Pay by Design: Teacher Performance Pay Experiment in Rural China

Pre-Analysis Plan

Prashant Loyalka
Sean Sylvia
Chengfang Liu
James Chu
Yaojiang Shi
Linxiu Zhang
Scott Rozelle

This document outlines the pre-analysis plan for analyzing the impacts of the “Pay by Design: Teacher Performance Pay Experiment in Rural China” Trial. The goal of this document is to pre-specify our planned methodology and specifications for analyzing the impacts of the program. In addition to the ex-ante planned analysis, we may conduct additional exploratory analyses (as these are an important means to discovery). However, this document serves as a record of our planned analyses and all analyses not planned ex-ante will be marked as such when reporting results.

I. Overview of the Study

Growing evidence suggests that teachers in developing countries often have weak or misaligned incentives for improving student outcomes. In response, policymakers and researchers have proposed performance pay as a way to improve student outcomes by tying concrete measures like achievement scores to teacher pay (Lazear, 2003).

While evidence from randomized experiments generally indicates that performance pay programs are effective at improving student achievement in developing countries, there has been considerable variation in how much these programs affect student achievement. For example, a performance pay program for teachers in India had positive/long-lasting impacts on achievement scores (Muralidharan and Sundararaman, 2011; Muralidharan, 2011). Earlier studies from other developing countries also showed positive but much smaller impacts (Glewwe et al., 2003; Lavy 2009).

One potential reason there may be such high levels of variation in impacts is because each performance pay program was designed differently. Although, taken together, the evidence suggests that performance pay holds promise for improving student achievement in developing countries, the variation in impact suggests that there are more or less effective ways to design

---

1 This plan was written before the endline data were examined, cleaned and analyzed.
2 In contrast, experimental studies of performance pay for teachers in developed countries show little or no impact on student achievement (Fryer, 2011, Springer et al., 2010).
performance pay. Because existing studies were conducted in various contexts and differ in a number of dimensions, at present we are not able to determine how specific design features affect the results. The first question pursued by our study is therefore the following: How should a school system design performance pay programs to maximize gains in student achievement?

Another largely unanswered question is how performance pay programs affect the achievement of different types of students. There are strong theoretical reasons to believe that certain types of teacher incentives could benefit students at certain parts of the achievement distribution (e.g. lower achieving students) more than other students (Neal, 2011). Neal and Schanzenbach (2010) do show that the introduction of test-based school accountability systems benefit students who are in certain segments of the achievement distribution more than others (e.g. higher or middle achieving students compared to lower achieving students). However, few if any studies explore how different performance pay programs affect students at certain segments of the achievement distribution more than others.

The questions about how to design performance pay programs to maximize achievement gains, in general, and how to design a program to help lower achieving students, in particular, are especially salient for China. The Teacher Performance Pay Policy of 2009 (State Council, 2009) raised rural teacher salaries. Moreover, the Policy stipulated that 30% of the raise should be based on performance measures. Unfortunately, education officials were silent on exactly how teachers should be rewarded. Consequently, the implementation of the policy varies widely. By testing alternative design features of performance pay, this study can provide guidance on how to raise the returns to the financial resources that support the Teacher Performance Pay Policy.

Performance Pay Designs

Our study focuses on two major design features of performance pay programs. We choose these design features since we believe that they have never been evaluated systematically in an experimental context and are likely to have large implications. The first design feature concerns the way in which teacher performance is evaluated in determining performance pay. The second design feature concerns the payout size—or strength—of performance pay.

In regards to the first design feature (the way in which teacher performance is evaluated in determining performance pay), we consider alternative ways of linking performance pay to student achievement (a commonly used measure of teacher performance). In many developing countries, such as China, teachers have traditionally been rewarded for the levels of their students’ achievement (e.g., the high the average achievement scores, the more a teacher was paid—State Council, 2009). However, a trend in performance pay programs in developed countries is to reward teachers for average gains in achievement scores. Performance pay for gains could improve student achievement more than performance pay for levels, since gains better reflect teacher effort (Hanushek et al., 2010). We, therefore, first consider examine whether linking teacher pay to student achievement gains is more effective than linking pay to levels.
Beyond using student achievement gains or levels, a third way of linking performance pay to student achievement adjusts for student background when creating performance measures. Programs that do not compare students that start with similar levels of academic achievement (when measuring teacher performance) may not maximize student achievement gains (Neal, 2011). These programs may also create incentives for teachers to avoid certain types of students (Neal, 2011). Specifically, performance pay that does not compare students of similar achievement levels can lead teachers to focus on students in the middle of the achievement distribution at the cost of lower achieving students.

To address this concern, researchers have suggested using performance pay designs that relate teacher rewards to the achievement gains of their students within appropriately defined comparison sets (Neal, 2011). In particular, a pay for percentile program has been shown (theoretically) to discourage teachers from focusing on children at the higher and middle segments of the achievement distribution at the expense of children at the lower end of the distribution (Neal, 2011). While the theory behind pay for percentile is compelling, it remains empirically untested.

A second performance pay design feature is the degree to which the size of the incentive payout—or strength of the incentive—matters in a performance pay program. On the one hand, incentivizing teachers with potentially larger payouts may be more effective than incentivizing teachers with smaller payouts. On the other hand, incentivizing teachers with smaller payouts may be just as effective (and therefore more cost-effective). This would be the case, for example, if there were diminishing marginal returns to providing teachers with performance pay.

With these three major design features in mind, the goals of our study are to:

(a) examine the impacts of different teacher performance pay designs (the evaluation method and payout size) on student achievement, both for the average student and for students across the baseline achievement distribution;

(b) examine the mechanisms through which different teacher performance pay designs affect student achievement (for the average student and for students across the baseline achievement distribution).

II. Research Design

To test the impacts of the different teacher performance pay designs discussed above, we designed a cluster-randomized controlled trial. In this trial, schools were randomly allocated to different treatment arms in a four arm by two arm (4X2) crosscutting experimental design. The two dimensions of the experiment are as follows:
The first dimension of the experimental design randomly allocated schools into the following four evaluation method design feature arms:

TREATMENT ARM A - CONTROL (52 schools)
Grade 6 teachers received NO performance pay contract;

TREATMENT ARM B - LEVELS (54 schools)
Grade 6 teachers received a performance pay contract stipulating rewards based on student achievement levels on endline tests;

TREATMENT ARM C - GAINS (56 schools)
Grade 6 teachers received a performance pay contract based on student achievement gains from baseline and endline tests;

TREATMENT ARM D - PAY FOR PERCENTILE (54 schools):
Grade 6 teachers received a performance pay contract stipulating rewards based on student growth percentiles (see Betebenner, 2011) estimated using baseline and endline tests.

The second dimension (orthogonal to the first dimension) of the experimental design further randomly assigned half of the schools in each of Treatment Arms B, C, and D to one of the following two payout size design feature arms:

TREATMENT ARM X – LARGE INCENTIVE PAYOUT (26 schools in Arm B, 26 schools in Arm C, and 26 schools in Arm D = 78 schools in total)
Grade 6 math teachers received a large (payout) incentive contract. The large incentive contract would pay teachers 7000 yuan if they ranked in the top percentile of participating teachers and decline by 70 yuan for each percentile thereafter. Teachers ranking in the bottom percentile would receive 70 yuan.

TREATMENT ARM Y – SMALL INCENTIVE PAYOUT (28 schools in Arm B, 30 schools in Arm C, and 28 schools in Arm D = 86 schools in total)
Grade 6 math teachers received a small (payout) incentive contract. The small incentive contract would pay teachers 3500 yuan if they ranked in the top percentile of participating teachers and decline by 35 yuan for each percentile thereafter. Teachers ranking in the bottom percentile would receive 35 yuan.
The diagram below summarizes the design of the randomized experiment:

<table>
<thead>
<tr>
<th>Teacher incentive treatment (outcome-based design feature x payout size design feature)</th>
<th>X. Large incentive payout</th>
<th>Y. Small incentive payout</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Control</td>
<td>A. 52 schools</td>
<td></td>
</tr>
<tr>
<td>B. Levels incentive</td>
<td>BX. 26 schools</td>
<td>BY. 28 schools</td>
</tr>
<tr>
<td>C. Gains incentive</td>
<td>CX. 26 schools</td>
<td>CY. 30 schools</td>
</tr>
<tr>
<td>D. Pay for percentile incentive</td>
<td>DX. 26 schools</td>
<td>DY. 28 schools</td>
</tr>
</tbody>
</table>

Note that the number of schools differ per treatment arm because our randomization was stratified by counties that had varying numbers of schools.

Program (Treatment) Administration

The performance pay contracts (the treatments described above) were given to teachers in September 2013. Teachers first filled out a baseline survey. Teachers then (depending on which treatment arm they were assigned to) received a detailed training on their performance pay contract (for more details, see Details about the Treatment/Interventions below).

Sampling

The experiment was conducted in rural China. Specifically, we randomly sampled 216 schools from 16 nationally-designated “poverty” counties in Yulin Prefecture (Shaanxi Province) and Tianshi Prefecture (Gansu Province) to participate in the experiment. In each school, we randomly sampled one grade 5 mathematics class. All students in the selected grade 5 mathematics classes were tested (in mathematics) and surveyed during two baseline waves. The first baseline wave was at the start of grade 5 (September 2012) and the second baseline wave was at the end of grade 5 (May 2013). The randomized experiment followed up the sampled grade 5 mathematics students into their grade 6 mathematics classes. Altogether, we sampled 8,892 rising grade 6 students and their grade 6 teachers.

Power Calculations

Power calculations were conducted before the beginning of the trial using Optimal Design software (Spybrook et al., 2009). Parameter estimates (to be inputted into the Optimal Design software) were obtained from the two baseline survey waves of grade 5 students during the 2012-2013 school year. Specifically, we used the baseline data to estimate the following parameters for the study:

- Intraclass correlation coefficient (adjusted for county fixed effects): 0.11

---

3 Whereas in most cases the class of grade 5 students in each sample school transferred as an intact group to the grade 6 class, on rare occasions the class of grade 5 mathematics students would be split into multiple grade 6 mathematics classes. In such rare cases, we tracked all of the grade 5 mathematics students into multiple grade 6 classes. We also surveyed and treated the grade 6 teachers for each of the multiple classes.

4 This was done to take into account potential gains from stratification of the randomization procedure by county. See Randomization Procedure below.
• Average number of rising grade 6 students per school: 42
• R-squared of 0.3. This conservative R-squared estimate was obtained from a regression of end-of-grade 5 mathematics test score (in wave 2 of the baseline survey) and wave 1 baseline variables (including the start-of-grade 5 students’ mathematics test scores from wave 1, other wave 1 student and school covariates, and county fixed effects).

As is standard in much of the economics and education literature, we set alpha = 0.05 and beta = 0.8. Based on the above parameters, we estimated that we needed at least 39 schools per treatment arm to detect a minimum detectable effect size (MDES) of 0.2.

The above power calculations do not take into account the following issues (which may slightly reduce our statistical power):

a) We will lose some statistical power due to any student attrition (from the end of grade 5 to the end of grade 6). Using information from past large-scale surveys in rural schools in Shaanxi (including from the prior two waves of baseline data), we assume that the attrition rate will be approximately 5%.

b) Testing multiple hypotheses (comparisons): During the analysis, we will test multiple hypotheses (see the Adjusting for Multiple Hypotheses section below).

Randomization Procedure (including Stratification/Blocking)

We designated each of the 16 counties in our sample as strata or blocks. The general idea behind the randomization procedure was to allocate schools within these counties/strata to one of eight different treatment conditions (BX = levels, big payout incentive; BY = levels, small payout incentive; CX = gains, big payout incentive; CY = gains, small payout incentive; DX = levels, big payout incentive; DY = levels, small payout incentive; AX = no incentives; AY = no incentives). Note that 2 of the 8 conditions (AX and AY) are identical (control group) conditions.

In the event that the number of schools in a stratum was not divisible by 8, remaining schools were randomly allocated to conditions. This allocation was done proportionally to achieve as much balance as possible in the number of schools allocated to each condition. Taken together, the stratification plus randomization procedure ensured that the schools in our sample (the original 216 schools) had an equal probability of being assigned to one of the eight treatment conditions within each county.

---

5 Specifically, we adjusted for the fact that the number of schools in a given county \( N_c \) was not always easily divisible by 8 (i.e. \( N_c \mod 8 \) did not equal 0) as follows: before the randomization step, we added “blank schools” \( b \) within each county such that \( N_c + b \mod 8 = 0 \). After adding in these blank schools, we then randomized the schools to one of the eight treatment conditions. After the randomization step was completed, we dropped the blank schools from the sample.
Baseline Balance

Table A1 (in Appendix A of this document) presents tests for balance on observables across the treatment arms. The table presents the results from a total of 186 tests comparing average variable values across arm pairs (31 variables * the 6 pairs—AB, AC, AD, BC, BD, CD). These tests were conducted by regressing each baseline variable on treatment group indicators and controlling for randomization strata. For tests of student-level variables, standard errors are clustered at the school level.

Out of the 84 tests, 2 (or 2/84 = 2.4%) were statistically significant (different from zero) at the 10% level, 4 (or 4/84 = 4.8%) were statistically significant at the 5% level), and 0 (or 0/84 = 0%) were statistically significant at the 1% level. The results from Table A1 indicate that balance was achieved across the different treatment arms (A, B, C, and D), especially as a small number of significant differences is to be expected (by random chance). Furthermore, none of the more important baseline covariates (in particular, the baseline test scores in wave 1 and wave 2 as well as gains in math scores between wave 1 and wave 2) were statistically different between the treatment arms (even at the 10% level).

We present tests for differences between student and school-level baseline covariates across the Arms X, Y, and A in Table A1 (in Appendix A of this document). Out of 42 total tests comparing average variable values across arm pairs (14 variables * the 3 pairs—XY, XA, YA), 2 (or 2/42= 4.8%) were statistically significant (different from zero) at the 10% level, 0 (or 0/42 = 0%) were statistically significant at the 5% level), and 1 (or 1/42 = 2.4%) was statistically significant at the 1% level. The results from Table A2 indicate that we achieved fairly good balance across the different treatment arms (X, Y, and A), especially as a small number of significant differences is to be expected (by random chance). Furthermore, none of the more important baseline covariates (in particular, the baseline test scores in wave 1 and wave 2 as well as gains in math scores between wave 1 and wave 2) were statistically different between the treatment arms (even at the 10% level). We did not have enough statistical “power” to easily interpret differences between the individual cells of the 4X2 experimental design (e.g. AX vs. BX and so on). We therefore do not present balance tests comparing the distribution of baseline covariates across the individual cells (we regard analyses that compare impacts across the individual cells as exploratory analyses only).

Data Collection

Baseline Surveys

Prior to the beginning of the trial (which took place when sample students were in their sixth grade year), two baseline survey waves collected information on students. The first wave of the student baseline survey was conducted in October 2011 (when sample students were at the start of their 5th grade year) and the second wave was conducted in May 2012 (at the end of sample students’ 5th grade year). During the each wave, the students in our analytical sample were given
a survey to collect detailed information about student and household characteristics (such as age, gender, parental education, parental occupation, family assets, and siblings). Students were also given a 30 minute standardized exam in Math. During each survey wave, we also collected school-level information from school administrators. For example, we collected information on school enrollments, facilities, and distance from the county seat.

A baseline survey of teachers was conducted in September 2013 (at the start of grade 6 for the students in our analytical sample). The survey collected information on teacher background, including information on teacher gender, ethnicity, age, teaching experience, teaching credentials, attitudes toward performance pay, and current performance pay. The teacher survey also included psychometric scales to measure social preferences including prosociality and inequality aversion. We also asked the teacher to indicate which of the sixth grade students he or she was teaching and subjective expectations about each student’s potential achievement gains. The teacher baseline survey took place before we provided the grade 6 teachers with performance pay contracts (in October 2013). Note that control group teachers did not receive a performance pay contract.

**Endline Survey and Primary Outcomes**

We conducted our endline (post-treatment evaluation) survey in May 2014 (at the end of grade 6 for the students in our analytical sample). The endline survey collected detailed information from students, teachers, and school administrators. The information collected from students included student, teacher and household behavioral responses to the teacher incentive pay program (e.g. perceptions of teacher care/effort, teachers ability to manage classroom, teacher practices (towards student), attitudes about math (math anxiety, math self-concept), time spent on math studies each week, class ranking in math, curricula exposure in math, work in other classes outside of math, parent involvement in schoolwork, seat position in class, etc). The information collected from teachers was similar to the information collected during the baseline survey. A principal survey was also used to collect additional information on performance pay policies and attitudes toward performance pay.

The primary outcome variables for the trial are student mathematics achievement at the end of grade 6 (the end of primary school). Mathematics achievement was measured during the endline survey using a 35 minute mathematics test. The mathematics test was constructed by trained psychometricians using a multiple-stage process. In the first stage, a large pool of mathematics test items was selected from the standardized mathematics curricula for grade 6 students in China (and Shaanxi and Gansu provinces in particular). In the second stage, the content validity of these test items was checked by multiple experts (grade 6 mathematics teachers and mathematics teaching professionals). In the third stage, the psychometric properties (reliability, difficulty, differential item functioning, and so on) of each test item were checked using data from a second (extensive) phase of pilot testing. In the fourth stage, a final end-of-grade 6 mathematics test was constructed using the items with “good” psychometric properties. In the fifth and final stage, the
final version of the test was also checked for “good” psychometric properties (unidimensionality, and so on).

In the analyses, we will convert the mathematics test scores from the endline survey into $z$-scores (subtracting each students’ endline mathematics score by the average endline mathematics score in the sample and dividing the standard deviation of the endline mathematics score in the sample).

III. Empirical Approach

General Econometric Framework

Our general approach for estimating treatment effects of the interventions will be to regress outcomes measured at follow-up on dummy variables indicating treatment assignment, baseline controls and strata (county) fixed effects using the following model:

$$Y_{ij} = \alpha_0 + \alpha_1 D_j + X_{ij}\alpha + \tau_c + \epsilon_{ij}$$

where $Y_{ij}$ is the outcome of interest measured at endline for student $i$ in school $j$; $D_j$ is one or more dummies indicating the school treatment assignment of school $j$; $X_{ij}$ is a vector of baseline control variables and $\tau_c$ is a set of county (strata) fixed effects. In all specifications, $X_{ij}$ will include the baseline value of the dependent variable (when available). We will also estimate treatment effects with an expanded set of baseline controls. For student-level outcomes, this expanded set of controls will additionally include student age, student gender, parent educational attainment, a household asset index (constructed using polychoric principal components analysis), class size, teacher experience, and teacher base salary. For outcomes measured at the teacher level, student and household controls will be omitted.

Standard Errors

When the dependent variable is defined at the student, teacher or household level, inference will be conducted clustering the standard errors at the school level using the cluster-corrected Huber-White estimator. For dependent variables measured at the teacher level, standard errors will be computed robust to heteroskedasticity (using the standard formula).

Distributional Analyses

The nature of the intervention and the incentives put in place imply that it is plausible that students at different levels of the achievement (or ability) distribution were more affected than others. To test for impacts on students at different parts of the achievement distribution, we will use two approaches. The first approach will test for the presence of heterogeneous effects by baseline mathematics achievement. The first approach will involve two types of specifications:
(1) Assuming linearity: we will introduce the continuous baseline mathematics attainment variable into the regression together with the interaction of this variable with each of the treatment arms. The baseline mathematics attainment variable will be defined in two ways: a) as a student’s normalized score from the second baseline survey wave and b) as the average of scores from the first and second baseline waves.

(2) Allowing for non-linear effects: we will create 2 new binary variables from the continuous baseline mathematics variables. The first binary variable will take value 1 if the value of the continuous variable is in the top tercile, and 0 otherwise. The second binary variable will take value 1 if the value of the continuous variable is in the middle tercile and 0 otherwise. We will introduce in regression (1) these new binary variables together with each of the interactions between such variable and each treatment arm binary variable (we will also remove the continuous wave 2 baseline mathematics achievement covariate).

In addition, to examining heterogeneous treatment effects by baseline achievement in relation to the rest of the sample (using normalized baseline scores or tercile of the baseline sample distribution), we will also test heterogeneity by student achievement in relation to their class peers. For the continuous case, we will test heterogeneity as described above but replacing the baseline normalized score with the percentile of a student’s score within their class. To allow for non-linear effects, we will use two binary variables indicating whether a student’s score is in the top tercile or middle tercile of the baseline distribution of scores within their class.

The second approach will use a quantile regression specification analogous to equation (1) to estimate treatment effects at quintiles of the endline test score distribution. Note that this will not yield treatment effect estimates at different quantiles of the baseline distribution (since rank order is not preserved). Rather, it will provide a comparison of resulting endline distributions of test scores. We are not aware of analytic methods to compute standard errors that account for clustering using quantile regressions. Hence, inference will be carried out using block-bootstrap (having the school as the block).

Subgroup Analyses and Interaction Effects

Some hypothesis will require to test whether the effect of any of the treatment arms is heterogenous according to some observable characteristic. In case this observable characteristic is binary, we will test for the presence of heterogenous effects by including the corresponding binary variable in regression (1) together with the multiplication of this binary variable with each of the treatment arm binary variables. If the observable characteristic is continuous, we will test for heterogenous effects in two different ways:

(1) Assuming linearity: we will introduce the continuous variable in the regression together with the interaction of this variable with each of the treatment arms.
Allowing for non-linear effects: we will create a binary variable that is equal to 1 if the value of the continuous variable is above the median, and 0 otherwise. We will introduce in regression (1) this new binary variable together with each of the interactions between such variable and each treatment arm binary variable (we will also remove the continuous covariate from the specification).

The specific variables that will be used in subgroup analyses are listed below under “Testing for heterogeneous effects (beyond distributional effects by baseline mathematics achievement).”

Adjusting for Multiple Outcomes

To examine the mechanisms through which teacher incentives do or do not have an effect as well as effects on additional outcomes of interest (see Mechanism Hypotheses section IV.D below), we will test for effects on additional secondary and intermediate outcomes (beyond student achievement). In many cases, multiple, interrelated indicators or survey items were used to measure a given outcome. Therefore to adjust for multiple inference in our analysis of these outcomes, we will use the following procedure: First, we will construct a summary index for each outcome using the indicators or survey items identified with that outcome. The grouping of individual indicators included in the same index (or corresponding to the same hypothesis) is defined in section IV.D below. If there is not an established method of constructing a summary index for a given set of indicators (as is the case for educational or psychological construct scales, for example) we will construct an index using the GLS weighting procedure described in Anderson (2008). If tests using a given index are significant, we will conduct further tests on the individual indicators making up the index to facilitate exploratory analyses. When testing individual indicators we will report the standard p-value for each indicator and, in addition, present inference adjusting for multiple tests (controlling the False Discovery Rate – see Anderson 2008). Tests of heterogeneous treatment effects will each be treated as independent hypotheses (and not adjusted).

Dealing with Attrition/Missing Data

We will address missing data in two ways. First, observations with missing data will be discarded from the analysis (listwise deletions). Second, when missingness is greater than 10%, we will conduct additional analysis using values imputed using multiple imputation techniques (the ice command in STATA, using 8 repetitions).

The degree of differential attrition across trial arms will also be assessed. Students and teachers will be considered to have attritted if they were not present at the endline survey. Note that, as attrition is itself an outcome of the study, information will be collected on individuals who attrit and the reasons for attrition analyzed. If it is found that the rate of attrition differs significantly across trial arms, we will use the Lee trimming method (Lee, 2009) to construct bounds around estimates of treatment effects on primary outcomes.
IV. Specific Analyses/Tests

A. Test the average impacts of different approaches to evaluate teachers for the determination of performance pay (levels, gains, pay for percentile) on student achievement

**Test A1:** Test the impact of any incentive (164 schools) versus no incentive (52 schools).

- Study Arm Comparison(s):
  - Levels, Gains, and Pay for Percentile arms Combined (B, C, and D combined) versus no-incentive arm (A)

- Outcome:
  - Student mathematics achievement (on the end-of-grade 6 endline test)

**Test A2:** Test the impact of the each of the teacher evaluation designs for performance pay (B, C, D—54, 56, 54 schools respectively) against no incentive (52 schools).

- Study Arm Comparison(s):
  - Levels incentive arm (B) versus no-incentive arm (A)
  - Gains incentive arm (C) versus no-incentive arm (A)
  - Pay for percentile incentive arm (C) versus no-incentive arm (A)

- Outcome:
  - Student mathematics achievement (on the end-of-grade 6 endline test)

**Test A3:** Test the impact of the Gains incentive (56 schools) versus the Levels incentive (54 schools).

- Study Arm Comparison(s):
  - Gains arm (C) versus Levels arm (B)

- Outcome:
  - Student mathematics achievement (on the end-of-grade 6 endline test)

**Test A4:** Test the impact of the Pay for percentile incentive (54 schools) versus the Gains incentive (56 schools).

- Study Arm Comparison(s):
• Pay for percentile arm (D) versus Gains arm (C)

Outcome:

• Student mathematics achievement (on the end-of-grade 6 endline test)

B. Test the **distributional impacts** (heterogeneous impacts by baseline student achievement) of different approaches to evaluate teachers for the determination of performance pay (levels, gains, pay for percentile) on student achievement

**Test B1: Test the impact of any incentive (164 schools) versus no incentive (62 schools).**

Study Arm Comparison(s):

• Levels, Gains, and Pay for Percentile arms Combined (B, C, and D combined) versus no-incentive arm (A)

Outcome:

• Student mathematics achievement (on the end-of-grade 6 endline test)

**Test B2: Test the impact of the each of the teacher evaluation designs for performance pay (B, C, D—54, 56, 54 schools respectively) against no incentive pay (52 schools).**

Study Arm Comparison(s):

• Levels incentive arm (B) versus no-incentive arm (A)
• Gains incentive arm (C) versus no-incentive arm (A)
• Pay for percentile incentive arm (C) versus no-incentive arm (A)

Outcome:

• Student mathematics achievement (on the end-of-grade 6 endline test)

**Test B3: Test the impact of the Gains incentive (56 schools) versus the Levels incentive (54 schools).**

Study Arm Comparison(s):

• Gains arm (C) versus Levels arm (B)

Outcome:

• Student mathematics achievement (on the end-of-grade 6 endline test)
**Test B4:** Test the impact of the Pay for percentile incentive (54 schools) versus the Gains incentive (56 schools).

Study Arm Comparison(s):

- Pay for percentile arm (D) versus Gains arm (C)

Outcome:

Student mathematics achievement (on the end-of-grade 6 endline test)

C. Test the average impacts of the “payout size” designs of teacher performance pay (large versus small incentive payouts) on student achievement

**Test C1:** Test for differences in the impacts of large payout incentives (78 schools) versus small payout incentives (86 schools) versus no incentives versus no incentive (52 schools).

Study Arm Comparison(s):

- Large Incentive arm (X) versus Small Incentive arm (Y)
- Large Incentive arm (X) versus no incentive arm (A)
- Small Incentive arm (Y) versus no incentive arm (A)

Outcome:

- Student mathematics achievement (on the end-of-grade 6 endline test)

D. Mechanism Hypotheses: Test the mechanisms by which different approaches to evaluate teachers for the determination of performance pay (levels, gains, pay for percentile) affect student achievement

**Secondary and Mediating Outcomes (from student endline survey (es), teacher baseline (bt), and teacher endline (et) surveys)**

<table>
<thead>
<tr>
<th>Secondary Outcome</th>
<th>Survey items</th>
<th>Operationalization (and additional notes)</th>
</tr>
</thead>
</table>

\(^6\) For outcomes 1-4, we may construction alternative indices in line with that traditionally used in the education literature. This additional index will be scaled using a weighted maximum likelihood estimate (WLE) (Warm 1989),
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Mathematics anxiety</td>
<td>es_61, es_63, es_65, es_68, es_10</td>
<td>Construct index: GLS weighting</td>
</tr>
<tr>
<td>3. Intrinsic motivation for mathematics</td>
<td>es_53, es_55, es_56, es_58</td>
<td>Construct index: GLS weighting</td>
</tr>
<tr>
<td>4. Instrumental motivation for mathematics</td>
<td>es_54, es_57, es_59, es_60</td>
<td>Construct index: GLS weighting</td>
</tr>
<tr>
<td>5. Students’ class rank after one semester of grade 6</td>
<td>es_79</td>
<td>Change the variable so that it represents “reverse rank” (so that the highest number is the first rank and 1 is the lowest rank). Treat the variable as continuous.</td>
</tr>
<tr>
<td>6. Change in teacher’s perception of student’s class rank in mathematics</td>
<td>et_stu_situ_7, bt_stu_situ_7</td>
<td>Convert the endline and baseline variables (separately) into percentile ranks (by comparing students’ endline and baseline ranks with endline and baseline classmates respectively). The outcome variable will equal the endline percentile rank minus the baseline percentile rank.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Note: This analysis only applies to a subset of 12 randomly selected students in each grade 6 class (for which teachers answered these questions for individual students)</td>
</tr>
<tr>
<td>7. Change in teacher’s perception of student’s ability to make gains in mathematics</td>
<td>et_stu_situ_8, bt_stu_situ_8</td>
<td>The outcome variable = et_stu_situ_8 – bt_stu_situ_8.</td>
</tr>
</tbody>
</table>

using a one-parameter item response model (a partial credit model will be used in the case of items with more than two categories).
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>achievement</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Note: This analysis only applies to a subset of 12 randomly selected students in each grade 6 class (for which teachers answered these questions for individual students)</td>
<td></td>
</tr>
<tr>
<td>8. Change in teacher’s perception of how much they could help student make gains in mathematics achievement (through extra tutoring)</td>
<td>et_stu_situ_9, bt_stu_situ_9</td>
<td>The outcome variable = et_stu_situ_9 - bt_stu_situ_9.</td>
</tr>
<tr>
<td></td>
<td>Note: This analysis only applies to a subset of 12 randomly selected students in each grade 6 class (for which teachers answered these questions for individual students)</td>
<td></td>
</tr>
</tbody>
</table>
| 9. Student’s taught easy, middle, advanced mathematics curricula by grade 6 teacher | es_82 thru es_90 inclusive (each are dummy variables taking values of 1 = taught, 0 = not taught) | a) The average of es_82, es_83, es_84 represents degree to which teacher taught easy curricula  
|   |   | b) The average of es_85, es_86, es_87 represents degree to which teacher taught medium level of difficulty curricula  
|   |   | c) The average of es_88, es_89, es_90 represents degree to which teacher taught advanced curricula |
| 10. Student time spent on math | es_71 and es_72, es_73, es_76, es_77 | Construct an index using GLS weighting procedure  
<p>|   |   | When analyzing indicators individually, treat individual variables as ordinal category outcome variables (run an |</p>
<table>
<thead>
<tr>
<th>11. Student perception of teacher practices</th>
<th>es_31 through es_44 inclusive</th>
<th>Construct an index using GLS weighting procedure. When analyzing indicators individually, construct individual dummy variables corresponding to e_31 thru e_44. Specifically, for each variable split the ordinal values into two categories so as to minimize number of responses between each category.</th>
</tr>
</thead>
<tbody>
<tr>
<td>12. Changes in seating arrangement: Sitting in the front (y/n) Change in seating from the start of the year</td>
<td>es_23 (first two rows, y/n) es_25 (change to front or back, y/n)</td>
<td>No index</td>
</tr>
<tr>
<td>13. Student perception of teacher care/effort</td>
<td>es_26, es_27, es_28, es_29, es_30</td>
<td>Construct an index using GLS weighting procedure. When analyzing indicators individually, construct individual dummy variables corresponding to e_26 thru e_30. Specifically, for each variable split the ordinal values into two categories so as to minimize number of responses between each category.</td>
</tr>
<tr>
<td>14. Student perception of teachers’ ability to manage classroom</td>
<td>es_45, es_46, es_47, es_48</td>
<td>Construct an index using GLS weighting procedure. When analyzing indicators individually, construct individual dummy variables</td>
</tr>
</tbody>
</table>
15. Student perception of student-teacher communication

<table>
<thead>
<tr>
<th>Corresponding variables</th>
<th>Construct an index using GLS weighting procedure when analyzing indicators individually, construct individual dummy variables corresponding to e_49 thru e_52. Specifically, for each variable split the ordinal values into two categories so as to minimize number of responses between each category.</th>
</tr>
</thead>
<tbody>
<tr>
<td>es_49, es_50, es_51, es_52</td>
<td></td>
</tr>
</tbody>
</table>

16. Substitution of student time away from non-math subjects

<table>
<thead>
<tr>
<th>Variables</th>
<th>Construct an index using GLS weighting procedure when analyzing indicators individually, construct individual dummy variables corresponding to e_91 thru e_99. Specifically, for each variable split the ordinal values into two categories so as to minimize number of responses between each category.</th>
</tr>
</thead>
<tbody>
<tr>
<td>es_91 thru es_99</td>
<td></td>
</tr>
</tbody>
</table>

17. Parents involved in schoolwork

<table>
<thead>
<tr>
<th>Variables</th>
<th>Construct an index using GLS weighting procedure when analyzing indicators individually, construct individual dummy variables corresponding to e_19 thru e_22. Specifically, for each variable split the ordinal values into two categories so as to minimize number of responses between each category.</th>
</tr>
</thead>
<tbody>
<tr>
<td>es_19, es_20, es_21, es_22</td>
<td></td>
</tr>
</tbody>
</table>

18. Teacher understanding of incentive pay contract and

<table>
<thead>
<tr>
<th>Variables</th>
<th>Number answered correctly (out of 3 questions). Analyze using ordered probit model</th>
</tr>
</thead>
<tbody>
<tr>
<td>et_section9_70-71, et_section9_72</td>
<td></td>
</tr>
<tr>
<td>19. Salience of contract</td>
<td>et_section_9_69, et_section9_73</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>20. Teacher effort</td>
<td>et_section5_38-45</td>
</tr>
<tr>
<td>21. Teacher attention to test preparation</td>
<td>et_section6_49-50</td>
</tr>
<tr>
<td>22. Teacher opinion of performance pay</td>
<td>et_section3_question14-25</td>
</tr>
</tbody>
</table>

**For each of the above secondary outcomes, test the following for average impacts:**

**Test 1: Test the impact of any incentive (164 schools) versus no incentive (52 schools).**

Study Arm Comparison(s):

- Levels, Gains, and Pay for Percentile arms Combined (B, C, and D combined) versus no-incentive arm (A)

Outcome:

- Each of the five secondary outcomes (on the end-of-grade 6 survey)

**Test 2: Test the impact of the each of the teacher evaluation designs for performance pay (B, C, D—54, 56, 54 schools respectively) against no incentive (52 schools).**

Study Arm Comparison(s):

- Levels incentive arm (B) versus no-incentive arm (A)
- Gains incentive arm (C) versus no-incentive arm (A)
- Pay for percentile incentive arm (C) versus no-incentive arm (A)

Outcome:

- Each of the five secondary outcomes (on the end-of-grade 6 survey)
**Test 3:** Test the impact of the Gains incentive (56 schools) versus the Levels incentive (54 schools).

Study Arm Comparison(s):

- Gains arm (C) versus Levels arm (B)

Outcome:

- Each of the five secondary outcomes (on the end-of-grade 6 survey)

**Test 4:** Test the impact of the Pay for percentile incentive (54 schools) versus the Gains incentive (56 schools).

Study Arm Comparison(s):

- Pay for percentile arm (D) versus Gains arm (C)

Outcome:

- Each of the five secondary outcomes (on the end-of-grade 6 survey)

**Test 5:** Test for differences in the impacts of large payout incentives (78 schools) versus small payout incentives (86 schools) versus no incentives versus no incentive (52 schools).

Study Arm Comparison(s):

- Large Incentive arm (X) versus Small Incentive arm (Y)
- Large Incentive arm (X) versus no incentive arm (A)
- Small Incentive arm (Y) versus no incentive arm (A)

Outcome:

- Each of the five secondary outcomes (on the end-of-grade 6 survey)

For each of the above secondary outcomes, test the following for distributional impacts (by achievement – see Distributional Analyses above):

**Test 1:** Test the impact of any incentive (164 schools) versus no incentive (52 schools).

Study Arm Comparison(s):
• Levels, Gains, and Pay for Percentile arms Combined (B, C, and D combined) versus no-incentive arm (A)

Outcome:

• Each of the five secondary outcomes (on the end-of-grade 6 survey)

**Test 2: Test the impact of the each of the teacher evaluation designs for performance pay (B, C, D—54, 56, 54 schools respectively) against no incentive (52 schools).**

Study Arm Comparison(s):

• Levels incentive arm (B) versus no-incentive arm (A)
• Gains incentive arm (C) versus no-incentive arm (A)
• Pay for percentile incentive arm (C) versus no-incentive arm (A)

Outcome:

• Each of the five secondary outcomes (on the end-of-grade 6 survey)

**Test 3: Test the impact of the Gains incentive (56 schools) versus the Levels incentive (54 schools).**

Study Arm Comparison(s):

• Gains arm (C) versus Levels arm (B)

Outcome:

• Each of the five secondary outcomes (on the end-of-grade 6 survey)

**Test 4: Test the impact of the Pay for percentile incentive (54 schools) versus the Gains incentive (56 schools).**

Study Arm Comparison(s):

• Pay for percentile arm (D) versus Gains arm (C)

Outcome:

• Each of the five secondary outcomes (on the end-of-grade 6 survey)

**Test 5: Test for differences in the impacts of large payout incentives (78 schools) versus small payout incentives (86 schools) versus no incentives versus no incentive (52 schools).**
Study Arm Comparison(s):

- Large Incentive arm (X) versus Small Incentive arm (Y)
- Large Incentive arm (X) versus no incentive arm (A)
- Small Incentive arm (Y) versus no incentive arm (A)

Outcome:

- Each of the five secondary outcomes (on the end-of-grade 6 survey)
### E. ADDITIONAL HETEROGENEOUS EFFECTS ANALYSES (FOR LOOKING AT DISTRIBUTIONAL EFFECTS):

<table>
<thead>
<tr>
<th></th>
<th>Question</th>
<th>Code</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Teacher’s preferred ability distribution (inequality aversion)</td>
<td>bt_section12_2_1</td>
<td>Incentivized Bayesian Truth Serum question</td>
</tr>
<tr>
<td>2.</td>
<td>Teacher’s reported effort on the top 1/3, middle 1/3, bottom 1/3 of achievers</td>
<td>bt_section12_1_1a-1c</td>
<td>Incentivized Bayesian Truth Serum question</td>
</tr>
<tr>
<td>3.</td>
<td>Teacher experience</td>
<td>bt_section2_question 30</td>
<td>in years</td>
</tr>
<tr>
<td>4.</td>
<td>Teacher risk aversion</td>
<td>bt_section6_questions77-79</td>
<td>Dummy variables will be defined for high risk aversion (the top 2 of 4 categories).</td>
</tr>
<tr>
<td>5.</td>
<td>Teacher baseline intrinsic motivation</td>
<td>bt_section10_questions103, 105, 107, 109</td>
<td>Construct an index using GLS weighting procedure</td>
</tr>
<tr>
<td>6.</td>
<td>Teacher baseline prosocial motivation</td>
<td>bt_section10_questions102, 104, 106, 108</td>
<td>Construct an index using GLS weighting procedure</td>
</tr>
<tr>
<td>7.</td>
<td>Teacher baseline perception of ability to improve student scores (on average)</td>
<td>bt_stu_situ_9</td>
<td>Average for all students</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>Note</strong>: This analysis only applies to a subset of 12 randomly selected students in each grade 6 class (for which teachers answered these questions for individual students)</td>
</tr>
<tr>
<td>8.</td>
<td>Teacher baseline perception of ability to improve student scores (for bottom, middle, top students)</td>
<td>bt_stu_situ_9</td>
<td>Average for top, middle, bottom students separately</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>Note</strong>: This analysis only applies to a subset of 12 randomly selected students in each grade 6 class (for which teachers answered these questions for individual students)</td>
</tr>
</tbody>
</table>
### Teacher experience with performance pay

**Variables:**
- bt_section2_question35b, bt_section2_question35
- bt_section2_question35b / bt_section2_question35 = % of total salary that is performance pay (per month in the previous semester)

### Teacher base salary

**Variables:**
- bt_section2_question35
- in yuan/month

### Teacher opinion of performance pay at baseline

**Variables:**
- bt_section3_question43-51
- Construct an index using GLS weighting procedure
- Question 51 will be defined based on proportion assigned to test scores

---

**Test E1: Test the heterogeneous impacts of any incentive (164 schools) versus no incentive (52 schools).**

Study Arm Comparison(s):
- Levels, Gains, and Pay for Percentile arms Combined (B, C, and D combined) versus no-incentive arm (A)

Outcome:
- Student mathematics achievement (on the end-of-grade 6 endline test)

**Test E2: Test the heterogeneous impacts of the each of the teacher evaluation designs for performance pay (B, C, D—54, 56, 54 schools respectively) against no incentive (52 schools).**

Study Arm Comparison(s):
- Levels incentive arm (B) versus no-incentive arm (A)
- Gains incentive arm (C) versus no-incentive arm (A)
- Pay for percentile incentive arm (C) versus no-incentive arm (A)

Outcome:
- Student mathematics achievement (on the end-of-grade 6 endline test)
**Test E3: Test the heterogeneous impacts of the Gains incentive (56 schools) versus the Levels incentive (54 schools).**

Study Arm Comparison(s):

- Gains arm (C) versus Levels arm (B)

Outcome:

- Student mathematics achievement (on the end-of-grade 6 endline test)

**Test E4: Test the heterogeneous impacts of the Pay for percentile incentive (54 schools) versus the Gains incentive (56 schools).**

Study Arm Comparison(s):

- Pay for percentile arm (D) versus Gains arm (C)

Outcome:

- Student mathematics achievement (on the end-of-grade 6 endline test)
References


Appendix A: Balance Tests across Treatment Arms

Table A1: Testing for Balance in Baseline Covariates across Treatment Arms: A (control), B (levels), C (gains), and D (pay for percentile)

<table>
<thead>
<tr>
<th></th>
<th>control_levels</th>
<th>control_gains</th>
<th>control_p4p</th>
<th>levels_gains</th>
<th>levels_p4p</th>
<th>gains_p4p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wave 2 math</td>
<td>0.045</td>
<td>-0.047</td>
<td>0.024</td>
<td>-0.104</td>
<td>0.013</td>
<td>0.110</td>
</tr>
<tr>
<td>Gains math</td>
<td>0.012</td>
<td>-0.005</td>
<td>-0.029</td>
<td>-0.054</td>
<td>-0.009</td>
<td>0.024</td>
</tr>
<tr>
<td>Wave 1 math</td>
<td>0.032</td>
<td>-0.041</td>
<td>0.056</td>
<td>-0.046</td>
<td>0.023</td>
<td>0.084</td>
</tr>
<tr>
<td>Age</td>
<td>-0.083</td>
<td>-0.138***</td>
<td>-0.086</td>
<td>-0.054</td>
<td>0.010</td>
<td>0.077</td>
</tr>
<tr>
<td>Dad’s educ</td>
<td>-0.006</td>
<td>-0.028</td>
<td>0.002</td>
<td>-0.027</td>
<td>0.003</td>
<td>0.020</td>
</tr>
<tr>
<td>Female</td>
<td>0.008</td>
<td>-0.005</td>
<td>-0.004</td>
<td>-0.013</td>
<td>-0.009</td>
<td>0.011</td>
</tr>
<tr>
<td>Mom’s educ</td>
<td>-0.007</td>
<td>-0.019</td>
<td>-0.020</td>
<td>-0.021</td>
<td>-0.004</td>
<td>0.011</td>
</tr>
<tr>
<td>Asset index</td>
<td>-0.017</td>
<td>-0.026</td>
<td>-0.039</td>
<td>0.000</td>
<td>-0.036</td>
<td>-0.030</td>
</tr>
<tr>
<td>Class size</td>
<td>2.233</td>
<td>-2.980</td>
<td>2.138</td>
<td>-5.080**</td>
<td>0.158</td>
<td>5.444**</td>
</tr>
<tr>
<td>Num classes</td>
<td>1.190</td>
<td>1.001</td>
<td>0.610</td>
<td>0.565</td>
<td>-0.229</td>
<td>-0.495</td>
</tr>
<tr>
<td>Num contract teachers</td>
<td>-0.421</td>
<td>-0.128</td>
<td>-0.007</td>
<td>0.199</td>
<td>0.174</td>
<td>0.069</td>
</tr>
<tr>
<td>Num teacher</td>
<td>2.400</td>
<td>3.395</td>
<td>2.167</td>
<td>3.279</td>
<td>-0.261</td>
<td>-1.545</td>
</tr>
<tr>
<td>Teacher exp</td>
<td>-2.965*</td>
<td>-2.546*</td>
<td>-4.333**</td>
<td>-0.308</td>
<td>-1.383</td>
<td>-0.828</td>
</tr>
<tr>
<td>Teacher salary</td>
<td>2677.406</td>
<td>2941.015</td>
<td>3374.101</td>
<td>826.579</td>
<td>4079.425</td>
<td>2857.886</td>
</tr>
</tbody>
</table>
Table A2: Balance across Treatment Arms X (big incentive payout), Y (small incentive payout), and A (control)

<table>
<thead>
<tr>
<th></th>
<th>control_small</th>
<th>control_big</th>
<th>small_big</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wave 2 math</td>
<td>-0.015</td>
<td>0.012</td>
<td>0.014</td>
</tr>
<tr>
<td>Gains math</td>
<td>-0.020</td>
<td>-0.011</td>
<td>0.020</td>
</tr>
<tr>
<td>Wave 1 math</td>
<td>0.006</td>
<td>0.024</td>
<td>-0.008</td>
</tr>
<tr>
<td>Age</td>
<td>-0.111*</td>
<td>-0.097</td>
<td>0.010</td>
</tr>
<tr>
<td>Dad’s educ</td>
<td>-0.003</td>
<td>-0.021</td>
<td>-0.014</td>
</tr>
<tr>
<td>Female</td>
<td>0.002</td>
<td>-0.004</td>
<td>0.002</td>
</tr>
<tr>
<td>Mom’s educ</td>
<td>-0.017</td>
<td>-0.010</td>
<td>0.015</td>
</tr>
<tr>
<td>Poverty level</td>
<td>0.001</td>
<td>-0.048</td>
<td>-0.054</td>
</tr>
<tr>
<td>Class size</td>
<td>1.924</td>
<td>-0.910</td>
<td>-2.951</td>
</tr>
<tr>
<td>Num classes</td>
<td>1.421</td>
<td>0.512</td>
<td>-0.474</td>
</tr>
<tr>
<td>Num contract teachers</td>
<td>-0.163</td>
<td>-0.147</td>
<td>0.115</td>
</tr>
<tr>
<td>Num teacher</td>
<td>5.343</td>
<td>0.263</td>
<td>-3.745</td>
</tr>
<tr>
<td>Teacher exp</td>
<td>-4.128***</td>
<td>-2.552*</td>
<td>2.270</td>
</tr>
<tr>
<td>Teacher salary</td>
<td>1096.025</td>
<td>5116.916</td>
<td>4676.637</td>
</tr>
</tbody>
</table>