Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement

Prashant Loyalka, Sean Sylvia, Chengfang Liu, James Chu, Yaojiang Shi*

June 11, 2015

PRELIMINARY AND INCOMPLETE

ABSTRACT: There is growing interest in strengthening teacher incentives by tying pay to performance measures based on student achievement. Yet, there is little empirical evidence on how teachers may respond to specific design features of performance pay schemes. Theoretically appealing but relatively complex schemes may not outperform less appealing but simple schemes in practice. In this paper, we present the results of a randomized trial designed to test alternative approaches of mapping student achievement into rewards for teachers. Teachers in western China were randomly assigned to participate in rank-order tournaments in which teacher rankings were determined as a function of their students' scores on standardized exams by one of three different methods of defining teacher performance. We find that teachers offered *pay-for-percentile* incentives (based on the scheme described in Barlevy and Neal 2012) outperform teachers offered two simpler schemes based on year-end class average achievement *levels* or average *gains* over the course of a school year. Achievement gains under pay-for-percentile were mirrored by meaningful changes in the intensity of teaching. Moreover, we find that pay-for-percentile incentives lead to broad based gains, improving outcomes for students across the achievement distribution within the class. Our finding that teachers respond to a relatively intricate feature of an incentive scheme highlights the importance of close attention to performance pay design.

Keywords: Teacher Performance Pay, Distributional Effects, China
JEL Codes: I24, O15, J33, M52

* Loyalka, Stanford University; Sylvia, Renmin University of China; Liu, Chinese Academy of Sciences; Chu, Stanford University; Shi, Shaanxi Normal University
Corresponding Author: Sean Sylvia; Address: School of Economics, Renmin University of China, No. 59 Zhongguancun Ave., Beijing, China 200872; Email: ssylvia@ruc.edu.cn

Pay by Design: Teacher Performance Pay Design and the Distribution of Student Achievement

Teachers often work in environments where they face incentives that are weak or misaligned with improving student outcomes (Lazear, 2003). Teacher salaries, for instance, are often most closely related to teacher attributes such as education and experience, which tend not to be strongly associated with student achievement (Hanushek and Rivkin, 2010; Podgursky and Springer, 2007, Rivkin, Hanushek, and Kain, 2005). Possibly due to a lack of explicit incentives to improve student outcomes, teacher absenteeism is pervasive in many parts of the world (Banerjee and Duflo, 2006; Chaudhury et al., 2006; Kremer et al., 2005) and teachers often fail to teach effectively when present (Chaudhury et al., 2006; Staiger and Rockoff, 2010). In response, a growing movement seeks to better align teacher incentives by linking teacher pay more directly with student achievement and performance pay programs are increasingly implemented in both developed and developing countries (Hanushek and Woessmann, 2011; Bruns et al., 2011; Woessmann, 2011; OECD, 2009).

Whether performance pay schemes can improve student outcomes, however, may depend critically on their design (Neal, 2011; Bruns et al., 2011). Schemes in which rewards are not closely linked to teacher effort may fail to provide strong incentives. For example, performance pay schemes that are involve performance cutoffs have been shown to provide weak incentives for teachers starting away from those cutoffs (Neal and Schanzenbach, 2010; Lazear, 2006). Moreover, schemes employing performance measures that are misaligned with ultimate student outcomes of interest may lead to strategic behavior and other unintended consequences (Holmstrom and Milgrom 1991; Baker 1992; Dixit 2002).

While studies have highlighted particular weaknesses in particular design features of performance pay schemes, many important aspects of design have yet to be explored empirically. Few empirical studies directly compare the effects of alternative design features of performance pay schemes.[1] Although theoretically appealing (and often more

[1] Important exceptions include Muralidharan and Sundararaman (2011) who compare individual and group incentives for teachers in India and Fryer et al. (2012) who compare both incentives designed to exploit loss aversion and a more traditional incentive scheme as well as individual and group incentives. Behrman et al. 2015 present an experiment in Mexico comparing incentives for teachers to incentives for students

complex) designs exist that are meant to address common failures, there is little evidence to suggest whether these outperform less appealing but simpler schemes in practice.

In this paper, we study incentive design directly by comparing performance pay schemes that vary in how student achievement is mapped onto teacher rewards. Specifically, we test alternative ways of using student achievement scores from standardized exams to measure and reward teacher performance. How student scores are used to measure teacher performance and mapped onto rewards may affect the strength of incentive schemes and lead teachers to devote more or less effort toward improving student outcomes (Bruns et al., 2011; Neal, 2011; Neal and Schanzenbach, 2010). A particular challenge is in devising a single measure of performance using the achievement scores of individual students in a teacher's class (Neal 2011, Barlevy and Neal 2012). How measures of achievement for individual students are combined into an index of teacher performance in the determination of rewards may – in addition to overall incentive strength – affect how teachers choose to allocate effort and attention across students in the classroom by explicitly or implicitly weighting some students in the class more than others (a version of the multitasking problem – Holmstrom and Milgrom 1991).

We compare alternative performance pay designs through a large-scale randomized trial in western China. Math teachers in 216 primary schools were randomly placed into a control group or one of three different rank-order tournaments that varied in how the achievement scores of individual students were combined into an index of teacher performance and used to rank and reward teachers. Teachers in half of the schools in each of these treatment groups were then randomly allocated to a small incentive treatment or a large incentive treatment (where rewards were twice as large, but remain within policy-relevant levels).

We present three main findings. First, we find that teachers offered a *pay-for-percentile* ("pay-for-percentile") scheme (based on the scheme described in Barlevy and Neal (2012)) outperformed teachers offered two more simple schemes based on class average achievement levels ("levels") at the end of the school year or class average achievement gains ("gains") from the start to the end of the school year. On average, the pay-for-percentile scheme led to a 0.13 to 0.15 standard deviation (SD) increase in

---

and joint incentives for students, teachers and school administrators.

student achievement relative to the control group, while levels and gains incentives led to no significant improvements on average. Achievement gains under pay-for-percentile were mirrored by meaningful changes in the intensity of teaching as evidenced by students experiencing greater curricular coverage, being taught more advanced curricula, and being more likely to correctly answer difficult exam items.

Second, while levels and gains incentives had little effect on students at any part of the within class distribution of student achievement, we find that pay-for-percentile led to broad-based impacts across students within a class. Specifically, we find that the pay-for-percentile scheme resulted in significant and comparable gains for students across the within-class distribution of baseline achievement scores. This is in line with how the pay-for-percentile scheme more symmetrically rewards student gains across students within a class.

Third, we find suggestive evidence of complementarity between the strength of incentive design and reward size. Specifically, while the effects of levels and gains remain small and insignificant when potential rewards are increased, the estimated impact of pay-for-percentile incentives is notably larger with larger rewards (and is in fact insignificant with smaller rewards). This suggests that, in our context, both a strong incentive design and sufficiently large rewards may be required to produce meaningful gains for students.

Our study makes several contributions to the literature. Most directly, we contribute to a growing literature on the effectiveness of teacher performance pay. Overall, results from previous well-identified studies have been decidedly mixed. Several studies have found teacher performance pay to be effective at improving student achievement, particularly in developing countries where hidden action problems can be pernicious (Glewwe et al. 2010, Duflo and Hanna 2005, Lavy 2002, Lavy 2009, Muralidharan and Sundararaman, 2011, Fryer et al. 2012; Dee and Wyckoff, 2015).[2,3] For instance, impressive evidence comes from a large-scale experiment in India which found large and long-lasting effects of teacher performance pay tied to student achievement on

[2] Glewwe et al. (2010) finds that teacher incentives in Kenya led to improvements in student achievement after 2 years, but that these effects faded after three years.

[3] In a follow-up to his 2009 study, Lavy (2015) shows that a teacher performance pay program in Israel affected long run student outcomes including college attendance and earnings 15 years after the original program.

math and language scores (Muralidharan and Sundararaman, 2011; Muralidharan, 2011). In contrast, other recent studies in developed and developing countries have found that teacher performance pay does not have significant effects on student achievement (Behrman et al. 2015; Springer et al. 2010; Fryer, 2013).

Beyond providing more evidence on the effectiveness performance pay, we contribute to this literature in three ways. Our primary contribution to this literature is by directly comparing alternative methods of measuring and rewarding teacher performance as a function of student achievement. In addition to context, previous studies of teacher performance pay vary widely in the overall design of incentive schemes and in how these schemes measure teacher performance in particular.[4] Only two previous studies directly test design features of incentive schemes. Muralidharan and Sundararaman (2011) compare group and individual incentives and find that individual incentives are more effective after the first year. Fryer et al. (2012) compare incentives designed to exploit loss aversion with more traditional incentives and find loss aversion incentives to be substantially more effective. Fryer et al. (2012) also compare individual and group incentives and find no significant differences. Our results in this paper highlight that how the achievement scores of individual students are combined into an index of teacher performance matters—independent of other design features. Second, this study is also the first of which we are aware to experimentally compare varying sizes of monetary rewards for teachers (adding to two recent studies testing incentive reward size in health delivery – Ashraf, Bandiera and Jack (2014) and Luo et al. (2015)).[5] Third, by directly studying how incentive design affects students with different levels of achievement within a class,

[4] Muralidharan and Sundararaman (2011) study a piece rate scheme tied to average gains in student achievement. The scheme studied in Behrman et al. (2015) rewarded and penalized teachers based on the progression (or regression) of their students (individually) through proficiency levels. The scheme studied in Springer et al. (2010) rewarded math teachers bonuses if their students performed in the 80th percentile, 90th percentile or 95th percentile. Fryer (2013) studies a scheme in New York City that paid schools a reward, per union staff member, if they met performance targets set by the Department of Education and based on school report card scores. Lavy (2009) studies a rank order tournament among teachers with fixed rewards of several levels. Teachers were ranked based on how many students passed the matriculation exam, as well as the average scores of their students. In Glewwe, Ilias and Kremer (2010) bonuses were awarded to schools for either being the top scoring school or for showing the most improvement. Bonuses were divided equally among all teachers in a school who were working with grades 4-8.

[5] Both of these studies of incentives in health delivery (Ashraf, Bandiera and Jack (2014) and Luo et al. (2015)) compare small rewards with substantially larger ones. Ashraf, Bandiera and Jack (2014) compare small rewards with large rewards that are approximately nine times greater and Luo et al. (2015) compare small rewards with larger rewards that are ten times greater. Here, we compare small rewards with larger rewards that are only two times greater.

we add to evidence from two recent studies (Neal and Schanzenbach (2010) and Duflo, Dupas and Kremer (2011) that show how incentives can affect teacher instructional focus and allocation of effort across students within a class.

Our work also contributes to literatures outside of education. In general, our results add to a growing number of studies that use field experiments to evaluate performance incentives in organizations (Bardach et al., 2013; Bandiera et al., 2007, 2005; Cadsby et al., 2007). We also contribute to the literature on tournaments, in particular by testing the effects of different size rewards. Although there is evidence from the lab (see Freeman and Gelber 2010), we are aware of no field experiments that have tested the effect of varying tournament prize structure. Finally, we show that teachers can respond to relatively complex features of reward schemes. Growing evidence (generally from tax structures) suggests that individuals have difficulty reacting optimally to complex incentives (Saez 2010; Abeler and Jäger 2013). While we cannot say if teachers respond optimally to the incentives they were given, we find that they do respond more to pay-for-percentile incentives than gains or levels incentives, which are arguably more simple schemes. Inasmuch as our results show that teachers respond to relatively intricate features of incentive contracts, they suggest considerable room for these features to affect welfare (for better or worse) and highlight the importance of close attention to incentive design.

The rest of the paper is organized as follows. Section 2 presents our experimental design and data. We share our results in Section 3. Section 4 discusses the results and concludes.

2. Experimental Design & Data

2.1. School Sample

Our study sample is located in two prefectures in western China. The first prefecture is located in Shaanxi Province (ranked 16 out of 31 in terms of GDP per capita in China), and the second is located in Gansu Province (ranked 27 out of 31—NBS, 2014). Within these two prefectures, we included 16 nationally-designated poverty counties in our sample. Within each of these counties, we conducted a canvass survey to

construct a list of all rural elementary schools meeting our inclusion criteria.[6] We then applied two exclusion criteria to this list. We then randomly selected 216 schools for inclusion in the study.

2.2. Randomization and Stratification

We designed our study as a cluster-randomized trial using a partial cross-cutting design (Table 1). The 216 schools included in the study were first randomized into a control group (52 schools) and three incentive design groups: a "levels" incentive group (Group B – 54 schools), a "gains" incentive group (Group C – 56 schools), and a "pay-for-percentile" group (Group D – 54 schools).[7] Across these three incentive treatments, we orthogonally assigned schools to incentive size groups: a "small" incentive group (Group X – 78 schools) and a "large" incentive group (Group Y – 86 schools). All sixth grade math teachers in a school were assigned to the same treatment.

To improve power, we used a stratified randomization procedure. Specifically, we stratified the randomization procedure by county (yielding 16 total strata). Our analysis takes this randomization procedure into account. In particular, we condition on stratum fixed effects (Bruhn and McKenzie 2009).

2.3. Incentive Design and Treatments

2.3.1 Common Rank-Order Tournament Structure

While the incentive design treatments vary in how teacher performance is defined in the determination of rewards, all incentive treatments have a common underlying rank-order tournament structure. When informed of their incentive, teachers were told that they would compete with sixth grade math teachers in other schools (both within and outside their prefecture). The competition would be based on their students' performance

[6] We applied three exclusion criteria to our sampling frame. First, because our substantive interest is in poor areas of rural China, we excluded elementary schools located in urban areas (the county seats). Second, when rural Chinese elementary schools serve areas with low enrollment, they may close higher grades (5^{th} and 6^{th} grade) and send eligible students to neighboring schools. We excluded these "incomplete" elementary schools. Third, we excluded elementary schools that had enrollments smaller than 120 (i.e. enrolling an average of fewer than 20 students per grade). Because the prefecture departments of education informed us that these schools would likely be merged or closed down in following years, we decided to exclude these schools from our sample.

[7] Note that the number of schools across treatments is unequal due to the number of schools available per county (strata) not being evenly divisible.

on common standardized math exams.[8] According to their percentile ranking among other teachers in the program, each teacher would be given a cash reward (transferred to their bank account) within two months after the end of the school year. Teachers were not told the total number of teachers who would be competing in the tournament.

Rewards were structured to be linear in percentile rank as:

$$Bonus = R - (99 - PercentileRank) \times b$$

where R is the reward for teachers ranking in the top percentile and b is the incremental reward for each percentile rank. In the small incentive treatment (Group X), teachers ranking in the top percentile received 3500 *yuan* ($547) and the incremental reward per percentile rank was 35 *yuan*.[9] In the large incentive treatment (Group Y), teachers ranking in the top percentile received 7000 *yuan* ($1,094) and the incremental reward per percentile rank was 70 *yuan*. These reward amounts were calibrated so that the top reward was equal to approximately one month salary in the small incentive treatment and was equal to two months salary in the large incentive treatment.[10]

Note that this structure departs from more traditional tournament schemes which typically have a less differentiated reward structure. Specifically, tournament schemes more often have fewer reward levels and only reward top performers (the tournament studied in Lavy (2009) for example has only four reward levels). By setting rewards to be linearly increasing in percentile rank, the underlying reward structure that we use in this study is similar to the incentive scheme studied in Knoeber and Thurman (1994).[11] We

[8] Only 11 schools in our sample had multiple sixth grade math teachers. When there was more than one sixth grade math teacher, teachers were ranked together and were explicitly told that they would not be competing with one another.

[9] Rewards were structured such that all teachers received some reward. Teachers ranking in the bottom percentile received 70 yuan in the large incentive treatment and 35 yuan in the small incentive treatment.

[10] While there was no explicit penalty if students were absent on testing dates, contracts stated that teachers would be disqualified if there was evidence that students were purposefully kept from sitting exams. In practice, teachers also had little or no warning of the exact testing date at the end of the school year. We find no evidence that lower achieving students were less likely to sit exams at the end of the year.

[11] Knoeber and Thurman (1994) also study a similar "linear relative performance evaluation" (LRPE) scheme that, instead of rewarding percentile rank, bases rewards on a cardinal distance from mean output. Bandiera et al. (2005) compare an LRPE scheme with piece rates in a study of fruit pickers in the UK.

chose to use this linear structure in order to minimize distortions in incentive strength due to non-linearities.[12]

Relative rewards schemes such as rank-order tournaments have a number of potential advantages over piece rate schemes. First, tournaments provide budget certainty as teachers compete for a fixed pool of money (Neal 2011; Lavy 2009). This may make this sort of system more attractive to policymakers. Neal (2011) notes that tournaments may also be less subject to political pressures that seek to flatten rewards. For risk-averse agents, tournaments are also more robust to common shocks (across all participants).[13] Teachers may also be more likely to trust the outcome of a tournament that places them in clear relative position to their peers rather than that of a piece-rate scheme which places teacher performance on an externally derived scale based on student test scores (teachers may doubt that the scaling of the tests leads to consistent teacher ratings, for example—Briggs and Weeks 2009). On the other hand, tournaments may be subject to dynamic gaming when rankings are determined by a gain measure (Macartney 2014; Heinrich and Marschke 2010). This may be less of a concern in our case, however, given the linear relationship in our schemes between teacher ranks and rewards.[14] Tournaments may also be less efficient when teachers are of heterogeneous ability and the tournament is not seeded based on ability (Lazear and Rosen, 1981).

2.3.2 Defining Teacher Performance

Our primary interest is in evaluating designs that use alternative ways of defining teacher performance as a function of student achievement. Specifically, we vary how achievement scores of individual students in each teacher's class are combined into an

[12] Tournament theory suggests a tradeoff between the size of reward increments between reward levels (which increase the monetary size of rewards) and weakened incentives for individuals far enough away from these cutoffs. In a recent lab experiment, Freeman and Gelber (2010) find that a tournament with multiple, differentiated prizes led to greater effort than a tournament with a single prize for top performers, holding total prize money constant.

[13] Although it is difficult to say whether common or idiosyncratic shocks are more or less important in the long-run, one reason we chose to use rank order tournaments over a piece rate schemes based on student scores is that relative reward schemes would likely be more effective if teachers were uncertain about the difficulty of exams (one type of common shock).

[14] Bandiera et al. (2005) find that piece rate incentives outperform relative incentives in a study of fruit pickers in the UK. Their findings suggest, however, that this is due to workers' desire to not impose externalities on co-workers under the relative scheme by performing better. This mechanism may be less important in our setting as competition was purposefully designed to be between teachers across different schools.

index of teacher performance. The index of teacher performance is subsequently used to rank teachers in the tournament. The three designs that we evaluate are as follows:

Levels Incentive (Group B): In the "levels" incentive treatment, teacher performance was defined as the class average of student scores on a standardized exam at the end of the school year. Thus, teachers are ranked in the tournament and rewarded based on this year-end class average achievement. Evaluating teachers based on *levels* of student achievement (average student exam performance at a given point in time) is common in China and other developing countries (Murnane and Ganimian 2014).

Gains Incentive (Group C): Teacher performance in the "gains" incentive treatment was defined as the class average gain in student achievement from the start to the end of the school year. Individual student achievement gains were measured as the difference in their score on a standardized exam administered at the end of the school year minus that student's performance on a similar exam at the end of the previous school year.

Compared to levels incentives, gains incentives may be more effective if teachers perceive the change in student achievement scores to be more closely related to their own effort. In a tournament context, ranking teachers according to average gains (rather than levels) in student achievement can help (in part) to address weakened incentives in classes with low or high baseline levels of achievement relative to reward cutoffs. Although this should be less of a factor given the linearity of our underlying reward structure, teacher decisions themselves may be nonlinear in payoffs (teachers may only incorporate these incentives into their decision making if perceived rewards a large enough, for example). On the other hand, rewarding gains incentives may provide weaker incentives if teachers recognize that gains measurements are more subject to statistical noise (Murnane and Ganimian 2014).

An issue with both levels and gains incentives is that, by averaging scores across students in a class, they implicitly weight some students in the class more than others. Because rewards are a function of these averages, teachers seeking to maximize rewards will optimally allocate effort across students in the class according to expected marginal returns and costs of effort (where the margin returns to effort are in terms of achievement gains on a given assessment scale). Teachers may, for instance, neglect students scoring

at the top of an assessment scale because further achievement gains would not be measured (or rewarded)—there is no room for improvement within the measured scale.

Pay-for-Percentile Incentives (Group D): The third way of defining teacher performance that we examine is through a "pay-for-percentile" approach based on the method described in Barlevy and Neal (2012). In this treatment, teacher performance was calculated as follows. First, all students were placed in comparison groups based on their score on the baseline exam conducted at the end of the previous school year. Within each of these comparison groups students were then ranked by their score on the endline exam and assigned a percentile score, equivalent to the fraction of students in a student's comparison group whose score was lower than that student. A teacher's performance index (percentile performance index) was then determined by the average percentile rank taken over all students in his or her class. This percentile performance index can be interpreted as the fraction of contests that students of a given teacher win when compared to students who are taught by other teachers and yet began the school year at similar achievement levels (Barlevy and Neal 2012).

The distinguishing feature of this pay-for-percentile scheme is that it avoids the need to compute and reward an aggregate statistical measure of total classroom performance based on individual student achievement scores (Barlevy and Neal 2012). Teachers essentially compete in as many contests as there are students in her class. Because these contests are symmetric and ordinal (independent of assessment scale), rewarding teachers according to the percentile performance index largely avoids implicit weighting that can occur when teacher performance is defined as an aggregate statistical measure (such as an average) of individual student scores. Under the right (theoretical) conditions, the pay for percentile scheme can strengthen incentives for teachers to focus instruction and attention more broadly across students within a classroom.[15]

Compared to levels and gains incentives, there are at least two reasons why rewarding teacher performance based on pay-for-percentile may create stronger incentives on average. First, pay-for-percentile incentives better account for the composition of students taught by teachers in determining their performance relative to

[15] Note that, our pay-for-percentile treatment does not control for all the factors proposed by Barlevy and Neal (2012). We chose to control for baseline student achievement and not other baseline factors (such as family background) to increase the transparency of the incentive scheme for the treated teachers.

peers. In a relative reward scheme, this can create a closer relationship between rewards and teacher effort. In particular, compared to gains, pay-for-percentile adjusts for not only the classroom average baseline achievement, but also the distribution of baseline achievement (and potentially other characteristics) within a class. A second possibility is that pay-for-percentile, since it rewards effort focused on any student, gives teachers more flexibility to allocate effort where the returns to their effort are highest in terms of generating human capital for their students. They can teach to their comparative advantage rather than absolute gains on an assessment scale.

Pay-for-percentile, however, could be less effective to the extent that it also rewards teachers for more broadly focused effort across students and could penalize teachers for focusing on any subset of students. Given the symmetry of individual student contests, teachers are penalized for neglecting any student. Depending on how student human capital is produced as a function of teacher effort (specifically, to what degree individual student outcomes are jointly produced by teacher effort), however, this feature could lead teachers to spread effort across students to the degree that student gains are limited on average.

Pay-for-percentile incentives could also fail to outperform levels and gains incentives in practice due to their perceived relative complexity and less transparency. A growing body of research, mainly from tax incentives, suggests that people have difficulty responding optimally to complex incentives (Saez 2010; Abler and Jäger 2013). If pay-for-percentile contracts are perceived as complex and monetary incentives are not worth teacher effort required to figure out an optimal response and incorporate this into their teaching practice, pay-for-percentile incentives may be ineffective.

2.3.3 Implementation

Following an initial survey (described below), teachers in all incentive arms were presented performance pay contracts stipulating details of their assigned incentive scheme. These contracts were signed and stamped by the Chinese Academy of Sciences and were presented with officials from the local bureaus of education. Before signing the contract, teachers were provided with materials explaining the details of the contract and how rewards would be calculated. To better ensure that teachers understood the incentive

structure and contract terms, they were also given a training session lasting approximately 2 hours covering the same material. A short quiz was also given to teachers to check and correct misunderstanding of the contract terms and reward determination.

2.4. Data Collection

Our data collection efforts entailed several survey rounds and focused on students in the sixth grade during the 2013/2014 school year. First, we conducted two baseline survey waves in the 216 schools included in the study, one at the beginning (September) and one at the end (May) of the 2012/2013 school year (when the children were in fifth grade). These surveys collected detailed information on student, teacher and school characteristics. Students were also administered standardized exams in math. The use of two baseline surveys affords us with additional precision in controlling for prior achievement, as well as a measure for how much students were learning prior to our experiment. At the beginning of the 2013/2014 school year, we conducted a detailed survey of all sixth grade math teachers. A follow-up survey collecting information on students, teachers and schools was conducted in May 2014, at the end of the 2013/2014 school year.

Student Surveys. Surveys were administered to students in September 2012, May 2013 and May 2014 (at the beginning and end of their fifth grade year and at the end of their sixth grade year). During each wave, surveys asked for basic student and household characteristics (such as age, gender, parental education, parental occupation, family assets, and number of siblings). During the endline survey, students were also asked detailed questions covering their attitudes about math (anxiety, self-concept, intrinsic and instrumental motivation scales); math curricula that teachers covered with students during the school year; time spent on math studies each week; perceptions of teacher teaching practices, teacher care, teacher management of the classroom, teacher communication; parent involvement in schoolwork; and time spent on subjects outside of math.

Teacher Surveys. We conducted a baseline survey of all sixth grade mathematics teachers (who taught our sample students) in September 2013. 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 psychological scales to measure social preferences including prosocial motivation 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 teachers with performance pay contracts (in October 2013). In May 2014, we surveyed the teachers again using the same questions as in the baseline survey.

Standardized Math Exams. Our primary outcome for the trial is student mathematics achievement scores. Math achievement was measured during the endline survey using a 35 minute mathematics test. The mathematics test was constructed by trained psychometricians. Mathematics test items were first selected from the standardized mathematics curricula for primary school students in China (and Shaanxi and Gansu provinces in particular) and the content validity of these test items was checked by multiple experts. The psychometric properties (unidimensionality, reliability, difficulty, differential item functioning, and so on) of the test were then checked using data from extensive pilot testing. In the analyses, we normalize mathematics exam scores using the mean and distribution in the control group. Estimated effects are therefore expressed in standard deviations.

2.5. Balance and Attrition

Summary statistics and tests for balance across study arms are shown in Appendix Table 1. Panel A shows student-level characteristics, Panel B shows teacher and class characteristics and Panel C shows school level characteristics. The first column gives the mean in the control group. Columns 2-4 and 6-7 show coefficients on treatment variables estimated using Equation (1) with the baseline covariate at left as the dependent variable and controlling only for randomization strata (county) dummies. Columns (5) and (8) show p-values from a test that coefficients are jointly zero for each statistic. Only six of the eighty coefficients estimated are significant at 10% or less and only one test of joint equality is rejected at 10%.[16]

[16] Note that teacher level characteristics in this table differ from those in our pre-analysis plan, which used

The overall attrition rate between September 2013 and May 2014 (beginning and end of the school year of the intervention) was 6.4% in our sample. Defining attrition as missing mathematics scores at the endline for students with a baseline measurement, Appendix Table 2 shows that there were no meaningful differences in attrition across treatment groups in the full sample (Columns 1 & 2). We do find that attrition was significantly less for pay-for-percentile within the small incentive groups, but this magnitude is small (Column 3, Row 6).[17]

2.6. Empirical Strategy

Given the random assignment of schools to treatment cells as shown in Table 1, comparisons of outcome variable means across treatment groups provide unbiased estimates of the effect of each experimental treatment. However, to increase power (and to account for our stratified randomization procedure), we condition our estimates on strata (county) dummy variables and also present results adjusted for additional covariates. With few exceptions, all of the analyses presented (including outcome variables, regression specifications, and hypotheses tested) were pre-specified in a pre-analysis plan written and filed before endline data were available for analysis.[18] In reporting results below, we explicitly note analyses that deviate from the pre-analysis plan.

As specified in advance, we use ordinary least-squares (OLS) regression to estimate the effect of teacher incentive treatments on student outcomes with the following specification:

$$Y_{ijc} = \alpha + T_i'\beta + X_{ijc}'\gamma + \tau_c + \varepsilon_{ijc} \quad (1)$$

where Y_{ijc} is the outcome for child i in school j in county c; T_j is a vector of dummy variables indicating the treatment assignment of school j; X_{ijc} is a vector of control variables and τ_c is a set of county (strata) fixed effects. In all specifications, X_{ijc} includes baseline student scores. We also estimate treatment effects with an expanded set of

---

teacher characteristics from the previous year. The characteristics used here are for teachers who were present in the baseline and thus part of the experiment. This balance table also uses the post-attrition sample.

[17] This turns out to be inconsequential for our results.

[18] This analysis plan was filed with the American Economic Association RCT Registry at https://www.socialscienceregistry.org/trials/411.

controls. For student-level outcomes, this includes student age, student gender, parent educational attainment, a household asset index (constructed using polychoric principal components), class size, teacher experience, and teacher base salary. We adjusted our standard errors for clustering at the school level by using the cluster-corrected Huber-White estimator.

In addition to estimating effects on our primary outcome (year-end standardized exam scores normalized by the control group distribution), we use the same specification to estimate effects on secondary outcomes to examine the mechanisms underlying changes in exam scores. For these secondary outcomes, we focus our analysis on summary indices constructed using groups of closely-related outcome variables (as we specified in advance). To construct these indices, we used the GLS weighting procedure described by Anderson (2008). For each individual, we constructed a variable $\bar{s}_{ij}$ as the weighted average of k normalized outcome variables in group (y_{ijk}). The weight placed on each outcome variable is the sum of its row entries in the inverted covariance matrix for group j such that:

$$\bar{s}_{ij} = \left(\mathbf{1}'\widehat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{1}\right)^{-1}\left(\mathbf{1}'\widehat{\boldsymbol{\Sigma}}_j^{-1}\boldsymbol{y}_{ij}\right)$$

where $\mathbf{1}$ is a column vector of 1s, $\widehat{\boldsymbol{\Sigma}}_j^{-1}$ is the inverted covariance matrix, and $\boldsymbol{y}_{ij}$ is a column vector of all outcomes for individual i in group j. Because each outcome is normalized (by subtracting the mean and dividing by the standard deviation in the sample), the summary index, $\bar{s}_{ij}$, is in standard deviation units. In addition to reducing the number of tests required, this weighting procedure can improve efficiency by placing less weight on outcomes that are highly correlated and more weight on those less correlated. The summary index variable can also be created for individuals with a subset of missing outcomes (these outcomes simply receive less weight in the construction of the index).

3. Results

In this section, we present four sets of results. First, we present results on the average impacts of incentives designs on student achievement (Section 3.1). Second, we present results on the average impacts of incentives on student secondary outcomes and teacher behavior (Section 3.2). Third, we present results on the distributional impacts of

incentives on achievement, focusing especially on incentives that use different ways of defining teacher performance (Section 3.3). Finally, we show how results differ by teacher and school characteristics, focusing on those characteristics that theoretically may be important for the incentive designs to be effective (Section 3.4).

3.1 Average Impacts of Incentives on Achievement

The first six rows (Panel A) of Table 2 report estimates for the different incentive treatments (any incentive, those based on different teacher performance indices, and those based on different reward size). As specified in our pre-analysis plan, we report estimates using Equation (1) and two different sets of controls: a limited set of controls (controlling only for two waves of baseline standardized math exam scores and strata fixed effects) as well as estimates from regressions that include an expanded set of controls (additionally controlling for student gender, age, parental educational attainment, a household asset index, class size, teacher experience and teacher base salary). Panel B of Table 2 reports estimated differences in impacts between different treatments (with corresponding p-values).

Any incentive. We find weak evidence that having *any incentive* (pooling all incentive treatments) modestly increases student achievement at the endline. The specification including the expanded set of controls shows that having *any incentive* increases student achievement by 0.074 SDs (Table 2, Panel A, Row 1, Column 2). The result is statistically significant at the 10% level.

Teacher performance measures. Although the effect of teachers having *any incentive* is modest, the effect of incentives depends on different ways of defining teacher performance. Most notably, we find that *pay-for-percentile* incentives have a significant and meaningful effect on student achievement. We estimate that *pay-for-percentile* incentives raise student scores by 0.128 SDs (in the basic specification) to 0.148 SDs (in the specification with additional controls—Table 2, Panel A, Row 2, Columns 3 and 4). Both estimates are statistically significant at the 5% level. By contrast, we find no significant effects from offering teachers *levels* or *gains* incentives (Table 2, Panel A, Rows 3-4, Columns 3-4). Furthermore, when we compare the estimates across the treatments, we find that *pay-for-percentile* significantly outperforms *gains* (0.147 SDs, p-

value=0.023 with the expanded control set). The point estimate for pay-for-percentile is also larger than that for *levels* though not significant (0.064 SDs, p-value=0.292).

The effect of *pay-for-percentile* relative to *levels* and *gains* can be more clearly seen in Figure 1A, which plots the cumulative distributions of endline achievement by incentive type. Figure 1A shows that distribution of student achievement in the *pay-for-percentile* arm first-order stochastically dominates that for the *levels*, *gains*, and *control* arms. Notably, the distribution for the *pay-for-percentile* group is shifted to the right of the distribution for the *levels* group (Kolmogorov-Smirov test p-value=.073).

This result that *pay-for-percentile* outperforms *gains* incentives and (marginally) *levels* incentives shows that, even with the same underlying (rank tournament) incentive structure, the way the teacher performance index is defined matters. Moreover, this effect comes at no or little added cost as monitoring costs (to collect underlying assessment data) and the total amount of rewards paid was held constant. Given that *gains* and *levels* are arguably much simpler schemes this result also hints that, at least in our context, teachers respond to relatively complex features of incentive schemes.

Small Rewards, Large Rewards and the Complementarity of Reward Size with Teacher Performance Measurement. When pooling across the teacher performance index treatments, the difference between large and small incentives is small and insignificant. Both are also insignificantly different from zero (Table 2, Columns 4 & 5). However, when estimating the effects of the performance index treatments separately in the small and large reward group, pay-for-percentile incentives seem to be more effective (and are only significant) when teachers are offered larger rewards (0.165 SD larger, Table 2, Panel B, Rows 4 and 13-14, Columns 7-10).[19] The larger difference between pay-for-percentile and other treatment groups with large rewards can also be seen in Figure 1B, which plots the cumulative distribution of endline scores by performance index arm using large reward schools only. While we cannot reject the hypothesis that the effect of *pay-for-percentile* with small rewards is the same as the effect of the *pay-for-percentile* with larger rewards at conventional levels, there may be a non-trivial difference in effect sizes that we are unable to detect due to a lack of statistical power (we did not power the study

[19] Note that the study was not powered to test the interaction between the teacher performance index treatments and incentive size. Testing this interaction was therefore not pre-specified.

ex-ante to test interactions across treatments). Given that the effects of levels and gains incentives are unaffected by increasing reward size but the effect of pay-for-percentile is greater with larger rewards implies that there may be complementarity between the strength of design and size of rewards. In our context, both seem necessary to generate significant improvements in student achievement.

3.2. Impacts of Incentives on Teacher Behavior and Secondary Student Outcomes

We next examine the effects of incentives on secondary student outcomes and teacher behavior, as these effects may explain the changes in endline achievement that we describe in Section 3.1. To estimate the effects, we run regressions analogous to equation 1, but substitute endline achievement for secondary student outcome and teacher behavior variables. Outcomes representing "curricular coverage" were measured by asking students whether they had been exposed to specific examples of curricula material in class during the school year. Students were given three such examples of curricula material from the last semester of grade five ("easy" material), three from the first semester of grade 6 ("medium" material) and three from the second semester of grade 6 ("hard material). Students' binary responses were averaged for the easy, medium, and hard categories. Most of the other secondary outcomes (math self-concept, math anxiety, math intrinsic and instrumental motivation, student perception of teacher practice, teacher cares, teacher can manage, teacher communication) are indices that were created from a family of outcome variables using the GLS weighting procedure described in Anderson (2008). The indices are standardized so that they have a mean of 0 and a SD of 1.

We find significant impacts of teacher incentives on teaching practice. In particular, students with teachers that receive *any incentive* report being covered more curricula content at the easy level (significant at the 10% level) and medium levels (significant at the 5% level) and the same amount at the hard level (not statistically significant at the 10% level) compared to students in the control group (Table 3, Panel A, Row 1, Columns 5-7). The magnitude and significance of the effects appear to hold regardless of the size of the incentive (although the effect of being taught more curricula at the easy level is not statistically significant for the large incentive group—Table 3, Panel B, Rows 2-3, Columns 5-7). However, the effects vary across the teacher

performance index treatments. In particular, while students in the *levels* group also report being taught more curricula at the easy and medium levels (see Table 3, Panel B, Row 4, Columns 5-7) and students in the *gains* group report being taught more curricula at the medium level, students in the *pay-for-percentile* group report being taught more medium (significant at the 5% level) and hard (significant at the 1% level) curricula.

These impacts on curricular coverage suggest that teachers covered more of the curriculum, however this could come at the expense of reduced intensity of instruction. Teachers could respond to incentives by teaching at a faster pace in order to cover as much of the curriculum as possible. To test this, we estimate treatment effects on subsets of test items categorized into easy, medium and hard questions (Table 4).[20] Test items were categorized into easy, medium and hard questions (10 items each) using the frequency of correct responses in the control group. Similar to student's responses, we find that pay-for-percentile led to sizeable gains in all three categories, particularly when incentives were large. Pooling across small and large incentives, pay-for-percentile increased the easy question sub-score by 0.204 SD, the medium question sub-score by 0.211 SD (this is marginally insignificant), and the hard question sub-score by 0.335 SD. Here, we also see more consistent impacts of pay-for-percentile when incentives are large (Table 4, Columns 7-9). Importantly, these results show that pay-for-percentile incentives increased both the coverage and intensity of instruction.

Despite these effects on curricular coverage and intensity, we find little effect on other types of teacher behavior (Table 3, Columns 9-14). There are no statistically significant impacts from any of the incentive arms on teachers' classroom engagement, care, classroom management, or communication as reported by student and no significant effect on self-reported teacher effort. The finding of little impact on these dimensions of teacher behavior in the classroom is similar to results in Glewwe et al. (2010) and Muralidharan and Sundararaman (2011) who find little impact of incentives on classroom processes. These studies, however, do find changes in teacher behavior outside of the classroom. While we do find impacts of all types of incentives on student-reported times

[20] Note that analysis of test items was not pre-specified in our analysis plan. This analysis should therefore be considered exploratory.

being tutored outside of class,[21] these do not explain the significantly larger impact of pay-for-percentile. In our case, it seems that incentives worked largely through curricular coverage and instructional intensity.

We also find little evidence that incentives of any kind affect students' secondary learning outcomes. Effects on indices representing math self-concept, math anxiety, instrumental motivation in math, and student time spent on math are all insignificant (Table 3, Columns 1, 2, 4, 8). Importantly we also find little evidence that any type of incentives led to increased math test preparation or substitution away from non-math subjects (results omitted for the sake of brevity).

3.3. Effects on the Distribution of Student Achievement

Table 5 reports distributional effects by baseline achievement for the three incentives which use different ways of defining teacher performance. The first two rows (Panel A) present estimates of effects along the baseline distribution of student achievement in the full sample, while second two rows (Panel B) of Table 5 present effects along the baseline achievement within the classroom ("class rank"). The first row in each panel proxies baseline achievement using the second wave baseline exam score only, whereas the second row calculates baseline achievement by averaging the first and second wave baseline exam scores. In accordance with the pre-analysis plan, we report estimates both without assuming linearity (impacts by tercile of the baseline distribution) and estimates of the linear interaction between treatments and baseline achievement. We estimate effects by tercile of the baseline distribution by estimating Equation (1) but including dummy variables for the second and third terciles and interactions with indicators for the *levels*, *gains*, and *pay-for-percentile* incentive arms. Linear interactions with baseline achievement are estimated analogously, but instead including the baseline score and interactions with treatment arm indicators directly. All regressions control for strata (county) fixed effects, student gender, age, parental educational attainment, a household asset index, class size, teacher experience and teacher base salary.

[21] We do not individually report these results as they are part of an index upon which the effect of incentives was insignificant and thus are not pre-specified.

When we examine distributional effects across baseline achievement for the full sample (Table 5, Panel A), we find that the *pay-for-percentile* incentives have the largest effect on students at the middle of the distribution (increasing scores by 0.187 or 0.228 SDs depending on the baseline distribution used – Table 5, Panel A, Rows 3-4, Column 12). The *levels* and *gains* incentives, however, have little effects at any point along the full sample baseline distribution.

The effects of pay-for-percentile, however, are more broad-based when looking within classrooms (Panel B, Columns 11-13). Pay-for-percentile has significant effects for students within each tercile of baseline achievement regardless of how baseline achievement is measured. Point estimates are slightly larger for the bottom tercile (0.172 SDs to 0.188 SDs) than the middle and top terciles (0.12 to 0.14 SDs). Apart from some indication of a positive effect of levels incentives on the bottom tercile of the within class distribution (0.113 SD - Table 5, Panel B, Row 4, Column 1), levels and gains incentives had no significant effects for any tercile of the within class distribution. Broad based gains attributable to pay-for-percentile within classrooms suggests that larger effects of pay-for-percentile in the middle of the distribution for the full sample are due to heterogeneous effects across classes. We explore this further below.

3.4. Heterogeneous Effects by Teacher and Class Characteristics

We focus our analysis of heterogeneous effects on teacher and class characteristics that may in theory affect how teachers respond to tournament-based incentives generally and, within that structure, to the alternative ways of defining teacher performance. To estimate heterogeneous effects on student achievement scores, we run regressions analogous to Equation (1), but add interactions between incentive arms (levels, gains, pay-for-percentile) and teacher and class characteristics measured at baseline. For baseline characteristics that are continuous, we create dummy variables indicating whether that characteristic is above the median of the distribution in the sample. The results are reported in Table 6. Each column in Table 6 shows coefficients on treatment indicators and interactions with baseline covariates listed at the top of the column (across the top row).

Basic Teacher Characteristics. There is little evidence of differential impacts by teacher experience, teacher base salary, or teacher gender for the *levels*, *gains*, and *pay-for-percentile* incentives (Table 6, Rows 2, 4, 6, Columns 1-3).

Teacher Social Preferences. Although we find little evidence of heterogeneity by basic teacher characteristics, impacts of pay-for-percentile incentives do vary substantially by elements of teacher's social preferences as measured at baseline. First and strikingly, the impact of pay-for-percentile among teachers with higher than median pro-social motivation (motivation seeded in a desire to help others, measured using psychological scales based on Grant (2008)) is 0.295 SDs larger (significant at 1%, Table 6, Row 6, Column 5). The interaction between pay-for-percentile and intrinsic motivation (motivation seeded in pleasure of doing the job itself, also measured using psychological scales based on Grant (2008)) is also positive, but insignificant (Table 6, Column 4). This result contrasts with concerns, particularly in public service contexts, that external monetary incentives may dampen the effects of internal motivation, at least for pay-for-percentile incentives (see Frey and Jegen (2001) and Kamenica (2012) for reviews). [22]

As to why *pay-for-percentile* incentives are substantially more effective among more pro-socially motivated teachers compared to *levels* or *gains*, we speculate that *pay-for-percentile* incentives may be perceived as more supportive and less controlling, in which case incentives can actually crowd-in motivation (Frey and Jegen 2001). Pay-for-percentile incentives may be viewed as less controlling, for example, because teachers feel that – given the way *pay-for-percentile* incentives reward effort focused across students in the class – they may have more flexibility to decide upon their allocation of effort and instructional focus.

In addition to pro-social motivation, we also find that pay-for-percentile incentives were substantially less effective among teachers that value equity over

[22] We also note that – inasmuch as prosocial motivation reflects other-regarding preferences more generally – this finding suggests that the effect of incentives was not reduced by teachers internalizing externalities imposed on other teachers as a result of the relative nature of the underlying rank order tournament structure of the incentives (Bandiera, Barankay and Rasul 2005). Bandiera, Barankay and Rasul (2005) find that relative incentives underperform piece rates and evidence attributing this underperformance to workers' internalizing externalities that are imposed on other workers under relative incentives. They, however, also find that this is only the case when workers and monitor other workers and be monitored. Given that our tournaments intentionally did not place teachers in the same school in competition with one another our finding is not inconsistent with their context.

efficiency (Table 6, Column 7). Specifically, as part of the baseline survey, teachers were presented with an incentivized Bayesian Truth Serum (BTS) question (Prelec 2004) that asked:

> "Suppose the following are distributions of scores for your students at the end of the 2013-2014 school year (on a 100 point exam). Which distribution, A, B, or C, would you prefer?"

Grade Distribution	Average for TOP 1/3 of Students	Average for MIDDLE 1/3 of Students	Average for BOTTOM 1/3 of Students
A	99 Points	72 Points	33 Points
B	94 Points	68 Points	42 Points
C	78 Points	64 Points	51 Points

The options (A, B, and C) progressively alter the hypothetical score distribution in two ways: a) the class average achievement of the class declines and b) the distribution becomes more equitable (dispersion decreases). We find that pay-for-percentile incentives were 0.248 SD *less* effective among teachers who chose option C at baseline. Given how this question pits teacher preferences for efficiency against equity, that teachers choosing options A and B were more responsive to pay-for-percentile than those choosing C may simply reflect that teachers preferring efficiency in student outcomes also have the propensity to make more efficiency-minded decisions in their work. A second, possibility is that teachers who are inequality averse already tend to avoid focusing their effort on some students in his or her class at the expense of others.

Teacher Self-Perceived Value Added. The importance of distributional considerations in explaining the effectiveness of *pay-for-percentile* incentives is also supported by the fact that teachers with higher "self-perceived value added", particularly for children at the low end of the class distribution, are more responsive to pay-for-percentile (Table 6, Row 6, Columns 8-11). We elicited teacher self-perceived value added in the baseline survey by presenting teachers with a list of twelve students from their current class (chosen randomly, stratified by terciles of the baseline exam score) and

asking teachers to guess each student's score on a year-end standardized math exam based on the national curriculum, once without any mention of other factors and then again assuming that the teacher gave that student one additional hour of out-of-class tutoring per week for the entire school year. Column 8 uses the average value added to an hour of tutoring (as calculated by the difference between the two exam scores) over all twelve students on the list. Columns 9, 10 and 11 use the average of the bottom, middle and top 3rd of students on this list, respectively, as ranked by ability as reported by the teacher.

We find that the average impact of *pay-for-percentile* – on all students – was substantially larger among teachers whose self-perceived value added was above the median in the sample with respect to students in the bottom tercile of the class (0.314 SDs, significant at the 1% level—Table 6, Row 6, Column 9) and the middle tercile of the class (0.255 SDs, significant at the 5% level—Table 6, Row 6, Column 10). The average impacts of *pay-for-percentile* are also larger, but insignificantly so, among teachers that have a high self-perceived value added for students in the top tercile (Table 6, Row 6, Column 11). There is little evidence of heterogeneity along these dimensions for *levels* and *gains*. These results are consistent with the hypothesis that, because pay-for-percentile rewards teacher effort across the class distribution (as opposed to levels and gains that place more weight on students with a higher value added in absolute terms) teachers could better teach to their comparative advantage. An alternative hypothesis is that teachers who believe their own value added is high generally (relative to other factors affecting student outcomes) are more responsive to sufficiently well-designed incentives because they believe that rewards more closely reflect their own effort (and variation in teacher perceptions of value added is larger for relatively underperforming students).

Class and Grade Characteristics. In the distributional effects in discussed in Section 3.3, we found that pay-for-percentile had a larger effect in the middle of the distribution for the full sample, but had comparable effects across the within-class distribution, suggesting a large amount of variation in effects across classes. We see this confirmed here as we test for heterogeneity by average classroom achievement at baseline (Table 6, Columns 12-14). We find the impact of pay-for-percentile is

substantially larger among teachers with class averages that are in the middle tercile of the full sample (0.324 SDs, significant at the 1% level—Table 6, Row 6, Column 13). Although point estimates for gains and levels follow a similar pattern, none of these are significant (Table 6, Rows 2 and 4, Column 13). This result could be related to the underlying tournament structure if factors other than teacher effort have (or are perceived by teachers to have) a relatively larger influence on student outcomes at the ends of the distribution.

Finally, we find that the effects of both gains incentives and pay-for-percentile are larger in schools where the size of the sixth grade class is above the median (Table 6, Column 15). Differences in effects by class size may be expected because a larger class may entail more teacher effort. As noted above, a subset of schools have more than two sixth grade teachers which could be correlated with grade size. However, there is not evidence that effects vary with the number of sixth grade math teachers per school (Table 6, Column 16).

4. Discussion & Conclusion

This paper provides evidence on the relative effectiveness of different designs of teacher performance pay. Specifically, we test alternative ways of using student achievement scores to measure teacher performance in the determination of rewards as well as how the effects of incentives vary with reward size. We highlight three key findings. First, we find that pay-for-percentile incentives, based on the scheme described in Barlevy and Neal (2012), led to larger gains in student achievement than two alternative schemes that rewarded teachers based on class-average student achievement on a year-end exam and the class-average gains in student achievement over the school year. Pay-for-percentile incentives, but not the other two designs, increased both the coverage and intensity of classroom instruction. Second, in line with the design of the pay-for-percentile scheme – which rewards teachers for effort devoted to each student in the class – we find broad-based gains across the distribution of students in the class. Third, we do not find a significant difference between small and large incentives on average pooling across designs, but do find suggestive evidence of complementarity between incentive design strength and reward size. As noted by Bruns et al. (2011),

complementarity between incentive design strength and reward size has important implications for cost effectiveness. In our context, we find that both a stronger design and relatively large (but policy relevant) rewards are required to produce meaningful gains in student achievement.

With our results we offer a number of caveats. First, we only study the effects of incentives over one year. It is likely that impacts will change as teachers become accustomed to incentive schemes. Note, however, that most multi-year studies of teacher incentives have shown larger effects after the first year. Second, our study was not powered to ex-ante to study the interaction between different ways of measuring teacher performance and incentive size. Although we find suggestive evidence, future studies explicitly powered to test the complementarity between incentive design strength and reward size will be valuable. Finally, as with all randomized trials, results will not necessarily hold other contexts or if incentive schemes are implemented on a very large scale. A particular consideration for teacher incentives that we do not consider, for instance, is how incentive schemes may affect how individuals select into the teaching profession.

Despite these caveats, we believe that these results clearly illustrate that incentive design matters. Moreover, teachers in our context respond to a relatively intricate design feature. This suggests the need for further research to identify the features of incentive design that matter in practice as well as how design features interact. It also suggests more generally that there is substantial scope for the design of incentives to affect welfare.

References

Abeler, J., Jäger, S., 2013. Complex Tax Incentives-An Experimental Investigation. CESifo Working Paper (No. 4231).

Ashraf, N., Bandiera, O., Jack, B.K., 2013. No Margin, No Mission? A Field Experiment on Incentives for Pro-Social Tasks. Working paper.

Baker, G.P., 1992. Incentive Contracts and Performance Measurement. *Journal of Political Economy,* 100, 598–614.

Bandiera, O., Barankay, I. and Rasul, I. 2007. Incentives for Managers and Inequality Among Workers: Evidence From a Firm Level Experiment. *Quarterly Journal of Economics*, 122, 729–775.

Bandiera, O., Barankay, I. and Rasul, I. 2005. Social Preferences and the Response to Incentives: Evidence From Personnel Data. *Quarterly Journal of Economics*, 120, 917–962.

Banerjee, A., Duflo, E., 2006. Addressing Absence. *The Journal of Economic Perspectives,* 20, 117–132.

Bardach, N. S., Wang, J. J., De Leon, S. F., Shih, S. C., Boscardin, W. J., Goldman, L. E., & Dudley, R. A. 2013. Effect of pay-for-performance incentives on quality of care in small practices with electronic health records: a randomized trial. *JAMA*, *310*(10), 1051-1059.

Barlevy, G. & Neal, D. 2012. Pay for percentile. *American Economic Review*, 102(5), 1805-31.

Behrman, J.R., Parker, S.W., Todd, Petra E., Wolpin, K.I., 2015. Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools. *Journal of Political Economy* 123, 325–364.

Briggs, D. C., & Weeks, J. P. 2009. The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy*, *4*(4), 384-414.

Bruhn, M., McKenzie, D., 2009. In Pursuit of Balance: Randomization in Practice in Development Field Experiments. American Economic Journal: Applied Economics 1, 200–232.

Bruns, B., Filmer, D., Patrinos, H.A., 2011. Making Schools Work: New Evidence on Accountability Reforms. The World Bank.

Cadsby, C.B., Song, F., & Tapon, F. 2007. Sorting and incentive effects of pay-for performance: An experimental investigation. *Academy of Management Journal*, 50, 387–405.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., & Rogers, F. H. 2006. Missing in action: teacher and health worker absence in developing countries. *The Journal of Economic Perspectives*, *20*(1), 91-116.

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*(2), 267-297.

Dixit, A., 2002. Incentives and Organizations in the Public Sector: An Interpretative Review. *The Journal of Human Resources*, 37, 696–727.

Duflo, E., Hanna, R., 2005. Monitoring Works: Getting Teachers to Come to School. National Bureau of Economic Research.

Duflo, E., Dupas, P., Kremer, M., 2011. Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. *American Economic Review*, 101, 1739–1774.

Freeman, R.B., Gelber, A.M., 2010. Prize Structure and Information in Tournaments: Experimental Evidence. *American Economic Journal: Applied Economics*, 2, 149–164.

Frey, B.S., Jegen, R., 2001. Motivation Crowding Theory. *Journal of Economic Surveys*, 15, 589–611.

Fryer, R. G. 2013. Teacher Incentives and Student Achievement: Evidence from New York City Public Schools. *Journal of Labor Economics*, *31*(2), 373-407.

Fryer Jr, R. G., Levitt, S. D., List, J., & Sadoff, S. 2012. Enhancing the efficacy of teacher incentives through loss aversion: A field experiment (No. w18237). National Bureau of Economic Research.

Glewwe, P., Ilias, N., & Kremer, M. 2010. Teacher Incentives. *American Economic Journal: Applied Economics*, 205-227.

Gneezy, U., Leonard, K.L., List, J.A., 2009. Gender Differences in Competition: Evidence From a Matrilineal and a Patriarchal Society. *Econometrica*, 77, 1637–1664.

Grant, A.M., 2008. Does intrinsic motivation fuel the prosocial fire? Motivational synergy in predicting persistence, performance, and productivity. Journal of *Applied Psychology*, 93, 48.

Hanushek, E.A., Rivkin, S.G., 2010. Generalizations about using value-added measures of teacher quality. *The American Economic Review*, 267–271.

Hanushek, E.A., Woessmann, L., 2011. Overview of the symposium on performance pay for teachers. *Economics of Education Review,* 30, 391–393.

Heinrich, C. J., & Marschke, G. 2010. Incentives and their dynamics in public sector performance management systems. *Journal of Policy Analysis and Management*, *29*(1), 183.

Holmstrom, B., Milgrom, P., 1991. Multitask Principal–Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, and Organization,* 7, 24–52.

Kamenica, E., 2012. Behavioral Economics and Psychology of Incentives. *Annual Review of Economics*, 4, 427–452.

Knoeber, C.R., Thurman, W.N., 1994. Testing the Theory of Tournaments: An Empirical Analysis of Broiler Production. *Journal of Labor Economics,* 12, 155–179.

Kremer, M., Chaudhury, N., Rogers, F.H., Muralidharan, K., Hammer, J., 2005. Teacher Absence in India: A Snapshot. *Journal of the European Economic Association*, 3, 658–667.

Lavy, V., 2002. Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement. *Journal of Political Economy*, 110, 1286–1317.

Lavy, V., 2009. Performance Pay and Teachers' Effort, Productivity, and Grading Ethics. *American Economic Review*, 99, 1979–2011.

Lavy, V., 2015. Teachers' Pay for Performance in the Long-Run: Effects on Students' Educational and Labor Market Outcomes in Adulthood (Working Paper No. 20983). National Bureau of Economic Research.

Lazear, E.P., 2003. Teacher incentives. *Swedish Economic Policy Review,* 10, 179–214.

Lazear, E. P., 2006. Speeding, Terrorism, and Teaching to the Test, *Quarterly Journal of Economics,* 121:3, 1029–1061.

Lazear, E. P., and Rosen, S.. 1981. Rank-Order Tournaments as Optimum Labor Contracts, *Journal of Political Economy*, 89:5, 841–864.

Luo, R., Miller, G., Rozelle, S., Sylvia, S., Vera-Hernandez, M. 2015. "Can Bureaucrats Really be Paid Like CEOs? School Administrator Incentives for Anemia Reduction in Rural China," NBER Working Paper.

Macartney, H. 2014. *The Dynamic Effects of Educational Accountability* (No. w19915). National Bureau of Economic Research.

Muralidharan, K. & Sundararaman, V. (2011). Teacher Performance Pay: Experimental Evidence from India. *Journal of Political Economy*, 119(1), 39 - 77.

Murnane, R.J., Ganimian, A.J., 2014. Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations (Working Paper No. 20284). National Bureau of Economic Research.

Neal, D. 2011. *The design of performance pay in education* (No. w16710). National Bureau of Economic Research.

Neal, D., & Schanzenbach, D. W. 2010. Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics*, *92*(2), 263-283.

Niederle, M., Vesterlund, L., 2007. Do Women Shy Away From Competition? Do Men Compete Too Much? *The Quarterly Journal of Economics,* 122, 1067–1101. doi:10.1162/qjec.122.3.1067

Organisation for Economic Co-operation and Development. 2009. *Evaluating and rewarding the quality of teachers: International practices.* Paris: OECD.

Podgursky, M. J., & Springer, M. G. 2007. Teacher performance pay: A review. *Journal of Policy Analysis and Management*, *26*(4), 909.

Prelec, D. (2004). A Bayesian truth serum for subjective data. *Science*, *306*(5695), 462-466.

Rivkin, S.G., Hanushek, E.A., Kain, J.F., 2005. Teachers, schools, and academic achievement. *Econometrica,* 417–458.

Saez, E., 2010. Do Taxpayers Bunch at Kink Points? *American Economic Journal: Economic Policy,* 2, 180–212.

Springer, M.G., Hamilton, L., McCaffrey, D.F., Ballou, D., Le, V.-N., Pepper, M., Lockwood, J.R., Stecher, B.M., 2010. Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching. National Center on Performance Incentives.

Staiger, D. O., & Rockoff, J. E. 2010. Searching for effective teachers with imperfect information. *The Journal of Economic Perspectives*, *24*(3), 97-117.

Woessmann, L., 2011. Cross-Country Evidence on Teacher Performance Pay. *Economics of Education Review,* 30, 404–418.

Figure 1: Cumulative Distribution of Standardized Exam Scores at Endline by Treatment Group

Figure 1A: Full Sample

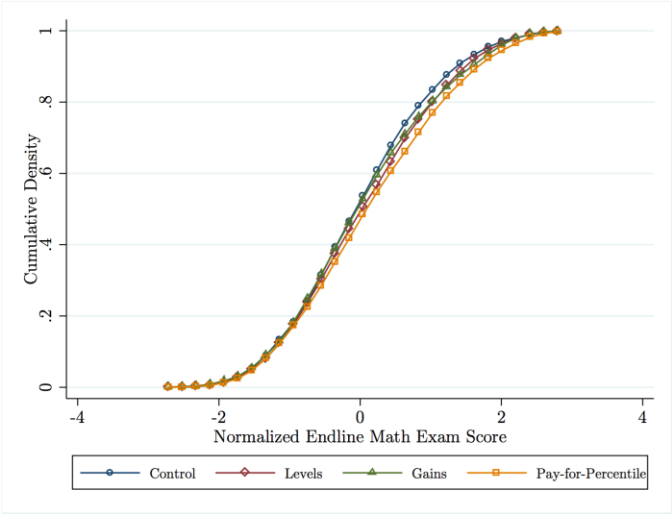

Figure 1B: Large Reward Schools

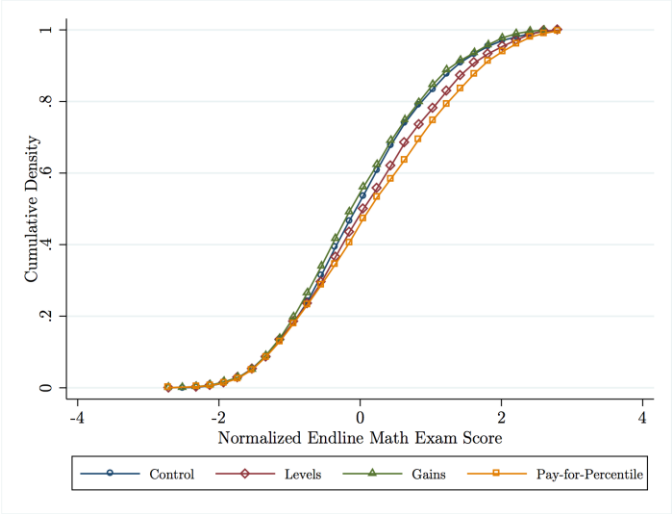

Figure 1B: Small Reward Schools

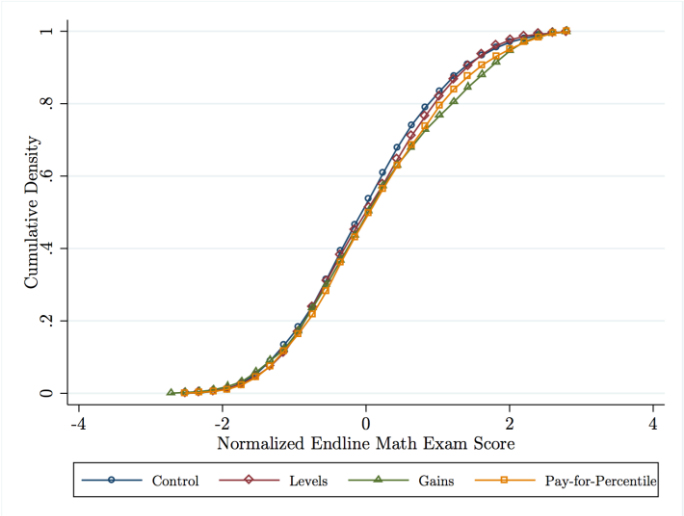

Table 1: Experimental Design

Teacher Performance Index Groups:	**Reward Size Groups:**	
	X. Large Reward	*Y. Small Reward*
A. Control	A. 52 schools	
B. Levels	BX. 26 schools	BY. 28 schools
C. Gains	CX. 26 schools	CY. 30 schools
D. Pay for percentile	DX. 26 schools	DY. 28 schools

Table 2: Impact of Incentives on Test Scores

		Full Sample						Small Reward Groups Only		Large Reward Groups Only	
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A. Impacts Relative to Control Group											
(1)	Any Incentive	0.063	0.074*								
		(0.043)	(0.044)								
(2)	Small Incentive			0.063	0.081						
				(0.053)	(0.055)						
(3)	Large Incentive			0.064	0.067						
				(0.045)	(0.046)						
(4)	Levels Incentive					0.056	0.084	0.046	0.080	0.064	0.081
						(0.048)	(0.052)	(0.059)	(0.067)	(0.059)	(0.061)
(5)	Gains Incentive					0.012	0.001	0.049	0.037	-0.033	-0.033
						(0.051)	(0.050)	(0.064)	(0.063)	(0.060)	(0.061)
(6)	Pay-for-Percentile Incentive					0.128**	0.148**	0.089	0.131	0.163***	0.165***
						(0.064)	(0.064)	(0.094)	(0.100)	(0.059)	(0.060)
(7)	Additional Controls	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
(8)	Observations	7454	7373	7454	7373	7454	7373	4655	4609	4678	4628
Panel B. Comparisons Between Incentive Treatments											
(9)	Large - Small			0.001	-0.014						
(10)	P-value: Large - Small			0.989	0.778						
(11)	Gains - Levels					-0.044	-0.083	0.003	-0.043	-0.096	-0.114
(12)	P-value: Gains - Levels					0.390	0.114	0.974	0.605	0.153	0.100
(13)	P4P - Levels					0.072	0.064	0.043	0.051	0.099	0.085
(14)	P-value: P4P - Levels					0.236	0.292	0.648	0.602	0.157	0.237
(15)	P4P - Gains					0.116	0.147	0.041	0.094	0.195	0.199
(16)	P-value: P4P - Gains					0.078	0.023	0.698	0.406	0.005	0.004

NOTES. Rows 1-6 in Panel A show estimated coefficients and standard errors obtained by estimating Equation 1. The dependent variable in each regression is normalized math exam scores at endline. Each regression controls for two waves of baseline standardized math exam scores and strata (county) fixed effects. Additional control variables (in even numbered columns) include student gender, age, parent educational attainment, a household asset index, class size, teacher experience and teacher base salary. Panel B presents differences between estimated impacts between incentive treatment groups and corresponding p-values. All tests account for clustering at the school level.*, **, and *** indicate significance at 10%, 5% and 1%.

Table 3: Impacts on Secondary Outcomes

Dependent Variable:	Math Self Concept (1)	Math Anxiety (2)	Math Intrinsic Motivation (3)	Math Instrumental Motivation (4)	Taught Easy Curriculum (5)	Taught Medium Curriculum (6)	Taught Hard Curriculum (7)	Student Time on Math (8)	Student Perception of Teacher Teaching Practice (9)	Teacher Cares (10)	Teacher Can Manage (11)	Teacher Communication (12)	Parents Help with Homework (13)	Teacher Self-reported Effort (14)
Panel A. Any Incentive														
(1) Any Incentive	0.008	0.009	0.070	0.016	0.015*	0.022**	0.011	0.012	0.026	0.008	0.010	0.029	0.012	0.010
	(0.034)	(0.031)	(0.047)	(0.036)	(0.009)	(0.009)	(0.012)	(0.043)	(0.033)	(0.055)	(0.043)	(0.045)	(0.039)	(0.064)
Panel B. Incentive Size														
(2) Small Incentive	0.014	0.015	0.061	0.030	0.019*	0.025***	0.019	0.012	0.002	-0.023	0.013	-0.006	0.006	0.064
	(0.037)	(0.035)	(0.053)	(0.040)	(0.010)	(0.010)	(0.013)	(0.051)	(0.036)	(0.063)	(0.048)	(0.053)	(0.046)	(0.069)
(3) Large Incentive	0.003	0.002	0.079	0.001	0.012	0.019**	0.003	0.013	0.050	0.039	0.008	0.065	0.018	-0.050
	(0.038)	(0.034)	(0.052)	(0.040)	(0.011)	(0.010)	(0.014)	(0.048)	(0.037)	(0.060)	(0.047)	(0.050)	(0.042)	(0.070)
Panel C. Incentive Design														
(4) Levels Incentive	0.023	0.009	0.029	-0.042	0.019*	0.020*	0.005	0.031	0.014	0.034	-0.004	-0.029	-0.059	0.055
	(0.040)	(0.039)	(0.056)	(0.046)	(0.012)	(0.010)	(0.015)	(0.056)	(0.040)	(0.063)	(0.049)	(0.055)	(0.049)	(0.078)
(5) Gains Incentive	0.012	0.024	0.093*	0.022	0.012	0.022**	-0.009	0.008	0.022	-0.003	0.001	0.043	0.062	0.003
	(0.039)	(0.034)	(0.054)	(0.039)	(0.012)	(0.010)	(0.014)	(0.055)	(0.036)	(0.066)	(0.052)	(0.048)	(0.046)	(0.075)
(6) Pay-for-Percentile Incentive	-0.011	-0.009	0.083	0.065	0.016	0.025**	0.040***	-0.001	0.040	-0.005	0.036	0.071	0.024	-0.024
	(0.043)	(0.040)	(0.063)	(0.047)	(0.012)	(0.011)	(0.014)	(0.054)	(0.045)	(0.073)	(0.055)	(0.067)	(0.048)	(0.076)
(7) Mean in Control Group	-0.009	0.003	-0.055	-0.019	0.853	0.856	0.788	-0.014	-0.024	-0.027	-0.013	-0.032	0.005	0.030
(8) Observations	7373	7373	7373	7373	7373	7370	7366	7373	7373	7372	7373	7373	7371	235

NOTES. Shows estimated coefficients and standard errors obtained by estimating regressions analogous to Equation 1. All outcome variables are summary indexes apart from columns 5-7. Summary indexes were constructed using the GLS weighting procedure in Anderson (2008). Each regression controls for two waves of baseline standardized math exam scores, strata (county) fixed effects as well as student gender, age, parent educational attainment, a household asset index, class size, teacher experience and teacher base salary. The outcome in column 14 is at the teacher level; regressions include school level aggregates of control variables. All standard errors account for clustering at the school level.*, **, and *** indicate significance at 10%, 5% and 1%.

Table 4: Impacts on Quesiton Difficulty Subscores

		Full Sample			Small Reward Groups Only			Large Reward Groups Only		
		Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	Levels Incentive	0.056	0.215*	0.157	0.076	0.17	0.158	0.025	0.245*	0.139
		(0.085)	(0.115)	(0.11)	(0.121)	(0.138)	(0.138)	(0.094)	(0.13)	(0.131)
(2)	Gains Incentive	-0.012	-0.022	0.04	0.021	0.093	0.074	-0.038	-0.125	-0.005
		(0.07)	(0.114)	(0.112)	(0.071)	(0.14)	(0.147)	(0.097)	(0.137)	(0.128)
(3)	Pay-for-Percentile Incentive	0.204**	0.211	0.335**	0.219*	0.17	0.274	0.203**	0.237*	0.399***
		(0.084)	(0.142)	(0.14)	(0.118)	(0.223)	(0.215)	(0.093)	(0.125)	(0.137)

NOTES. Table shows estimated coefficients and standard errors obtained by estimating Equation 1. The dependent variable in each regression is an endline exam subscore (for easy, medium and hard items). Test questions were classified as easy, medium and hard based on the rate of correct responses in the control group. Each regression controls for two waves of baseline standardized math exam scores, strata (county) fixed effects, student gender, age, parent educational attainment, a household asset index, class size, teacher experience and teacher base salary. All standard errors account for clustering at the school level.*, **, and *** indicate significance at 10%, 5% and 1%.

Table 5: Distributional Effects by Baseline Exam Score

	Levels					Gains					Pay-for-Percentile					
	Effect on:			Difference (Top - Bottom)	Linear Interaction	Effect on:			Difference (Top - Bottom)	Linear Interaction	Effect on:			Difference (Top - Bottom)	Linear Interaction	Observations
	Bottom 1/3 (1)	Middle 1/3 (2)	Top 1/3 (3)	(4)	(5)	Bottom 1/3 (6)	Middle 1/3 (7)	Top 1/3 (8)	(9)	(10)	Bottom 1/3 (11)	Middle 1/3 (12)	Top 1/3 (13)	(14)	(15)	(16)
Panel A. Across Full Sample Distribution																
(1) Second Baseline Exam Score	0.053	0.129	0.059	0.006	0.005	-0.022	0.041	-0.031	-0.010	-0.015	0.129	0.187***	0.112	-0.017	-0.015	7373
	0.065	0.070	0.070	(0.085)	(0.033)	0.060	0.063	0.074	(0.080)	(0.033)	0.096	0.069	0.076	(0.104)	(0.046)	
(2) Average of Baseline Scores	0.046	0.175**	0.058	0.012	-0.024	0.010	0.018	-0.027	-0.037	-0.042	0.111	0.228***	0.148*	0.038	-0.067	7373
	0.069	0.067	0.067	(0.087)	(0.095)	0.066	0.062	0.067	(0.085)	(0.098)	0.098	0.070	0.075	(0.107)	(0.123)	
Panel B. Within Class Distribution																
(3) Within Class Percentile Rank	0.067	0.102	0.078	0.011	-0.007	-0.009	0.024	-0.015	-0.006	-0.031	0.172*	0.126*	0.145**	-0.027	-0.024	7373
(Second Exam)	0.063	0.064	0.061	(0.066)	(0.037)	0.065	0.059	0.064	(0.070)	(0.038)	0.089	0.068	0.073	(0.086)	(0.054)	
(4) Within Class Percentile Rank	0.113*	0.104	0.045	-0.068	-0.065	0.038	-0.008	-0.023	-0.060	-0.062	0.188*	0.140*	0.124*	-0.063	-0.080	7373
(Both Exams)	0.061	0.066	0.061	(0.062)	(0.091)	0.060	0.064	0.063	(0.065)	(0.095)	0.083	0.079	0.069	(0.083)	(0.124)	

NOTES. Data source: Full sample. Each row shows effects of incentive designs on normalized math scores by the distribution of baseline scores at left. Panel A shows effects along the distribution of scores across the full sample of students and Panel B shows effects by within class percentile rank. Effects for each tercile of the distribution for a given baseline score were estimated using a single regression analogous to Equation (1), but including dummy variables for the second and third terciles and interactions with treatment arms. Columns 5,10 and 15 report the coefficient and standard error on an interaction between the corresponding treatment group and baseline score entered linearly (instead of second and third tercile dummies). Estimates in Row (1) do not control for second-wave baseline scores and estimates in Row 2 do not control for either wave of baseline scores.

Table 6: Heterogeneous Effects by Baseline Teacher and Class Characteristics

Baseline Variable (VAR):	Teacher Experience (>Median) (1)	Teacher Base Pay (>Median) (2)	Teacher Female (3)	Teacher Intrinsic Motivation High (>Median) (4)	Teacher Prosocial Motivation (>Median) (5)	Teacher Risk Averse (6)	Teacher Inequality Averse (7)	Teacher Self-perceived Value Added High (>Median) (8)	Teacher Self-perceived Value Added for Bottom Students High (>Median) (9)	Teacher Self-perceived Value Added for Middle Students High (>Median) (10)	Teacher Self-perceived Value Added for Top Students High (>Median) (11)	Class Average Baseline Scores in Top Tercile in Sample (12)	Class Average Baseline Scores in Middle Tercile in Sample (13)	Class Average Baseline Scores in Bottom Tercile in Sample (14)	Grade Size (15)	Number of 6th Grade Math Teachers in School (16)
(1) Levels Incentive	0.046	0.110	0.104	0.054	0.081	0.068	0.088*	0.031	-0.019	-0.001	0.004	0.110*	0.041	0.101	0.050	0.093
	(0.068)	(0.075)	(0.066)	(0.079)	(0.067)	(0.080)	(0.053)	(0.081)	(0.078)	(0.083)	(0.073)	(0.061)	(0.069)	(0.063)	(0.184)	(0.144)
(2) Levels × VAR	0.087	-0.035	-0.053	0.062	0.012	0.035	-0.033	0.082	0.170*	0.133	0.134	-0.086	0.117	-0.050	0.001	-0.009
	(0.092)	(0.103)	(0.100)	(0.101)	(0.095)	(0.104)	(0.128)	(0.096)	(0.097)	(0.095)	(0.093)	(0.113)	(0.110)	(0.115)	(0.004)	(0.123)
(3) Gains Incentive	-0.053	0.043	-0.093	-0.029	-0.083	-0.017	-0.003	-0.050	-0.021	-0.014	-0.005	0.031	-0.052	0.017	-0.394**	-0.059
	(0.064)	(0.075)	(0.063)	(0.086)	(0.074)	(0.073)	(0.051)	(0.079)	(0.079)	(0.082)	(0.075)	(0.058)	(0.063)	(0.064)	(0.162)	(0.146)
(4) Gains × VAR	0.119	-0.078	0.186*	0.059	0.167	0.034	0.156	0.111	0.036	0.031	0.025	-0.097	0.135	-0.053	0.008**	0.048
	(0.101)	(0.103)	(0.096)	(0.103)	(0.102)	(0.108)	(0.130)	(0.098)	(0.100)	(0.104)	(0.101)	(0.110)	(0.103)	(0.106)	(0.003)	(0.132)
(5) Pay-for-Percentile Incentive	0.230***	0.156	0.092	0.077	-0.000	0.138	0.212***	0.020	-0.011	0.027	0.076	0.197**	0.062	0.199***	-0.307*	0.040
	(0.088)	(0.095)	(0.089)	(0.081)	(0.065)	(0.093)	(0.071)	(0.076)	(0.074)	(0.079)	(0.070)	(0.086)	(0.080)	(0.074)	(0.164)	(0.148)
(6) Pay-for-Percentile × VAR	-0.190	-0.007	0.142	0.127	0.295***	0.022	-0.248*	0.264**	0.314***	0.255**	0.177	-0.151	0.324***	-0.126	0.010***	0.085
	(0.123)	(0.116)	(0.120)	(0.124)	(0.110)	(0.117)	(0.132)	(0.121)	(0.119)	(0.128)	(0.130)	(0.135)	(0.124)	(0.162)	(0.004)	(0.116)
(7) VAR	0.026	-0.102	-0.005	-0.100	-0.080	0.006	0.021	-0.146**	-0.160**	-0.127*	-0.070	0.084	-0.097	0.023	-0.003	0.046
	(0.077)	(0.092)	(0.073)	(0.071)	(0.063)	(0.073)	(0.096)	(0.067)	(0.066)	(0.069)	(0.068)	(0.085)	(0.073)	(0.085)	(0.003)	(0.115)
(8) Observations	7373	7373	7373	7373	7373	7373	7373	7273	7273	7249	7257	7373	7373	7373	7373	7373

NOTES. Shows estimated coefficients and standard errors obtained by estimating regressions analogous to Equation 1, but adding the baseline variable of interest and interactions with incentive treatment dummies. Each regression controls for two waves of baseline standardized math exam scores, strata (county) fixed effects as well as student gender, age, parent educational attainment, a household asset index, class size, teacher experience and teacher base salary. All standard errors account for clustering at the school level.*, **, and *** indicate significance at 10%, 5% and 1%.

Appendix Table 1: Descriptive Statistics and Balance Check

		Coefficient (standard error) on:				Coefficient (standard error) on:			
	Control Mean	Levels Incentive	Gains Incentive	Pay-for-Percentile Incentive	Joint Test P-value: All=0	Small Incentive	Large Incentive	Joint Test P-value: All=0	Observations
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Student Characteristics									
(1) Standardized Math Test Score, Beginning of Previous School	0.000	-0.045	-0.015	-0.094	0.739	-0.040	-0.061	0.751	7996
		(0.084)	(0.082)	(0.093)		(0.079)	(0.080)		
(2) Standardized Math Test Score, End of Previous School Year	0.000	-0.005	0.028	-0.038	0.894	0.015	-0.023	0.848	8136
		(0.082)	(0.091)	(0.088)		(0.080)	(0.081)		
(3) Female (0/1)	0.492	-0.010	-0.002	-0.011	0.893	-0.005	-0.010	0.816	7996
		(0.017)	(0.015)	(0.018)		(0.015)	(0.015)		
(4) Age (Years)	11.986	0.088	0.137**	0.082	0.225	0.104*	0.103*	0.176	7992
		(0.063)	(0.066)	(0.072)		(0.062)	(0.061)		
(5) Father Attended Secondary School (0/1)	0.517	0.005	0.028	0.005	0.686	0.007	0.019	0.700	7965
		(0.024)	(0.026)	(0.026)		(0.023)	(0.023)		
(6) Mother Attended Secondary School (0/1)	0.312	0.010	0.019	0.011	0.900	0.021	0.007	0.660	7929
		(0.026)	(0.026)	(0.026)		(0.024)	(0.023)		
(7) Household Asset Index	-0.636	0.025	0.014	0.041	0.865	-0.001	0.054	0.348	7996
		(0.046)	(0.048)	(0.050)		(0.042)	(0.042)		
Panel B. Teacher and Class Characteristics									
(8) Age (Years)	32.621	1.671	0.367	0.581	0.745	0.305	1.548	0.549	243
		(1.599)	(1.682)	(1.473)		(1.347)	(1.572)		
(9) Female	0.421	-0.019	0.095	-0.013	0.492	0.012	0.031	0.933	243
		(0.091)	(0.089)	(0.093)		(0.082)	(0.087)		
(10) Han (0/1)	0.947	0.010	-0.062*	-0.014	0.229	-0.042*	0.003	0.134	243
		(0.034)	(0.035)	(0.027)		(0.024)	(0.034)		
(11) Teaching Experience (Years)	11.605	1.858	0.844	-0.167	0.617	0.477	1.224	0.789	243
		(1.772)	(1.994)	(1.630)		(1.509)	(1.808)		
(12) Monthly Base Salary (Yuan)	2852.772	255.599*	-149.432	142.402	0.054	119.440	37.325	0.713	243
		(152.651)	(187.318)	(175.438)		(161.684)	(160.419)		
(13) Grade Size	43.346	-1.154	2.407	-3.430	0.300	-2.296	1.089	0.416	216
		(2.877)	(2.971)	(2.819)		(2.615)	(2.581)		
Panel C. School Characteristics									
(14) Number of Students	437.827	-59.555	-31.874	-46.852	0.807	-71.814	-16.537	0.270	216
		(62.562)	(60.861)	(65.916)		(58.522)	(60.857)		
(15) Number of Teachers	29.750	-0.447	-2.744	-0.979	0.859	-3.531	1.029	0.235	216
		(4.234)	(3.692)	(4.223)		(3.488)	(3.996)		
(16) Number of Contract Teachers	1.692	0.403	0.073	0.063	0.937	0.116	0.248	0.884	216
		(0.645)	(0.388)	(0.415)		(0.380)	(0.501)		

NOTES. Data source: baseline survey. The first column shows the mean in the control group. Columns 2-4 and 6-7 show coefficients and standard errors from a regression of each characteristic on indicators for incentive treatments, controlling for randomization strata. Columns 5 and 8 shows the p-value from a test that preceding coefficients are jointly zero. All tests account for clustering at the school level. *, **, and *** indicate significance at 10%, 5% and 1%.

Appendix Table 2: Attrition

		Full Sample		Small Incentive Groups Only	Large Incentive Groups Only
		(1)	(2)	(3)	(4)
(2)	Small Incentive	-0.004			
		(0.014)			
(3)	Large Incentive	-0.007			
		(0.014)			
(4)	Levels Incentive		0.008	0.028	-0.007
			(0.019)	(0.033)	(0.013)
(5)	Gains Incentive		-0.015	-0.014	-0.018
			(0.010)	(0.013)	(0.013)
(6)	Pay-for-Percentile		-0.008	-0.026*	0.009
	Incentive		(0.017)	(0.013)	(0.030)
(8)	Observations	9072	9072	5719	5607
(9)	Mean in Control			0.064	

NOTES. The dependent variable in each regression is a dummy variable indicating a student was absent from the endline survey. Each regression controls for strata (county) fixed effects. Standard errors account for clustering at the school level.*, **, and *** indicate significance at 10%, 5% and 1%.