.0156</td>
<td>0.169**</td>
<td>-0.0566</td>
<td>0.249***</td>
</tr>
<tr>
<td></td>
<td>(0.197)</td>
<td>(0.0688)</td>
<td>(0.117)</td>
<td>(0.0775)</td>
</tr>
<tr>
<td>Interaction</td>
<td>0.00111</td>
<td>0.00827</td>
<td>0.0159</td>
<td>0.142*</td>
</tr>
<tr>
<td></td>
<td>(0.00274)</td>
<td>(0.00503)</td>
<td>(0.0977)</td>
<td>(0.0763)</td>
</tr>
<tr>
<td>Interaction*Subjective Treatment</td>
<td>0.00205</td>
<td>-0.00883</td>
<td>0.148</td>
<td>-0.211**</td>
</tr>
<tr>
<td></td>
<td>(0.00420)</td>
<td>(0.00648)</td>
<td>(0.127)</td>
<td>(0.0910)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interaction</th>
<th>Age</th>
<th>Experience (years)</th>
<th>Female</th>
<th>Manager inaccuracy (z-score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clusters</td>
<td>255</td>
<td>255</td>
<td>255</td>
<td>255</td>
</tr>
<tr>
<td>Observations</td>
<td>440595</td>
<td>440595</td>
<td>440595</td>
<td>440595</td>
</tr>
</tbody>
</table>
Principal-Agent Model with Moral Hazard

Using Baker (2002), $V$, outcome (student learning), is a function of their teacher’s effort, $a$, the return to those actions $f$:

$$V(a, e) = f \cdot a + e = f_1 a_1 + f_2 a_2 + ... + e$$ (4)
Principal-Agent Model with Moral Hazard

Using Baker (2002), $V$, outcome (student learning), is a function of their teacher’s effort, $\vec{a}$, the return to those actions $\vec{f}$:

$$V(\vec{a}, e) = \vec{f} \cdot \vec{a} + e = f_1 a_1 + f_2 a_2 + ... + e$$  \hspace{1cm} (4)$$

Teacher’s pay under an incentive contract is a function of those actions, the piece rate for each action $\vec{g}$, and noise, $\phi$:

$$P(\vec{a}, \phi) = \vec{g} \cdot \vec{a} + \phi = g_1 a_1 + g_2 a_2 + ... + \phi$$  \hspace{1cm} (5)$$
Principal-Agent Model with Moral Hazard

Using Baker (2002), $V$, outcome (student learning), is a function of their teacher’s effort, $\vec{a}$, the return to those actions $\vec{f}$:

$$V(\vec{a}, e) = \vec{f} \cdot \vec{a} + e = f_1a_1 + f_2a_2 + ... + e$$  \hspace{1cm} (4)

Teacher’s pay under an incentive contract is a function of those actions, the piece rate for each action $\vec{g}$, and noise, $\phi$:

$$P(\vec{a}, \phi) = \vec{g} \cdot \vec{a} + \phi = g_1a_1 + g_2a_2 + ... + \phi$$  \hspace{1cm} (5)

Assuming a quadratic cost of effort, then optimal effort will be $\vec{a}^* = \vec{g}$, and average student learning will be:

$$E[V^*(\vec{a}^*, e)] = \vec{f} \cdot \vec{g} = |f||g|\cos\theta$$  \hspace{1cm} (6)
Principal-Agent Model with Moral Hazard

Taking the variance of (2), we have $\text{var}(P) = |g|^2 \text{var}(\vec{a}) + \sigma^2_\phi$. Re-arranging, we can substitute this in for $|g|$ in to (3).
Appendix

Principal-Agent Model with Moral Hazard

Taking the variance of (2), we have \( \text{var}(P) = |g|^2 \text{var}(\vec{a}) + \sigma_\phi^2 \). Re-arranging, we can substitute this in for \(|g|\) in to (3).

Average student learning under a given incentive scheme is:

\[
E[V^*(\vec{a}^*, e)] = |f| |g| \cos \theta
\]

\[
= |f| \sqrt{\text{var}(P) - \sigma_\phi^2} \cos \theta
\]

\[
= |f| \sqrt{\text{var}(\vec{a})} \cos \theta
\]

Key predictions – Student learning is:

(a) decreasing in contract noise, \( \sigma_\phi^2 \)

(b) increasing in alignment (lack of distortion), \( \cos \theta \)
Principal-Agent Model with Moral Hazard

Taking the variance of (2), we have \( \text{var}(P) = |g|^2 \text{var}(\hat{a}) + \sigma^2_\phi \). Re-arranging, we can substitute this in for \(|g|\) in to (3).

Average student learning under a given incentive scheme is:

\[
E[V^*(\hat{a}^*, e)] = |f||g| \cos \theta = |f| \frac{\sqrt{\text{var}(P) - \sigma^2_\phi}}{\sqrt{\text{var}(\hat{a})}} \cos \theta
\]

|f|, \( \text{var}(P) \) and \( \text{var}(\hat{a}) \) are constant across the two incentive schemes we'll compare (feature of any within-firm tournament)
Principal-Agent Model with Moral Hazard

Taking the variance of (2), we have \( \text{var}(P) = |g|^2 \text{var}(\vec{a}) + \sigma^2 \). Re-arranging, we can substitute this in for \( |g| \) in to (3).

Average student learning under a given incentive scheme is:

\[
E[V^*(\vec{a}^*, e)] = |f||g|\cos\theta
= |f|\sqrt{\text{var}(P) - \sigma^2_\phi} \cos\theta
\]

\(|f|, \text{var}(P)\) and \( \text{var}(\vec{a}) \) are constant across the two incentive schemes we'll compare (feature of any within-firm tournament)
Appendix

Principal-Agent Model with Moral Hazard

Taking the variance of (2), we have $\text{var}(P) = |g|^2 \text{var}(\vec{a}) + \sigma^2_\phi$. Re-arranging, we can substitute this in for $|g|$ in (3).

Average student learning under a given incentive scheme is:

\[
E[V^*(\vec{a}^*, e)] = |f||g| \cos \theta \\
= |f| \frac{\sqrt{\text{var}(P) - \sigma^2_\phi}}{\sqrt{\text{var}(\vec{a})}} \cos \theta
\]

$|f|$, $\text{var}(P)$ and $\text{var}(\vec{a})$ are constant across the two incentive schemes we'll compare (feature of any within-firm tournament)

Key predictions – Student learning is:

(b) increasing in alignment (lack of distortion), $\cos(\theta)$

(a) decreasing in contract noise, $\sigma^2_\phi$
## Example evaluation criteria

### Plan 1: Manager Appraisal of Effort

<table>
<thead>
<tr>
<th>Effort Criteria</th>
<th>Objective Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessment of student understanding (monitoring of student learning,</td>
<td>20</td>
</tr>
<tr>
<td>effective and timely copy checking)</td>
<td></td>
</tr>
<tr>
<td>Differentiated lessons for varying learning needs</td>
<td>30</td>
</tr>
<tr>
<td>Effectively delivering accurate and relevant content (effective</td>
<td>30</td>
</tr>
<tr>
<td>implementation of the curriculum)</td>
<td></td>
</tr>
<tr>
<td>Providing caring, supportive environment</td>
<td>20</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>
Contribution

- Financial Discrimination
  Demonstrate gender discrimination
  cites
Contribution

- **Financial Discrimination**
  Demonstrate gender discrimination
  cites

- **Statistical discrimination**
  Text
  Cites
Contribution

- **Financial Discrimination**
  Demonstrate gender discrimination
  cites

- **Statistical discrimination**
  Text
  Cites

- **Disparate outcomes**
  Model with
  Baker, 2002; Prendergast, 1999; Prendergast and Topel, 1993; Prendergast, 2007
Design - Timeline

Baseline survey
- Informed about incentive schemes
- Measure teacher characteristics
Design - Timeline

- Research team meets in person with managers
- School system HR does in person presentation at each school
- Email information
- Displayed on teacher’s dashboard

Performance Eval Info (Treatment 1)

- Managers set criteria for ALL teachers (not new to experiment)

Baseline survey

Perf. Criteria Set

Example criteria

= Data collection activity

= Treatment implementation

= Both
Appendix

Design - Timeline

- **Baseline survey**
- **Performance Eval Info (Treatment 1)**
- **Clock in/out data**
- **Classroom video-taping**

- Record 5 hours of class time for 1500 teachers
- Tripod in back of classroom to minimize interference

Example criteria

- Blue box = Data collection activity
- Yellow box = Treatment implementation
- Green box = Both
Appendix

Design - Timeline

- Baseline survey
- Performance Eval Info (Treatment 1)
- Classroom video-taping
- Clock in/out data
- Perf. Criteria Set
- Manager Observation (Treatment 2)
- Managers asked to observe random set of teachers
- Managers score teachers on criteria
- Perf. Eval

Example criteria

Example criteria

= Data collection activity
= Treatment implementation
= Both
Appendix

Design - Timeline

- Baseline survey
- Performance Eval Info (Treatment 1)
- Classroom video-taping
- Clock in/out data
- Perf. Criteria Set
- Manager Observation (Treatment 2)
- Perf. Eval
- Raises set
- Test designed and graded by research team
- Teacher/Manager Survey (Endline)

Example criteria

= Data collection activity
= Treatment implementation
= Both
# Teacher and Manager Sample

<table>
<thead>
<tr>
<th></th>
<th>Study Sample</th>
<th>US Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>St. Dev.</td>
</tr>
<tr>
<td><strong>Panel A. Teacher Characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>35.0</td>
<td>8.9</td>
</tr>
<tr>
<td>Female</td>
<td>0.80</td>
<td>0.40</td>
</tr>
<tr>
<td>Years of experience</td>
<td>5.1</td>
<td>5.2</td>
</tr>
<tr>
<td>Has Post BA Education</td>
<td>0.68</td>
<td>0.47</td>
</tr>
<tr>
<td>Salary, USD</td>
<td>4,000</td>
<td>1,700</td>
</tr>
<tr>
<td><strong>Panel B. Manager Characteristics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>44.9</td>
<td>9.2</td>
</tr>
<tr>
<td>Female</td>
<td>0.61</td>
<td>0.49</td>
</tr>
<tr>
<td>Years of experience</td>
<td>9.6</td>
<td>7.9</td>
</tr>
<tr>
<td><strong>Panel C. Manager Time Use</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total hours worked</td>
<td>47.2</td>
<td>16.3</td>
</tr>
<tr>
<td>Hours spent on:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Administrative tasks</td>
<td>18.5</td>
<td>10.3</td>
</tr>
<tr>
<td>- Teacher management and teaching</td>
<td>17.5</td>
<td>8.2</td>
</tr>
<tr>
<td>- Student and parent interactions</td>
<td>6.3</td>
<td>4.4</td>
</tr>
<tr>
<td>- Other tasks</td>
<td>6.9</td>
<td>12.3</td>
</tr>
</tbody>
</table>

Col. 3 and 4 Source: School and Staff Survey (National Center for Education Statistics)