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

Prashant Loyalka, Sean Sylvia, Chengfang Liu, James Chu, Yaojiang Shi[†]

October 16, 2016

ABSTRACT: We present results of a randomized trial testing alternative approaches of mapping student achievement into rewards for teachers. Teachers in 216 schools in western China were assigned to performance pay schemes where teacher performance was assessed by one of three different methods. We find that teachers offered "pay-for-percentile" incentives (Barlevy and Neal 2012) outperform teachers offered simpler schemes based on class average achievement or average gains over a school year. Moreover, pay-for-percentile incentives produced broad-based gains across students within classes. That teachers respond to relatively intricate features of incentive schemes highlights the importance of close attention to performance pay design.

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

[†] Loyalka: Stanford University, Encina Hall East Wing Room 401, 616 Serra St., Stanford, CA 94305 (email: aielman@stanford.edu); Sylvia (corresponding author): Renmin University of China, Mingde Building Room 611, 59 Zhongguancun Ave., Beijing 100872 (e-mail: ssylvia@ruc.edu.cn); Liu: Peking University, Wangkezhen Building Room 409, No. 5 Yiheyuan Road, Beijing 100871; Chu: Stanford University, Encina Hall East Wing Room 401, 616 Serra St., Stanford, CA 94305 (email: jchu1225@stanford.edu); Shi: Shaanxi Normal University, 620 Chang'an Road West, Xi'an 710119, China (e-mail:shiyaojiang7@gmail.com). We are grateful to Grant Miller, Karthik Muralidharan, Derek Neal, Scott Rozelle and Marcos Vera-Hernández for helpful comments on earlier versions of the manuscript and to Jingchun Nie for research assistance. We would also like to thank students at the Center for Experimental Economics in Education (CEEE) at Shaanxi Normal University for exceptional project support as well as the Ford Foundation and Xu Family Foundation for financing the project.

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 tied to teacher attributes such as education level and experience that are not strongly associated with student achievement (Rivkin, Hanushek, and Kain 2005; Podgursky and Springer 2007; Hanushek and Rivkin 2010). Possibly due to a lack of explicit incentives to improve student outcomes, teacher absenteeism is pervasive in many parts of the world (Kremer et al. 2005; Banerjee and Duflo 2006; Chaudhury et al. 2006) and teachers often fail to teach effectively when present (Chaudhury et al. 2006; Staiger and Rockoff 2010). Policies that unconditionally increase teacher salaries – but do not provide incentives – may further fail to improve teacher effort or student learning (de Ree et al. 2015). In response, a growing movement seeks to better align teacher incentives by linking teacher pay more directly to student achievement, and performance pay programs are increasingly common in both developed and developing countries (OECD 2009; Hanushek and Woessmann 2011; Bruns et al. 2011; Woessmann 2011).

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 productive teacher effort are likely ineffective. Schemes involving performance targets, for instance, can fail to motivate teachers who believe that they have little chance of reaching these targets or teachers for whom achieving these targets would require little effort (Neal 2011). How incentive schemes are designed can further lead to triage across students, strengthening incentives for teachers to focus on students whose outcomes are more closely linked to rewards while neglecting others (Neal and Schanzenbach 2010; Contreras and Rau 2012). Certain designs may also be more likely than others to encourage teachers to "teach to the test," or devote effort toward improving student performance measures rather than actual student learning (Holmstrom and Milgrom 1991; Baker 1992; Dixit 2002).

While studies have highlighted weaknesses in specific 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 on student outcomes.[1] An important question is to what degree more intricate features of design actually matter in practice. Although theoretically appealing (and often more complex) designs meant to address common failures exist, there is little evidence to suggest whether these outperform less appealing but simpler schemes in practice (Leigh 2013). Evidence from contexts outside of education suggests that individuals may not respond as intended when faced with complex incentives and price schedules; responding to average rather than marginal prices, for instance (Liebman and Zeckhauser 2004; Dynarski and Scott-Clayton 2006; Ito 2014; Abeler and Jäger 2015).[2] The complexity of incentive schemes may also reduce perceived transparency, perhaps an important factor when trust in implementing agencies is low (Muralidharan and Sundararaman 2011).

In this paper, we study incentive design directly by comparing performance pay schemes that vary in how student achievement (performance on standardized exams) is used to measure and reward teacher performance. How student achievement scores are used to measure teacher performance and mapped onto rewards can—independently of the size or amount of potential rewards—affect the strength of incentive schemes and hence effort devoted by teachers toward improving student outcomes (Neal and Schanzenbach 2010; Bruns et al. 2011; Neal 2011). We focus specifically on alternative ways of defining a measure of teacher performance using the achievement scores of the multiple students in a teacher's class. In addition to affecting the overall strength of a performance pay scheme, the way in which achievement scores of individual students are combined into a measure of teacher performance may also affect how teachers choose to allocate effort and attention across different students in the classroom by explicitly or implicitly weighting some students in the class more than others.

[1] An important exception is Fryer et al. (2012) who compare incentives designed to exploit loss aversion with a more traditional incentive scheme. There have also been several studies comparing incentive schemes that vary in who is rewarded. These include Muralidharan and Sundararaman (2011) who compare individual and group incentives for teachers in India (Fryer et al. (2012) also compares individual and group incentives); Behrman et al. (2015) who present an experiment in Mexico comparing incentives for teachers to incentives for students and joint incentives for students, teachers and school administrators; and Barrera-Osorio and Raju (2015) who compare incentives for school principals only, incentives for school principals and teachers together, and larger incentives for school principals combined with (normal) incentives for teachers in an experiment in Pakistan.

[2] Ito (2014), for instance, finds that individuals in the US respond to average rather than marginal prices for electricity (thus rendering nonlinear pricing schedules ineffective).

We compared 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 a measure of teacher performance used to rank and reward teachers (hereafter "incentive design" treatments). Teachers in half of the schools in each of these treatment groups were then randomly allocated to a small reward treatment or a large reward treatment (where rewards were twice as large, but remained within policy-relevant levels).

We present three main findings. First, we find that teachers offered "pay-for-percentile" incentives—which reward teachers based on the rankings of individual students within appropriately-defined comparison sets, based on the scheme described in Barlevy and Neal (2012)—outperformed teachers offered two simpler schemes that rewarded 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. Pay-for-percentile incentives increased student achievement by approximately 0.15 standard deviations on average. Tests of distributional treatment effects, which take into account higher-order moments of test score distributions (Abadie, 2002), show pay-for-percentile incentives significantly outperformed both gains and levels incentives, while levels incentives outperformed gains incentives. Achievement gains under pay-for-percentile were mirrored by meaningful increases in the intensity of teaching as evidenced by teachers covering more material, covering more advanced curricula, and students being more likely to correctly answer difficult exam items.

Second, we do not find that doubling the size of potential rewards (from approximately one month of salary to two months of salary on average) has a significant effect on student achievement. Taken together with findings for how effects vary across the incentive design treatments, these results are remarkable in that they suggest that in our context the design of the incentive—specifically how teachers are ranked and rewarded according to the achievement of their students—has a larger effect on student performance than doubling the size of potential rewards.

Third, we find evidence that—following theoretical predictions—levels and gains incentives led teachers to focus on students for whom they perceived their own teaching

effort would yield the largest gains in terms of exam performance while pay-for-percentile incentives did not. This aligns with how the pay-for-percentile scheme rewards achievement gains more symmetrically across students within a class. For levels and gains incentives, focus on higher value-added students did not, however, translate into varying effects along the distribution of initial achievement within classes. Levels and gains incentives had no significant effects for students at any part of the distribution. Pay-for-percentile incentives, by contrast, led to broad-based gains along the distribution.

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 mixed. On the one hand, several studies have found teacher performance pay to be effective at improving student achievement, particularly in developing countries where hidden action problems tend to be more prevalent (Lavy 2002; Lavy 2009; Glewwe et al. 2010; Muralidharan and Sundararaman 2011; Duflo et al. 2012; Fryer et al. 2012; Dee and Wyckoff 2015).[3,4] 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 math and language scores (Muralidharan and Sundararaman 2011; Muralidharan 2012). In contrast, other recent studies in developed and developing countries have not found significant effects on student achievement (Springer et al. 2010; Fryer 2013; Behrman et al. 2015; Barrera-Osorio and Raju 2015).

Beyond providing more evidence on the effectiveness of incentives generally, we contribute to the teacher performance pay literature in three ways. Our primary contribution is the direct comparison of alternative methods of measuring and rewarding teacher performance as a function of student achievement. 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.[5] Only two studies provide direct

[3] 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.

[4] 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.

[5] 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

experimental comparisons of design features of incentive schemes for teachers. 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 a measure of teacher performance matters—independent of other design features. Second, we provide evidence suggesting that incentive schemes can be designed so as to largely eliminate triage by shifting teachers' instructional focus and allocation of effort more equally across students within a class. This finding adds to evidence that teachers tailor the focus of instruction to different students in response to cutoffs in incentive schemes and in response to class composition (Neal and Schanzenbach 2011; Duflo, Dupas and Kremer 2011). Third, this study is the first of which we are aware that experimentally compares varying sizes of monetary rewards for teachers (adding to three recent experimental studies which test the impacts of incentive reward size in alternative contexts– Ashraf, Bandiera and Jack (2014), Luo et al. (2015), and Barrera-Osario and Raju (2015)).[6]

Our findings also contribute to literatures outside of education. In general, our

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. The scheme studied in Barrera-Osario and Raju (2015) rewarded teachers based on linear function of a composite score where the composite score is a weighted combination of exam score gains, enrollment gains, and exam participation rates.

[6] Ashraf, Bandiera and Jack (2014) and Luo et al. (2015) study incentives in health delivery, including comparisons of 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. Ashraf, Bandiera and Jack (2014) find that small and large rewards were both ineffective while Luo et al. (2015) finds that larger rewards have larger effects than smaller rewards. Barrera-Osario and Raju (2015) compare small and large rewards (twice the size) for school principals conditional on teachers receiving small rewards. They find that increasing the size of potential principal rewards when teachers also had incentives did not lead to improvements in school enrollment, exam participation or exam scores.

results add to a growing number of studies that use field experiments to evaluate performance incentives in organizations (Bandiera et al. 2005, 2007; Cadsby et al. 2007; Bardach et al. 2013). We also contribute to the literature on tournaments, particularly 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 reward structure. Finally, despite evidence from elsewhere that individuals do not react as intended to complex incentives and prices, our results indicate that teachers can respond to relatively complex features of reward schemes. While we cannot say if teachers responded optimally to the incentives they were given, we find that they did respond more to pay-for-percentile incentives than more simple schemes and that they allocated effort across students in line with theoretical predictions. Inasmuch as our results indicate that teachers respond to relatively intricate features of incentive contracts, they suggest room for these features to affect welfare 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

The sample for our study was selected from 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 16 nationally-designated poverty counties in these two prefectures, we conducted a canvass survey of all elementary schools. From the complete list of schools, we randomly selected 216 rural schools for inclusion in the study. [7]

[7] We applied three exclusion criteria before sampling from the complete list of schools. 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

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; 2,254 students) and three incentive design groups: a "levels" incentive group (54 schools; 2,233 students), a "gains" incentive group (56 schools; 2,455 students), and a "pay-for-percentile" group (54 schools; 2,130 students).[8] Across these three incentive groups, we orthogonally assigned schools to reward size groups: a "small" reward size group (78 schools; 3,465 students) and a "large" reward size group (86 schools; 3,353 students). 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 by controlling for stratum fixed effects (Bruhn and McKenzie 2009). Note that while our study design allows for testing interaction effects between reward size and incentive design, we only powered the study to test between incentive designs and between reward sizes separately. That is, we only powered the study to test across the column labeled "Total" and across the row labeled "Total" in Table 1.

2.3. Incentive Design and Conceptual Framework

2.3.1 Incentive Design Treatments

Our primary goal is to evaluate 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 a measure of teacher performance that is used to rank teachers in the tournament. The three incentive design treatments that we evaluate are as follows:

following years, we decided to exclude these schools from our sample.

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

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

Gains Incentive: Teacher performance in the "gains" incentive treatment was defined as the class average of individual student achievement gains from the start to the end of the school year. Individual student achievement gains were measured as the difference in a student's 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.

Pay-for-Percentile Incentives: The third way of measuring teacher performance was through the "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 according to their score on the baseline exam conducted at the end of the previous school year.[9] 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 measure (percentile performance index) was then determined by the average percentile rank taken over all students in his or her class.[10] This percentile performance index can be interpreted as the fraction of contests that students of a given teacher won when compared to students who were taught by other teachers and yet began the school year at similar achievement levels (Barlevy and Neal 2012).

2.3.2 Common Rank-Order Tournament Structure

While the incentive design treatments varied in how teacher performance was measured in the determination of rewards, all incentive treatments had a common

[9] Teachers were not told the baseline achievement scores of individual students in any of the designs.
[10] We used the average as per Neal (2011).

underlying rank-order tournament structure.[11] When informed of their incentive, teachers were told that they would compete with sixth grade math teachers in other schools in their prefecture,[12] and the competition would be based on their students' performance on common standardized math exams.[13] According to their percentile ranking among other teachers in the program, teachers were told they would be given a cash reward (transferred to their bank account) within two months after the end of the school year.

Rewards were structured to be linear in percentile rank as follows:

$$Reward = R_{Top} - (99 - PercentileRank) \times b$$

where R_{Top} is the reward for teachers ranking in the top percentile and b is the incremental reward for each percentile rank. In the small reward size treatment, teachers ranking in the top percentile received 3500 *yuan* ($547) and the incremental reward per percentile rank was 35 *yuan*.[14] In the large reward size treatment, 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's salary in the small reward treatment and two months' salary in the large reward treatment.[15]

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 (for example, the tournament studied in Lavy (2009) has only four reward levels). By setting rewards to be linearly increasing in percentile rank, the underlying reward structure that we used in this

[11] Using a common underlying rank-order tournament structure allowed us to directly compare incentive designs that used different ways of measuring and rewarding teacher performance. Direct comparison would not have been possible with a piece-rate scheme as the rewarded units would have necessarily differed.

[12] The two prefectures in the study each have hundreds of primary schools (751 in the prefecture in Shaanxi and 1200 in the prefecture in Gansu). Teachers were not told the total number of teachers who would be competing in the tournament.

[13] 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.

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

[15] While there was no explicit penalty if students were absent on testing dates, contracts stated we would check and that teachers would be disqualified if 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 found no evidence that lower achieving students were less likely to sit exams at the end of the year.

study is similar to the incentive scheme studied in Knoeber and Thurman (1994).[16] We chose to use this linear structure to minimize distortions in incentive strength due to non-linearities in rewards.[17]

Relative rewards schemes such as rank-order tournaments have a number of potential advantages over piece-rate schemes. First, tournaments provide the implementing agency with budget certainty, as teachers compete for a fixed pool of money (Lavy 2009; Neal 2011); 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. Importantly for risk-averse agents, tournaments are also more robust to common shocks across all participants.[18] 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).[19]

2.3.3 Implementation

Following a baseline survey (described below), teachers in all incentive arms were presented performance pay contracts stipulating the details of their assigned incentive scheme. These contracts were signed and stamped by the Chinese Academy of

[16] 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.

[17] 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. Moldovanu and Sela (2001) present theory suggesting that the optimal (maximizing the expected sum of effort across contestants) number of prizes is increasing with the heterogeneity of ability of contestants and in the convexity of the cost functions they face. 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.

[18] 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 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 potential common shock).

[19] 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 is less important in our setting as competition was purposefully designed to be between teachers across different schools.

Sciences (a government organization) and were presented with officials from the local prefecture bureaus of education. Before signing the contract, teachers were provided with materials explaining the details of the contract and how rewards would be calculated.[20] To better ensure that teachers understood the incentive structure and contract terms, they were also given a training session lasting approximately two hours covering the same material. A short quiz was also given to teachers to check misunderstanding of the contract terms and reward determination and correct responses were reviewed with teachers.

2.3.4 Conceptual Framework

Our goal is to evaluate how each of the three ways of ranking and rewarding teachers using student's achievement scores (levels, gains, and pay-for-percentile) affects two different aspects of teacher effort. First, we aim to understand the effect of each scheme on overall effort—that is, how effective each scheme is in motivating teachers to increase the amount of effort they provide. Second, we aim to understand how each scheme affects how teachers allocate effort across students in their classes — i.e. do teachers triage certain students due to how teacher performance is measured?

Strength of the Incentive Design

According to standard contest theory, the relative strength of the incentives depends on teachers' beliefs about the mapping between their effort and expected changes in their performance rank. Assuming that teachers choose effort to maximize their reward (rank) in the contest, ranking teachers according to pay-for-percentile should provide stronger incentives overall than ranking teachers according to levels or gains. This is because pay-for-percentile places teachers in more symmetric contests in which they compete with teachers that have students with the same levels of baseline achievement. This symmetry strengthens incentives by reducing differences across teachers in expected marginal returns to effort (in terms of expected tournament rank). That is, teachers are less likely to believe either they or their competitors have an

[20] Chinese and translated versions of these materials are available for download at http://reap.stanford.edu.

advantage and that rank in the contest is more directly a result of the relative effort provided.

Assuming that teachers do respond to relatively intricate features of incentive design, ranking and rewarding teachers based on levels or gains in student achievement should create a weaker incentive relative to pay-for-percentile because of greater asymmetry due to (a) variation in baseline student ability, (b) variation in potential growth (teacher returns to effort) as a function of baseline student ability, (c) additional noise due to measurement, and (d) teacher uncertainty related to seeding. The relative strength of levels versus gains incentives is less clear and depends on how teachers perceive that gains in student achievement vary across students with different levels of baseline achievement.

To illustrate, first consider the case in which each teacher has only one student. The endline test score of each teacher's student, a_j, is produced according to

$$a_j = a_{j(t-1)} + \gamma(a_{j(t-1)})e_j + \nu_j \quad (1)$$

where e_j is the effort of teacher j, $a_{j(t-1)}$ is the baseline test score of her student, and ν_j is a shock to the student's endline test score due to luck. The parameter $\gamma(a_{j(t-1)})$ allows the productivity of teaching effort to vary with baseline student achievement. In a contest with J teachers, each teacher will choose effort to maximize her expected reward (incrementally increasing in tournament rank by a parameter π) less her cost of effort, $c(e_j)$ (with $c'(e) > 0$ and $c''(e) > 0$, assumed constant across teachers for simplicity) as

$$\max_{e_j} \sum_{k \neq j} \pi F\left(a_{j(t-1)} + \widetilde{\gamma}_j(a_{j(t-1)})e_j - a_{k(t-1)} - \widetilde{\gamma}_j(a_{k(t-1)})e_k\right) - c(e_j) \quad (2)$$

where $F(\varepsilon_{jk})$ is the distribution of $\varepsilon_{jk} = \nu_j - \nu_k$ which is identically and independently distributed for all (j, k) pairs. $\widetilde{\gamma}_j(\cdot)$ is teacher j's perception of how the productivity of teaching effort varies with baseline student achievement. Each teacher's first order condition is

$$\sum_{k \neq j} \pi \widetilde{\gamma}_j(a_{j(t-1)}) f\left(a_{j(t-1)} + \widetilde{\gamma}_j(a_{j(t-1)})e_j - a_{k(t-1)} - \widetilde{\gamma}_j(a_{k(t-1)})e_k\right) = c'(e_j). \quad (3)$$

That is, teachers will chose effort such that their marginal return to effort in terms of the number of individual contests with other teachers that they "win" is equal to their marginal cost of effort. A teacher's marginal return to effort depends on how much effort contributes to the probability that her student will outperform competitors' students given

differences in student ability, other teachers' efforts and the realizations of the random shocks. When $a_{j(t-1)} = a_{k(t-1)}$, the contest is symmetric and the Nash Equilibrium of this game is where all teachers chose the same, efficient level of effort, $e^* = e_j = e_k$.[21] As $a_{j(t-1)}$ and $a_{k(t-1)}$ diverge, however, the symmetry of the contest is reduced as differences in student ability become more important relative to differences in teacher effort in determining the winner of the contest.

Under pay-for-percentile, $a_{j(t-1)} = a_{k(t-1)}$ by construction: teachers only compete with teachers that teach students with the same levels of baseline achievement. Thus, pay-for-percentile is more likely to elicit efficient and symmetric effort from all teachers.[22]

The symmetry in teacher beliefs required to elicit efficient effort is less likely in the case of levels or gains incentives. Because $a_{j(t-1)}$ is not the same across all teachers, and assuming that teachers take this into account, there will generally be no equilibrium where $e^* = e_j = e_k$.

With levels incentives, the symmetry of the contest (and hence the strength of the incentive) will depend on the difference between $a_{j(t-1)}$ and $a_{k(t-1)}$ as well as teacher's perceptions of how the parameter $\gamma(\cdot)$ changes with baseline student achievement. Teachers will decrease their effort from e^* as $\left|a_{j(t-1)} - a_{k(t-1)}\right|$ grows because their marginal return to effort decreases: their final ranking and reward becomes more a signal of differences in baseline student ability rather than teacher effort.

Teacher perceptions of $\gamma(\cdot)$ can either add to or reduce contest asymmetries which arise due to differences in baseline ability. If teachers believe that improving student achievement is easier (requires less effort) for students with higher levels of baseline achievement, asymmetry will be greater. However, if teachers believe that improving student achievement is easier for students with lower levels of baseline

[21] For the sake of simplicity, we have assumed that differences in a_j and a_k are the only potential sources of asymmetry in the discussion here. In reality, other factors that are not (perceived to be) evenly distributed between a teacher and her comparison teachers can introduce asymmetry and lead to deviations from efficient effort levels. A main example is differences in teacher's perceptions of their own teaching ability relative to others (Barlevy and Neal (2012)).

[22] Subject to additional assumptions concerning the seeding of the contest for teacher quality, class size and peer composition (Barlevy and Neal (2012)).

achievement, asymmetry will decrease. In other words, differences between $\widetilde{\gamma_j}(a_{k(t-1)})$ and $\widetilde{\gamma_j}(a_{j(t-1)})$ can offset asymmetry due to differences between $a_{k(t-1)}$ and $a_{j(t-1)}$. The parameter $\widetilde{\gamma_j}(\cdot)$ depends on (a.) teacher beliefs about the educational production function, specifically their perception of how teaching effort contributes to student learning for students with different levels of baseline achievement (i.e. whether the performance of initially low-achieving students responds more or less to a given level of teaching effort than high-achieving students) and (b.) their perception of how levels of learning are reflected in the assessment scale (e.g. whether there is top-coding in the test so that learning gains at the top of the distribution are not fully reflected in the test score measures).[23]

Rewarding teachers based on their ranking in terms of student gains will also generally fail to elicit efficient effort and lead teachers to supply effort that is less than that under pay-for-percentile. Although gains incentives potentially make the contest more "fair" (symmetric) compared to levels by partially adjusting for baseline levels in student achievement, asymmetry will nevertheless arise if teachers believe that improving student achievement requires more or less effort for students at different levels of initial achievement.[24] That is, with gains incentives, in which teachers are rewarded based on $a_{j,k} - a_{j,k(t-1)}$, $a_{j,k(t-1)}$ is differenced out and each teacher's first order condition becomes

$$\sum_{k \neq j} \pi \widetilde{\gamma_j}(a_{j(t-1)}) f\left(\widetilde{\gamma_j}(a_{j(t-1)})e_j - \widetilde{\gamma_j}(a_{k(t-1)})e_k\right) = c'(e_j). \quad (4)$$

The symmetry of the contest depends on teachers' perceptions of $\gamma(\cdot)$. The contest based on gains will be asymmetric as along as $\widetilde{\gamma_j}(a_{j,k(t-1)})$ is not constant (i.e. as long as it varies with $a_{(t-1)}$) and $a_{(t-1)}$ varies across classes.

Though not made explicit in this simple model, pay-for-percentile incentives may also outperform levels and gains incentives because symmetry under pay-for-percentile depends less on teacher beliefs about $\widetilde{\gamma_j}$ and the distribution of $a_{k(t-1)}$. In general, teachers may be reluctant to increase effort due to their uncertainty about these

[23] Note that there was no actual top-coding in the actual exam used in the study to assess student performance.

[24] We show evidence below (in section 3.3.1) that teachers do indeed believe that returns to their effort (in terms of a hypothetical assessment scale) are higher for students toward the bottom of the distribution.

parameters. This uncertainty is less of a factor under pay-for-percentile because teachers are compared to others with the same baseline achievement by construction.[25]

Whether gains incentives elicit more effort than levels incentives depends on the relative asymmetry due to i) differences in perceptions of $\gamma(\cdot)$ alone and ii) differences in perceptions of $\gamma(\cdot)$ and differences in $a_{j,k(t-1)}$ jointly (i.e. whether these two terms are complements or substitutes). If $\widetilde{\gamma_j}(\cdot)$ is decreasing in $a_{j,k(t-1)}$ fast enough, gains incentives could be less symmetric than levels incentives and weaker as a result. The strength of gains incentives may also be weakened if teachers recognize that gains measurements are more subject to statistical noise (Murnane and Ganimian 2014).

Although standard theory implies that the more symmetric contest under pay-for-percentile should elicit greater effort relative to levels and gains incentives, pay-for-percentile may nevertheless fail to outperform levels and gains in practice if teachers perceive pay-for-percentile incentives as relatively complex and less transparent. A growing body of research suggests that people may not respond or respond bluntly when facing complex incentives or price schedules, likely due to the greater cognitive costs of understanding complexity (Liebman and Zeckhauser 2004; Dynarski and Scott-Clayton 2006; Ito 2014; Abeler and Jäger 2015). Liebman and Zeckhauser (2004) refer to the tendency of individuals to "schmedule" – or inaccurately perceive pricing schedules when they are complex, causing individuals to respond to average rather than marginal prices, for example. If pay-for-percentile contracts are perceived as complex and rewards are not large enough to cover the (cognitive) cost of choosing an optimal response and incorporating this into their teaching practice, pay-for-percentile incentives may be ineffective. Incentive scheme complexity may also reduce perceived transparency, which may be an important factor in developing countries where trust in implementing agencies may be more limited (Muralidharan and Sundararaman 2011).

Triage

[25] This uncertainty will still matter under pay-for-percentile to the degree that i) teachers are uncertain about how other teachers' returns to effort differ from theirs for a student of a given level of baseline achievement and ii) teachers are uncertain about seeding based on student baseline achievement due to measurement error testing.

How teachers are ranked and rewarded using student achievement scores can affect not only how much effort teachers provide overall, but also how teachers allocate that effort across students (Neal and Schanzenbach 2010). The way in which the achievement scores of multiple students are used to define teacher performance can create incentives for teachers to "triage" certain students in a class at the expense of others. This is because by transforming individual student scores into a single measure, performance indexes can (implicitly or explicitly) weight some students in the classroom more than others. Teachers will allocate effort across students in the class according to costs of effort and expected marginal returns to effort given the performance index and the reward structure they face.

When teachers are ranked and rewarded according to class average levels or gains, teachers will optimally allocate effort across students in the class in order to maximize the class average score on the final exam.[26] Assuming costs of effort are similar across students, teachers will focus relatively more on students for whom the expected return to effort is highest in terms of gains on the standardized exam (until marginal returns are equalized across students). Teachers may, for instance, focus less on high-achieving students because they believe that these students' achievement gains are less likely to be measured (or rewarded) due to top-coding of the assessment scale (these students are likely to score close to full marks even without any extra instruction). Whether and how triage occurs depends on how teachers view the mapping between their own effort and student achievement scores – in particular how perceived returns to effort vary across students of different baseline achievement levels.[27]

In comparison, pay-for-percentile incentives should limit the potential for triage. This is because pay-for-percentile rewards teachers according to each student's performance in ordinal contests within their own comparison group and each of these contests are weighted equally. A teacher essentially competes in as many contests as there are students in her class that have comparison students in other schools and is

[26] This will be the same for gains and levels incentives because maximizing the average level score will, by construction, also maximize the average gain score.

[27] Teachers were not told the exact performance of each student at baseline; however, teachers own rankings of students within their class at baseline is well correlated with within-class rankings by baseline exam scores (correlation coefficient = 0.524, p-value = 0.000).

rewarded based on each student's rank in these contest independent of assessment scale. As a result, the returns to effort are more equal across students. While triage can still occur (due to differences in costs of effort across students, for example), the pay for percentile scheme should strengthen incentives for teachers to focus instruction and attention more broadly across students within a classroom.

2.4. Data Collection

Our data collection efforts entailed several survey rounds and focused on students that were 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. Controlling for two waves of baseline achievement provides additional statistical precision in our analyses. At the beginning of the 2013-2014 school year, we also 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). The baseline surveys collected information on 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 (self-concept, anxiety, intrinsic and instrumental motivation scales); the types of math problems that teachers covered with students during the school year (to assess curricular coverage across levels of difficulty); time students 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 math 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 a module designed to elicit teachers' perceived returns to teaching effort for individual students within the class (described in detail below). The teacher baseline survey took place before we provided the teachers with performance pay contracts (in October 2013). We administered a nearly identical survey to teachers again in May 2014 after the conclusion of the experiment.

Standardized Math Exams. Our primary outcome is student math achievement scores. Math achievement was measured during the endline and baseline surveys using 35-minute mathematics tests. The mathematics tests were constructed by trained psychometricians. Math test items for the endline and baseline tests 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 of the test were then validated using data from extensive pilot testing to ensure good distributional properties (no bottom or top-coding, for example).[28] In the analyses, we normalized each wave of mathematics achievement scores separately using the mean and distribution in the control group. Estimated effects are therefore expressed in standard deviations.

2.5. Balance and Attrition

Appendix Table 1 shows summary statistics and tests for balance across study arms. Due to random assignment, the characteristics of students, teachers, classes and schools are similar across the study arms. Variable-level tests for balance do not reveal more significant differences than would be expected by chance.[29] Additionally, omnibus tests across all baseline characteristics in Appendix Table 1 do not reject balance across

[28] In the endline exam, only 23 students (0.27%) received a full score and no students received a zero score.
[29] Note that teacher level characteristics in this table differ from those in our pre-analysis plan, which used 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.

the student arms.[30] Characteristics are also balanced across the incentive design arms within the small and large reward size groups.

The overall attrition rate between September 2013 and May 2014 (beginning and end of the school year of the intervention) was 5.6% in our sample.[31] Appendix Table 2 shows that there is no significant differential attrition across the incentive design treatment groups or the reward size groups in the full sample. Within the small reward group, students of teachers with a pay-for-percentile incentive were slightly less likely to attrit compared to the control group (by 2.6 percentage points, Row 3, Column 3).

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 – see Bruhn and McKenzie 2009), 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.[32] 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:

$$Y_{ijc} = \alpha + T_{jc}'\beta + X_{ijc}'\gamma + \tau_c + \varepsilon_{ijc} \quad (5)$$

[30] These tests were conducted by regressing treatment assignment on all of the baseline characteristics in Appendix Table 1 using ordered probit regressions and testing that coefficients on all characteristics were jointly zero. The p-value of this test is 0.758 for the incentive design treatments and 0.678 for the reward size treatments.

[31] Two primary schools were included in the randomization but chose not to participate in the study before the start of the trial. Baseline characteristics are balanced across study arms including and excluding these schools.

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

where Y_{ijc} is the outcome for student i in school j in county c; T_{jc} 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 the two waves of baseline achievement scores. We also estimate treatment effects with an expanded set of controls. For student-level outcomes, this includes student age, student gender, parent educational attainment, a household asset index (constructed using polychoric principal components—Kolenikov and Angeles, 2009), class size, teacher experience, and teacher base salary. We adjusted our standard errors for clustering at the school level using the cluster-corrected Huber-White estimator. For our primary estimates, we present results of significance tests that adjust for multiple testing (across all pairwise comparisons between experimental groups) using the step-down procedure of Romano and Wolf (2005) which controls the familywise error rate.

Given that the incentive designs are hypothesized to affect not only average student scores but also the distribution of scores, estimating differences in means across groups may fail to fully capture the effects of different incentive designs (Abadie 2002; Banerjee and Duflo 2009; Imbens and Rubin 2015). To examine differences in the full distributions of student outcomes we conduct Kolmogorov-Smirnov type tests as discussed in Abadie (2002) and Imbens and Rubin (2015).[33] For each pair of experimental groups, we calculate three test statistics. For two sets of scores corresponding to groups A and B, we first calculate unidirectional test statistics (in both directions) as $\sup\left(\mathrm{F}^{\mathrm{A}}(\mathrm{y}) - \mathrm{F}^{\mathrm{B}}(\mathrm{y})\right)$, where F is the cumulative density function, to test whether the distribution of scores in group A dominate those in group B. We also calculate a combined test statistic as $\sup\left|\mathrm{F}^{\mathrm{A}}(\mathrm{y}) - \mathrm{F}^{\mathrm{B}}(\mathrm{y})\right|$ to test the equality of the distributions. For inference, we cluster bootstrap test statics using 1,000 repetitions.

In addition to estimating effects on our primary outcome (year-end standardized exam scores normalized by the control group distribution), we use equation (5) 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

[33] This analysis was not pre-specified.

advance).[34] To construct these indices, we used the GLS weighting procedure described by Anderson (2008). Specifically, for each individual, we constructed a variable $\bar{s}_{ij}$ as the weighted average of k normalized outcome variables in each 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:

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

where $\mathbf{1}$ is a column vector of 1s, $\widehat{\mathbf{\Sigma}}_j^{-1}$ 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.

3. Results

In this section, we present three sets of results. First, we present results on the average impacts of the different incentives designs and reward sizes on student achievement (Section 3.1). Second, we examine impacts on the distribution of student scores (Section 3.2). Third, we present results for the average impacts of incentives on student secondary outcomes and teacher behavior (Section 3.3). Finally, we present results on the within-class distributional impacts of incentives on achievement (Section 3.4).

3.1 Average Impacts of Incentives on Achievement

The first six rows (Panel A) of Table 2 report regression estimates for the effect of the different incentive treatments (any incentive, those based on different teacher performance indices, and those based on different reward sizes) relative to the control group. As specified in our pre-analysis plan, we report estimates using Equation (5) 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

[34] Testing for impacts on summary indices instead of individual indices has several advantages (see Anderson, 2008). First, conducting tests using summary indices avoid over-rejection due to multiple hypotheses. Second, they provide a statistical test for the general effect of an underlying latent variable (that may be incompletely expressed through multiple measures). Third, they are potentially more powerful than individual tests.

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 incentive treatments.

Any incentive. First pooling all incentive treatments, we find weak evidence that having any incentive modestly increases student achievement at the endline. The specification including the expanded set of controls shows that having any incentive significantly increases student achievement by 0.074 SDs (Table 2, Panel A, Row 1, Column 2).

Teacher performance measures. Although the effect of teachers having *any incentive* is modest, the effects of the different incentive designs vary. We find that only 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 regression specification) to 0.148 SDs (in the specification with additional controls—Panel A, Row 4, Columns 3 and 4). By contrast, we find no significant effects from offering teachers levels or gains incentives based on regression estimates (Panel A, Rows 2 and 3, Columns 3 and 4).

Comparing across the incentive design treatment point estimates, pay-for-percentile significantly outperforms gains (by 0.147 SDs—Panel B, Row 15, Column 4). The point estimate for pay-for-percentile is also larger than that for levels, but the difference is not statistically significant (difference=0.064 SDs). A joint test of equality shows that the three coefficients on the incentive design treatments differ significantly from one another (p-value=0.065).

The result that pay-for-percentile outperforms gains incentives and levels incentives shows that the way the teacher performance index is defined matters independent of other design features. Moreover, these effects come at no or little added cost since monitoring costs (costs of collecting underlying assessment data) and the total amount of rewards paid are constant. Given that gains and levels are arguably much simpler schemes, these results also suggest that—at least in our context—teachers respond to relatively complex features of incentive schemes.

Small Rewards versus Large Rewards. We do not find strong evidence that larger rewards significantly outperform smaller rewards. When pooling across the incentive design treatments, the difference between large and small incentives is small and insignificant (Table 2, Columns 5 and 6). Moreover, although we find that pay-for-percentile incentives do have a larger effect (and are only significant) with larger rewards (0.16 SDs, Panel A, Row 4, Columns 9 and 10), 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 (p-value = 0.268).[35]

Taken together, these results are remarkable in that they suggest that the design of the incentive—specifically, how teachers are ranked and rewarded according to the achievement of their students—has a larger effect on student performance than doubling the size of potential rewards.

3.2 Distributional Treatment Effects of Incentive Designs

The separate incentive designs are hypothesized to affect not only average performance, but also have varying effects for low and high performers. In this section, we therefore examine differences in the full distribution of scores across the incentive design groups following Abadie (2002).

Figure 1 shows the cumulative distributions of student test performance across the experimental groups. For the full sample (Panel A), the small reward group only (Panel B), and the large reward group only (Panel C), we plot the distributions of student scores adjusted for the set of pre-specified covariates listed above.[36] The plots indicate that pay-for-percentile outperforms levels and gains incentives. In all three graphs, the distribution of scores for the pay-for-percentile group appears to stochastically dominate that of the other two incentive schemes and the control group, though differences appear larger with large rewards.

Table 3 presents results for K-S type tests between each distribution pair using the full sample. Panel A presents tests comparing each incentive design to the control group

[35] Note that the study was not ex-ante powered to test the interaction between the teacher performance index treatments and incentive size and this test was not pre-specified.
[36] These are adjusted by estimating Equation 5 without treatment dummies and saving predicted residuals.

and Panel B shows comparisons between each treatment pair. For each comparison we show results for three tests discussed in section 2.6: the two unidirectional tests and the non-directional combined test.

The results in Panel A show that the levels incentive and the pay-for-percentile incentive both outperform the control group. The p-value for whether the distribution of student scores under levels lies to the right of the distribution of student scores under no incentive is 0.077 (Table 3, Row 1). The results are stronger for pay-for-percentile; the p-value for the same test comparing pay-for-percentile to the control group is 0.018 (Table 3, Row 3). Moreover, the tests show that the distribution of scores under levels and pay-for-percentile both first-order stochastically dominate the distribution of scores in the control group. In both cases, the test statistic for the difference between the control distribution and the treatment distribution is zero, meaning that there is no point at which the cumulative density of the control distribution is larger. There is no detectable difference between the distribution of scores in the gains incentive group and that in the control group.

Tests between each incentive design group reported in Panel B show that levels incentives outperform gains incentives and pay-for-percentile incentives outperform both gains and levels incentives. The p-value for the difference between levels and gains is 0.037 (Table 3, Row 4). The p-values for the difference between pay-for-percentile and levels and gains are 0.068 (Table 3, Row 5) and 0.033 (Table 3, Row 6). In all three comparisons test statics show first-order stochastic dominance or very near first-order stochastic dominance.

3.3. 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 find. To estimate the effects, we run regressions analogous to Equation (5), but substitute endline achievement with secondary student outcomes and measures of teacher behavior.

The measures of secondary outcomes that we use were constructed as pre-specified in our analysis plan. Most of these measures (math self-concept, math anxiety, math intrinsic and instrumental motivation, student time on math, student perception of

teacher teaching practices, teacher care, teacher management of the classroom, teacher communication, and parent involvement in schoolwork, teacher self-reported effort) are indices that were created from a family of outcome variables using the GLS weighting procedure described in Anderson (2008) (see Section 2.6). These each have a mean of 0 and a SD of 1 in the sample. 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.[37] 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 to each example were averaged for all three categories together and the easy, medium, and hard categories separately.

We find that the different incentive design treatments had significant effects on teaching practice as measured by student-reported curricular coverage (Table 4, Columns 1 to 4). Pay-for-percentile also had a significant effect on curricular coverage overall (Row 3, Column 1) and this effect is larger than that of gains incentives (p-value: 0.074) and levels incentives (though not statistically significant, p-value: 0.238).[38] Compared to the control group, students in the gains group report being taught more curricula at the medium level (Row 2, Column 3); and students in the pay-for-percentile group report being taught more medium and hard curricula (Row 3, Columns 3 and 4). The effect of pay-for-percentile on the teaching of hard curricula is significantly larger than the effects of levels and gains on the teaching of hard curricula (p-value (levels): 0.022; p-value (gains): 0.001).

Although the positive impacts on curricular coverage suggest that incentivized teachers covered more of the curriculum, 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, leaving less time for students to master the subject matter. To test this, we estimate treatment effects on subsets of test

[37] Curricular coverage (or "opportunity to learn") is commonly measured in the education research literature (see, for example, Schmidt et al. 2015).
[38] Testing effects on overall curricular coverage (combining easy, medium and hard) was not included in the pre-analysis plan.

items categorized into easy, medium and hard questions (Table 4, Columns 5 to 13).[39] Test items were categorized into easy, medium and hard questions (10 items each) using the frequency of correct responses in the control group. Compared to the control group, students in classes where teachers had pay-for-percentile incentives had significantly higher scores in easy and hard difficulty categories. Pay-for-percentile incentives increased easy question sub-score by 0.105 SDs (Row 3, Column 5) and the hard question sub-score by 0.16 SDs (Row 3, Column 7). With large rewards, pay-for-percentile incentives increased the hard question sub-score by 0.191 SDs (Row 3, Column 13). By contrast, there were no significant impacts for the levels and gains incentive arms. Taken together, these results show that: 1) pay-for-percentile incentives increased both the coverage and intensity of instruction and 2) teachers with pay-for-percentile covered relatively more advanced curricula.

Despite the effects of pay-for-performance incentives on curricular coverage and intensity, we find little effect on other types of teacher behavior (Appendix Table 3). There are no statistically significant impacts from any of the incentive arms on time on math, perceptions of teacher teaching practices, teacher care, teacher management of the classroom, or teacher communication as reported by students 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 being tutored outside of class (Column 12), these do not explain the significantly larger differential impact of pay-for-percentile. In our case, it seems that pay-for-percentile incentives worked largely through increased 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

[39] Analysis of test items was not pre-specified in our analysis plan. This analysis should be considered exploratory.

(Appendix Table 3, Columns 1 to 5). There is also no evidence that any type of incentives led to increased substitution of time away from subjects other than math (Column 13).

3.4. Effects on the Within-class Distribution of Student Achievement

As discussed in the conceptual framework section (Section 2.3.4), the different incentive design treatments may affect not only how much effort that teachers provide overall, but also how they choose to allocate that effort across students within their class (or how they focus instruction). In contrast with pay-for-percentile, under levels and gains, teachers may be more likely to (initially) focus their effort more on students for whom they believe the return to effort (in terms of gains in standardized exam scores) is highest. In this section, we examine this hypothesis by first exploring teachers' perceptions of their own value-added and how this varies across students.[40] We then test how the effects of levels, gains, and pay-for-percentile incentives vary across the within-class distribution of teachers' perception of value-added for individual students and across the within-class distribution of baseline achievement.

3.4.1 Teachers' Perceptions of Own Value-added

Teachers' perceptions of their own value-added (of their "perceived value-added" for short) with respect to individual students in their class were elicited as part of the baseline survey. To elicit a measure of teacher's perceived value-added, teachers were presented with a randomly-ordered list of 12 students from their class.[41] The teachers were asked to rank the students in terms of math ability. For each student, they were then asked to give their expectation for by how much the student's achievement would improve both with and without one hour of extra personal instruction from the teacher per week.[42] A teacher's perception of their own value-added for each student is measured as

[40] This analysis was not pre-specified and should be considered exploratory.
[41] Four students were randomly selected within each tercile of the within-class baseline achievement distribution to ensure coverage across achievement levels.
[42] Precisely, for each student, teachers were asked: (a.) to rank the math achievement of the student compared to other students on the list; (b.) if this student were given curriculum-appropriate exams at the beginning and end of sixth grade, by how much would expect this student's score to change (in terms of percent of correct answers)?; and (c.) to suppose the student were given one extra hour of personal instruction from you per week. By how much would expect this student's score to change (in percent of correct answers)? A teacher's perception of their own value-added for each student is measured as the difference between (b) and (c). To standardize this measure across teachers, this difference is then

the difference between these scores, normalized by the distribution of teacher's reported expectation of gains across students.[43]

Table 5 shows how this measure of teachers' perceived value-added varies across students within the class. This table shows coefficients from regressions of our measure of teachers' perceived value-added for each student on students within-class percentile ranking by math ability at baseline and other student characteristics (gender, age, parent educational attainment, and a household asset index), controlling for teacher fixed effects. We estimate these regressions using two measures of students' within-class ranking: a.) the rank provided by the teacher in the baseline survey and b.) the rank of student performance on the standardized baseline exam.

This analysis yields two findings of note. First, on average, teachers' perceived value-added declines with students' improved ranking within the class (Table 5, Row 1). This result is consistent with both measures of within-class percentile rank (either using teacher's own ranking (Columns 1 and 2) or the ranking based on the baseline exam (Columns 5 and 6)). Examining how perceptions vary across terciles of the within-class distribution, however, shows that teachers' perceived value-added is similar for students in the bottom two terciles but are significantly lower for students at the top of the distribution (Columns 3-4 and 7-8). Teachers' perceived value-added is approximately 0.2 SD lower for students in the top third of the distribution compared to the bottom third based on their own ranking of their students. This result does, however, mask a great deal of heterogeneity in teacher perceptions of for what type of students their value-added is the lowest and highest. Forty-three percent of teachers report the lowest perceived returns for students in the top tercile, 31 percent report the lowest returns for the bottom tercile and 17 percent the lowest returns for the middle tercile. Teachers were nearly evenly split in reporting highest returns for the bottom, middle and top of the distribution.

---

normalized by the within-class distribution of (c) (normalizing by the distribution of (b) produces similar results). No information other than student names and gender was presented to teachers.

[43] Admittedly, this measure is not ideal in that it reflects perceived returns to personal tutoring time whereas, given the results above on curricular coverage, we may be more interested in how returns differ from tailoring classroom instruction. Moreover, this is only a measure of the perceived returns to an initial unit of "extra" effort and does not provide information on how teachers think returns change marginally as more effort is directed toward a particular student. Nevertheless, this measure should serve as a reasonable proxy for teachers' perceptions of how returns vary more generally across students. It was also deemed that attempting to measure perceived returns to subsequent units of effort directed toward a particular student would introduce too much noise into the measure.

Second, teachers' perceived value-added is not significantly related to any other student characteristics once student ranking within the class is accounted for. This suggests that teachers in our sample may think about returns primarily as a function of initial ability.

3.4.2 Within-class Distributional Effects of Incentives

Table 6 shows estimates of how the effects of levels, gains, and pay-for-percentile incentives on endline student achievement vary with teacher's perceived value-added and with the within-class ranking of students in terms of initial math ability/achievement. Our goal is to understand how teachers allocate effort across students in response to incentives (i.e. whether teachers triage some students at the expense of others) and how this allocation of effort affects students at different parts of the initial distribution of achievement. To do this, we estimate heterogeneous effects along three different variables: teachers' perceived value-added at the student level, teachers ranking of students by math ability, and the within-class ranking of students using performance on baseline standardized exams. We estimate effects by tercile of the distribution for each of these variables by estimating Equation (5) but including dummy variables for the middle and top terciles and interactions with indicators for the levels, gains, and pay-for-percentile incentive arms. All regressions are estimated with and without the pre-specified expanded set of control variables.

We find that the effects of levels and gains incentives are significantly higher among students for whom teachers had the highest perceived value-added, but the effects of pay-for-percentile do not vary significantly with perceived value-added (Columns 1 and 2). For students in the top tercile of teacher's perceived value-added, levels incentives had an approximately 0.2 SD larger effect than on students in the bottom tercile and gains incentives had an approximately 0.3 SD larger effect than on students in the bottom tercile (although total effects of incentives on these students is not significantly positive in either case).[44] We do note however that these results should be

[44] The coefficient on the interaction term between the top tercile of perceived value added and pay-for-percentile incentives in these regressions, however, is not statistically different from the coefficients on the interactions terms between the top tercile and levels incentives (p-value=0.224) or gains incentives (p-value=0.121).

interpreted somewhat cautiously as our power for detecting effects on exam scores is reduced using the subsample of students for whom we have measures for teachers perceived value-added.

Assuming that these effects on endline achievement reflect teachers' allocation of effort across students (or their focus of classroom instruction), these results are consistent with teachers responding to levels and gains incentives by focusing relatively more on students with the highest returns to teacher effort in terms of exam score gains. They also suggest that pay-for-percentile does lead to a more equal allocation of teacher effort across students.

Although the effects of incentives seem to vary with teacher's perceptions of value-added, we do not find any evidence that the effects of incentives vary significantly along the distribution of within-class baseline achievement (Columns 3 to 6). Levels and gains incentives do not have significant effects for students at any part of the baseline distribution. Columns 5 and 6 show that pay-for-percentile incentives, however, led to broad-based gains for students along the within-class distribution of initial achievement. Given the correlation between teacher perceptions of value-added and the within class ranking of student by initial ability, one may anticipate levels and gains incentives having a positive effect on students at the bottom of the distribution. It appears, however, that this effect was muted on average in the sample due to the large amount of heterogeneity in teachers' perceived returns.

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, we do not find a significant difference in the effects of small and large rewards (double the size), either pooling across incentive design treatments or within each incentive design individually. Although the effect of pay-for-percentile is larger with large rewards than with smaller rewards, the difference is not significant. Third, we find evidence that teachers focus on students for whom they perceive their effort has the highest value added in terms of exam scores gains under levels and gains incentives, but not under pay for percentile. This result is consistent with the way in which pay-for-percentile rewards teachers more equally for gains across students.

There are several caveats to our findings. Most importantly, we only study the effects of incentives over one year. It is possible that impacts could change as teachers become accustomed to incentive schemes. However, it seems unlikely that the ordering of effects we observe would change in subsequent periods for two reasons. First, if the dynamic effects of incentives are affected by how well realized rewards reflect teacher effort, the effects of pay-for-percentile are more likely to improve and less likely to diminish than those of levels and gains incentives. Second, any negative effects due to lack of transparency or trust in the implementing agency could diminish in subsequent periods. If these negative effects are larger for pay-for-percentile, performance may improve relative to levels and gains incentives over time. Moreover, an additional potential benefit of pay-for-percentile incentives that we are unable to explore is that incentives can be linked to different student assessments over time (Barlevy and Neal 2012). If teachers have no advanced knowledge of which assessment will be used, pay-for-percentile may be less likely to create incentives for teachers to teach to a particular test.

A second caveat is that our study was not powered to ex-ante to study the interaction between different incentive designs individually and reward size. While our study design allows for testing these interaction effects, we only powered the study to test between incentive designs and between reward sizes separately. Future studies explicitly powered to test the complementarity between incentive design and reward size would be useful. Third, as with most empirical studies, results will not necessarily hold in 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. Finally, the version of the pay-for-percentile scheme we used did not adjust for other factors, such as teacher ability. It is possible that the effect of pay-for-percentile could be improved further as more data are available to increase the symmetry of contests by adjusting for additional differences across teachers.

Despite these caveats, we believe that these results clearly demonstrate that the design of teacher incentives 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 different design features interact.

References

Abadie, A., 2002. Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models. Journal of the American Statistical Association 97, 284–292.

Abeler, J., Jäger, S. 2015. "Complex Tax Incentives." *American Economic Journal: Economic Policy*,7(3): 1–28.

Anderson, M. L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association*, *103*(484): 1481–1495.

Ashraf, N., Bandiera, O., Jack, B.K., 2014. "No Margin, No Mission? A Field Experiment on Incentives for Public Service Delivery." *Journal of Public Economics* 120:1–17.

Baker, G.P., 1992. "Incentive Contracts and Performance Measurement." *Journal of Political Economy* 100(3): 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(3): 917–962.

Banerjee, A., Duflo, E., 2006. "Addressing Absence." *The Journal of Economic Perspectives* 20(1): 117–132.

Banerjee, A.V., Duflo, E., 2009. The Experimental Approach to Development Economics. Annual Review of Economics 1, 151–178.

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.

Barrera-Osorio, Felipe; Raju, Dhushyanth. 2015. "Teacher Performance Pay: Experimental Evidence from Pakistan." Impact Evaluation Series,Washington, D.C.: World Bank Group Policy Research Working Paper 7307.

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(2): 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(4): 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(2): 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.

Contreras, D., Rau, T., 2012. "Tournament Incentives for Teachers: Evidence from a Scaled-Up Intervention in Chile." *Economic Development and Cultural Change* 61(1): 219–246.

de Ree, J., Muralidharan, K., Pradhan, M., & Rogers, H. 2015. "Double for nothing? The Effect of Unconditional Teachers' Salary Increases on Performance." National Bureau of Economic Research Working Paper No. 21806.

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(4): 696–727.

Duflo, E., Hanna, R., Ryan, S. 2012. "Incentives Work: Getting Teachers to Come to School." American Economic Review 102(4): 1241–1278.

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(5): 1739–1774.

Dynarski, S., Scott-Clayton, J., 2006. "The Cost Of Complexity In Federal Student Aid: Lessons From Optimal Tax Theory And Behavioral Economics," *National Tax Journal* 59(2): 319-356.

Freeman, R.B., Gelber, A.M. 2010. "Prize Structure and Information in Tournaments: Experimental Evidence." *American Economic Journal: Applied Economics* 2(1): 149–164.

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." National Bureau of Economic Research Working Paper No. 18237.

Glewwe, P., Ilias, N., & Kremer, M. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2(3): 205–227.

Hanushek, E.A., Rivkin, S.G., 2010. "Generalizations about Using Value-added Measures of Teacher Quality." *The American Economic Review* 100(2): 267–271.

Hanushek, E.A., Woessmann, L., 2011. "Overview of the Symposium on Performance Pay for Teachers." *Economics of Education Review* 30(3): 391–393.

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.

Imbens, G.W. and Rubin, D.B., 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

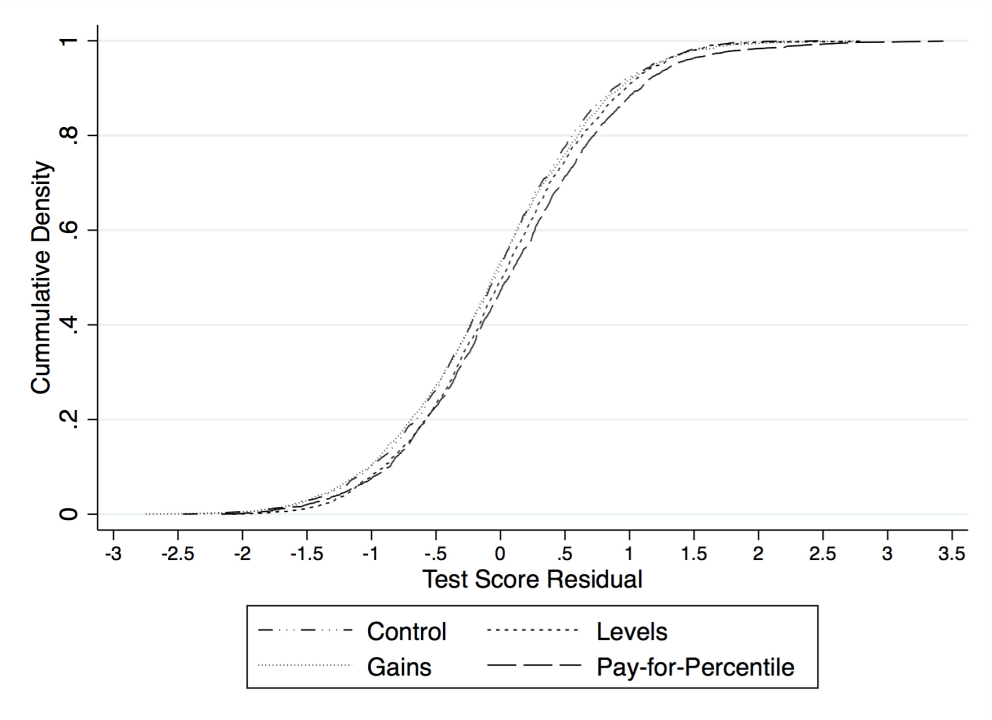
Ito, K., 2014. "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing." *The American Economic Review* 104(2): 537–563.

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.

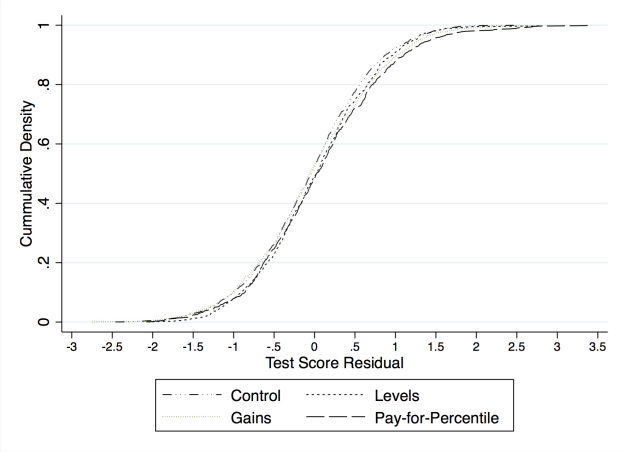
Kolenikov, S., & Angeles, G. 2009. "Socioeconomic Status Measurement with Discrete Proxy Variables: Is Principal Component Analysis a Reliable Answer?" *Review of Income and Wealth 55*(1): 128–165.
Kremer, M., Chaudhury, N., Rogers, F.H., Muralidharan, K., Hammer, J., 2005. "Teacher Absence in India: A Snapshot." *Journal of the European Economic Association* 3(2-3): 658–667.
Lavy, V., 2002. "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement." *Journal of Political Economy* 110(6): 1286–1317.
Lavy, V., 2009. "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics." *American Economic Review* 99(5): 1979–2011.
Lavy, V., 2015. "Teachers' Pay for Performance in the Long-Run: Effects on Students' Educational and Labor Market Outcomes in Adulthood" National Bureau of Economic Research Working Paper No. 20983.
Lazear, E.P., 2003. "Teacher Incentives." *Swedish Economic Policy Review* 10(2): 179–214.
Leigh, A., 2013. "The Economics and Politics of Teacher Merit Pay." *CESifo Economic Studies* 59(1): 1–33.
Liebman, J., Zeckhauser, R. 2004. "Schmeduling." Harvard University, Unpublished Manuscript.
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," National Bureau of Economic Research Working Paper No. 21302.
Moldovanu, B., Sela, A. 2001. "The Optimal Allocation of Prizes in Contests." *The American Economic Review* 91(3): 542–558.
Muralidharan, K. 2012. "Long-Term Effects of Teacher Performance Pay: Experimental Evidence from India." Unpublished Manuscript.
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." National Bureau of Economic Research Working Paper 20284.
National Bureau of Statistics of China. 2014. China Statistical Yearbook 2014. China Statistics Press: Beijing.
Neal, D. 2011. "The Design of Performance Pay in Education*."* Handbook of Economics of Education 4: 495–548.
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.
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): 909–949.
Rivkin, S.G., Hanushek, E.A., Kain, J.F., 2005. "Teachers, Schools, and Academic Achievement." *Econometrica 73(2):* 417–458.

Romano, J.P., Wolf, M., 2005. “Stepwise Multiple Testing as Formalized Data Snooping.” Econometrica 73, 1237–1282.
Schmidt, W. H., Burroughs, N. A., Zoido, P., & Houang, R. T. 2015. “The Role of Schooling in Perpetuating Educational Inequality An International Perspective.” *Educational Researcher 44*(7): 371-386.
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 (POINT).” Society for Research on Educational Effectiveness.
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(3): 404–418.

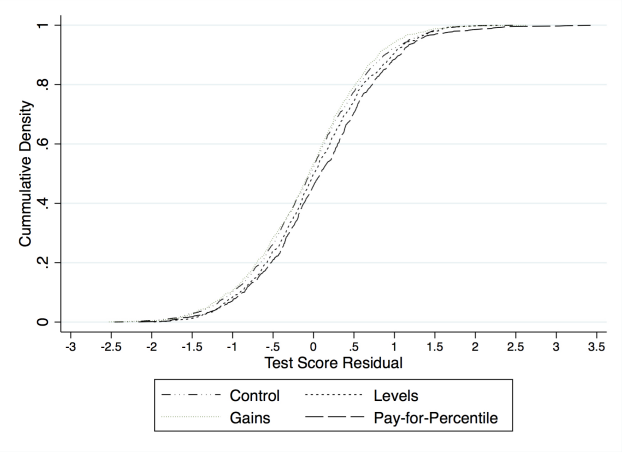
Figures and Tables



Panel A: Full Sample



Panel B: Small Reward Groups Only



Panel C: Large Reward Groups Only

Figure 1: Distribution of Test Scores across Groups

Notes: Figure shows estimated cumulative density functions of adjusted student scores across incentive treatment arms for the full sample, small reward schools only, and large reward schools only.

Table 1: Experimental Design

			Total
Control Group	52 (2,254)		52 (2,254)
	Reward Size Groups:		
Incentive Design Groups:	Large Reward	Small Reward	
Levels Incentive	26 (1,099)	28 (1,134)	54 (2,233)
Gains Incentive	26 (1,360)	30 (1,095)	56 (2,455)
Pay-for-percentile Incentive	26 (1,006)	28 (1,124)	54 (2,130)
Total	78 (3,465)	86 (3,353)	

Notes: Table shows the distribution of schools (students) across experimental groups. Note that the numbers of schools across treatments are unequal due to the number of schools available per county (strata) not being evenly divisible.

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)	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)
(3)	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)
(4)	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)
(5)	Small Reward					0.063	0.081				
						(0.053)	(0.055)				
(6)	Large Reward					0.064	0.067				
						(0.045)	(0.046)				
(7)	Additional Controls		X		X		X		X		X
(8)	Observations	7454	7373	7454	7373	7454	7373	4655	4609	4678	4628
Panel B. Comparisons Between Incentive Treatments											
(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
(17)	Large - Small					0.001	-0.014				
(18)	P-value: Large - Small					0.989	0.778				

Notes. Rows (1) to (6) (Panel A) show estimated coefficients and standard errors (in parentheses) obtained by estimating Equation 5. Standard errors account within schools The dependent variable in each regression is student standardized exam scores at endline normalized by the distribution in the control group. Each regression controls for two waves of baseline standardized math exam scores and strata (county) fixed effects. Additional control variables (included 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 (unadjusted) p-values. Significance stars indicate significance after adjusting for multiple hypotheses using the step-down procedure of Romano and Wolf (2005) to control the familywise error rate (FWER).

** Significant at the 5 percent level after adjusting for multiple hypotheses.

*Significant at the 10 percent level after adjusting for multiple hypotheses.

Table 3: Tests for Distributional Treatment Effects

		Test Statistic	*P-value*
	Test	(1)	(2)
Panel A. Relative to Control Group			
(1) Levels Incentive	$\sup\left(\mathbb{F}^{Levels} - \mathbb{F}^{Control}\right)$	0.036	0.077
	$\sup\left(\mathbb{F}^{Control} - \mathbb{F}^{Levels}\right)$	0.000	0.976
	$\sup\left\lvert\mathbb{F}^{Levels} - \mathbb{F}^{Control}\right\rvert$	0.036	0.045
(2) Gains Incentive	$\sup\left(\mathbb{F}^{Gains} - \mathbb{F}^{Control}\right)$	0.024	0.258
	$\sup\left(\mathbb{F}^{Control} - \mathbb{F}^{Gains}\right)$	0.024	0.188
	$\sup\left\lvert\mathbb{F}^{Gains} - \mathbb{F}^{Control}\right\rvert$	0.024	0.131
(3) Pay-for-Percentile Incentive (P4P)	$\sup\left(\mathbb{F}^{P4P} - \mathbb{F}^{Control}\right)$	0.071	0.018
	$\sup\left(\mathbb{F}^{Control} - \mathbb{F}^{P4P}\right)$	0.000	1.000
	$\sup\left\lvert\mathbb{F}^{P4P} - \mathbb{F}^{Control}\right\rvert$	0.071	0.013
Panel B. Between Incentive Treatments			
(4) Levels - Gains	$\sup\left(\mathbb{F}^{Levels} - \mathbb{F}^{Gains}\right)$	0.042	0.037
	$\sup\left(\mathbb{F}^{Gains} - \mathbb{F}^{Levels}\right)$	0.008	0.622
	$\sup\left\lvert\mathbb{F}^{Levels} - \mathbb{F}^{Gains}\right\rvert$	0.042	0.013
(5) P4P - Levels	$\sup\left(\mathbb{F}^{P4P} - \mathbb{F}^{Levels}\right)$	0.048	0.068
	$\sup\left(\mathbb{F}^{Levels} - \mathbb{F}^{P4P}\right)$	0.008	0.499
	$\sup\left\lvert\mathbb{F}^{P4P} - \mathbb{F}^{Levels}\right\rvert$	0.048	0.043
(6) P4P - Gains	$\sup\left(\mathbb{F}^{P4P} - \mathbb{F}^{Gains}\right)$	0.056	0.033
	$\sup\left(\mathbb{F}^{Gains} - \mathbb{F}^{P4P}\right)$	0.000	1.000
	$\sup\left\lvert\mathbb{F}^{P4P} - \mathbb{F}^{Gains}\right\rvert$	0.056	0.023

Notes. Panel A shows test statistics and p-values from Kolmogrov-Smirnov tests between the distribution of endline exam scores adjusted for baseline scores and strata fixed effects in each treatment group and the control group following Abadie (2002). Panel B shows test statistics and p-values from tests between treatment group pairs. P-values are calculated based on the distribution of 1,000 cluster bootstrap repetitions of the test statistic. The first two tests in each row are unidirectional tests that the values of exam scores in one group are larger (smaller) those in the other group. The third test is a combined test evaluating the equality of the distributions.

Table 4: Impacts on Question Difficulty Subscores and Curricula Coverage

		Curricular Coverage Full Sample				Difficulty Subscores Full Sample			Difficulty Subscores Small Reward Groups Only			Difficulty Subscores Large Reward Groups Only		
		Overall	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1)	Levels Incentive	0.015	0.019	0.020	0.005	0.029	0.094	0.075	0.039	0.074	0.076	0.013	0.107	0.066
		(0.010)	(0.012)	(0.010)	(0.015)	(0.044)	(0.050)	(0.052)	(0.062)	(0.060)	(0.066)	(0.049)	(0.057)	(0.062)
(2)	Gains Incentive	0.008	0.012	0.022*	-0.009	-0.006	-0.010	0.019	0.011	0.041	0.035	-0.019	-0.055	-0.002
		(0.009)	(0.012)	(0.010)	(0.014)	(0.036)	(0.050)	(0.053)	(0.037)	(0.061)	(0.070)	(0.050)	(0.060)	(0.061)
(3)	Pay-for-Percentile Incentive	0.027**	0.016	0.025*	0.040**	0.105**	0.092	0.160**	0.113	0.074	0.131	0.105	0.104	0.191**
		(0.011)	(0.012)	(0.011)	(0.014)	(0.043)	(0.062)	(0.067)	(0.061)	(0.097)	(0.103)	(0.048)	(0.055)	(0.065)
(4)	Observations	7373	7373	7370	7366	7373	7373	7373	4609	4609	4609	4628	4628	4628

Notes. Rows (1) to (3) show estimated coefficients and standard errors (in parentheses) obtained by estimating regressions analogous Equation 5. Standard errors account for clustering at the school level. The dependent variables in columns (1) to (4) are measures of curricular coverage (for all, easy, medium, and hard items) as reported by students. The dependent variables in columns (5) to (13) are endline exam subscores (for easy, medium and hard items) normalized by the distribution of control group scores. 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. Significance stars indicate significance after adjusting for multiple hypotheses using the step-down procedure of Romano and Wolf (2005) to control the familywise error rate (FWER).
** Significant at the 5 percent level after adjusting for multiple hypotheses.
* Significant at the 10 percent level after adjusting for multiple hypotheses.

Table 5: Correlation between Teacher Perception of Own Value-added and Student Characteristics

		Dependent Variable: Teacher Percieved Value Added							
Within-class Student Ranking used (Rows 1-3):		Teacher's Own Ranking of Students at Baseline				Ranking of Students by Baseline Exam Score			
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1)	Student Within-class Percentile Rank	-0.329***	-0.317***			-0.171*	-0.186**		
		(0.103)	(0.104)			(0.091)	(0.094)		
(2)	Student in Middle Tercile of Class (0/1)			-0.065	-0.053			-0.034	-0.045
				(0.052)	(0.053)			(0.046)	(0.047)
(3)	Student Top Tercile of Class (0/1)			-0.206***	-0.193***			-0.106*	-0.117*
				(0.071)	(0.071)			(0.062)	(0.064)
(4)	Female (0/1)		-0.032		-0.033		-0.044		-0.042
			(0.045)		(0.045)		(0.047)		(0.046)
(5)	Age (Years)		-0.026		-0.020		-0.019		-0.016
			(0.025)		(0.025)		(0.026)		(0.025)
(6)	Father Attended Secondary School (0/1)		-0.054		-0.058		-0.061		-0.062
			(0.049)		(0.049)		(0.049)		(0.050)
(7)	Mother Attended Secondary School (0/1)		-0.025		-0.027		-0.029		-0.030
			(0.039)		(0.039)		(0.039)		(0.038)
(8)	Household Asset Index		-0.019		-0.019		-0.019		-0.020
			(0.018)		(0.018)		(0.018)		(0.018)
(9)	Observations	2444	2347	2444	2347	2444	2347	2444	2347

Notes. Rows (1) to (8) show coefficients and standard errors (in parentheses) from regressions of teacher perceptions of their own value added at the student level on student characteristics at baseline. Teachers' perceptions of value added were measured as follows: During the baseline teacher survey (prior to random assignment) teachers were presented with a randomly-ordered list of 12 students randomly selected from a list of the students in their class. The selection of students to be included in the list was stratified by their performance on baseline exams. For each student on the list, teachers were asked (a.) to provide a rank based on ability in math among the students on the list, (b.) if this student were given an exam at the beginning of the school year and the end of the school year covering the sixth-grade curriculum, by how much would expect this student's score to change (in percent of correct answers)? (c.) Suppose this student were given one extra hour of personal instruction from you per week. What would you expect this student to score?. A teacher's perception of their own value added for each student is measured as the difference between (b) and (c), normalized by the distribution of (c). Teachers were provided no information on each student other than the student's name. In Columns (1) to (4) this measure of teachers' perception of value added is regressed on each student's within-class ranking (Rows 1-3) as provided by the teacher in question (a.). In Columns (5) to (8), Rows (1) to (3) are students' within-class ranking according to their performance on the baseline standardized exams. Each regression also controls for teacher fixed effects. Standard errors are clustered at the class level.

*** Significant at the 1 percent level

** Significant at the 5 percent level

* Significant at the 10 percent level.

Table 6: Within-class Distributional Effects

	Baseline Variable (VAR):	Teacher Perception of Own Value Added for Student		Teacher Ranking of Students at Baseline		Ranking of Students by Baseline Exam Score	
		(1)	(2)	(3)	(4)	(5)	(6)
(1)	Levels Incentive	-0.124	-0.133	0.051	0.053	0.092	0.091
		(0.087)	(0.087)	(0.082)	(0.083)	(0.056)	(0.058)
(2)	Gains Incentive	-0.185	-0.185	0.010	0.017	0.036	0.055
		(0.114)	(0.114)	(0.091)	(0.093)	(0.059)	(0.061)
(3)	Pay-for-Percentile Incentive	-0.020	-0.031	0.070	0.083	0.171**	0.174**
		(0.112)	(0.118)	(0.090)	(0.093)	(0.084)	(0.083)
(4)	VAR (Middle Tercile)	-0.077	-0.088	0.148*	0.136*	-0.179***	-0.176***
		(0.082)	(0.081)	(0.079)	(0.082)	(0.050)	(0.050)
(5)	VAR (Top Tercile)	-0.213**	-0.237**	0.424***	0.411***	-0.056	-0.056
		(0.096)	(0.096)	(0.079)	(0.081)	(0.068)	(0.068)
(6)	Levels X VAR (Middle Tercile)	0.053	0.066	-0.050	-0.042	-0.026	-0.009
		(0.111)	(0.110)	(0.100)	(0.102)	(0.059)	(0.060)
(7)	Levels X VAR (Top Tercile)	0.213*	0.262**	-0.091	-0.062	-0.071	-0.067
		(0.122)	(0.122)	(0.107)	(0.107)	(0.060)	(0.062)
(8)	Gains X VAR (Middle Tercile)	0.163	0.158	0.051	0.055	-0.031	-0.045
		(0.146)	(0.143)	(0.107)	(0.109)	(0.059)	(0.060)
(9)	Gains X VAR (Top Tercile)	0.333**	0.354**	-0.090	-0.091	-0.041	-0.060
		(0.152)	(0.151)	(0.113)	(0.113)	(0.064)	(0.065)
(10)	Pay-for-Percentile X VAR (Middle Tercile)	0.056	0.078	-0.022	-0.026	-0.055	-0.047
		(0.139)	(0.144)	(0.108)	(0.108)	(0.065)	(0.065)
(11)	Pay-for-Percentile X VAR (Top Tercile)	0.056	0.086	-0.069	-0.081	-0.063	-0.066
		(0.151)	(0.155)	(0.115)	(0.114)	(0.082)	(0.083)
(12)	Additional Controls		X		X		X
(13)	N	2238	2217	2415	2392	7454	7373

Notes. Rows (1) to (11) show estimated coefficients and standard errors (in parentheses) obtained by estimating regressions analogous Equation 5 but including the baseline variables listed at the top of the table and interactions with treatment arm indicators. The dependent variable in each regression is endline standardized math exam scores normalized by the distribution of control group scores. Each regression controls for two waves of baseline standardized math exam scores and strata (county) fixed effects. Additional control variables (included in even numbered columns) include student gender, age, parent educational attainment, a household asset index, class size, teacher experience and teacher base salary. See notes to Table 5 and text for a description of how teacher perceptions of value added were measured. All standard errors account for clustering at the school level.

*** Significant at the 1 percent level

** Significant at the 5 percent level

* Significant at the 10 percent level.

Appendix Table 1: Descriptive Statistics and Balance Check

			Coefficient (standard error) on:				Coefficient (standard error) on:			
		Control Mean (1)	Levels Incentive (2)	Gains Incentive (3)	Pay-for-Percentile Incentive (4)	Joint Test P-value: All=0 (5)	Small Incentive (6)	Large Incentive (7)	Joint Test P-value (8)	Obs. (9)
Panel A. Student Characteristics										
(1)	Standardized Math Test Score, Beginning of Previous School Year	0.00	-0.045 (0.084)	-0.015 (0.082)	-0.094 (0.093)	0.739	-0.040 (0.079)	-0.061 (0.080)	0.751	7996
(2)	Standardized Math Test Score, End of Previous School Year	0.00	-0.005 (0.082)	0.028 (0.091)	-0.038 (0.088)	0.894	0.015 (0.080)	-0.023 (0.081)	0.848	8136
(3)	Female (0/1)	0.49	-0.010 (0.017)	-0.002 (0.015)	-0.011 (0.018)	0.893	-0.005 (0.015)	-0.010 (0.015)	0.816	7996
(4)	Age (Years)	11.99	0.088 (0.063)	0.137** (0.066)	0.082 (0.072)	0.225	0.104* (0.062)	0.103* (0.061)	0.176	7992
(5)	Father Attended Secondary School (0/1)	0.52	0.005 (0.024)	0.028 (0.026)	0.005 (0.026)	0.686	0.007 (0.023)	0.019 (0.023)	0.700	7965
(6)	Mother Attended Secondary School (0/1)	0.31	0.010 (0.026)	0.019 (0.026)	0.011 (0.026)	0.900	0.021 (0.024)	0.007 (0.023)	0.660	7929
(7)	Household Asset Index	-0.64	0.025 (0.046)	0.014 (0.048)	0.041 (0.050)	0.865	-0.001 (0.042)	0.054 (0.042)	0.348	7996
Panel B. Teacher and Class Characteristics										
(8)	Age (Years)	32.62	1.671 (1.599)	0.367 (1.682)	0.581 (1.473)	0.745	0.305 (1.347)	1.548 (1.572)	0.549	243
(9)	Female	0.42	-0.019 (0.091)	0.095 (0.089)	-0.013 (0.093)	0.492	0.012 (0.082)	0.031 (0.087)	0.933	243
(10)	Han (0/1)	0.95	0.010 (0.034)	-0.062* (0.035)	-0.014 (0.027)	0.229	-0.042* (0.024)	0.003 (0.034)	0.134	243
(11)	Teaching Experience (Years)	11.61	1.858 (1.772)	0.844 (1.994)	-0.167 (1.630)	0.617	0.477 (1.509)	1.224 (1.808)	0.789	243
(12)	Monthly Base Salary (Yuan)	2852.77	255.599* (152.651)	-149.432 (187.318)	142.402 (175.438)	0.054	119.440 (161.684)	37.325 (160.419)	0.713	243
Panel C. School Characteristics										
(13)	Number of Students in Grade Six	43.35	-1.154 (2.877)	2.407 (2.971)	-3.430 (2.819)	0.300	-2.296 (2.615)	1.089 (2.581)	0.416	216
(14)	Number of Students in School	437.83	-59.555 (62.562)	-31.874 (60.861)	-46.852 (65.916)	0.807	-71.814 (58.522)	-16.537 (60.857)	0.270	216
(15)	Number of Teachers	29.75	-0.447 (4.234)	-2.744 (3.692)	-0.979 (4.223)	0.859	-3.531 (3.488)	1.029 (3.996)	0.235	216
(16)	Number of Contract Teachers	1.69	0.403 (0.645)	0.073 (0.388)	0.063 (0.415)	0.937	0.116 (0.380)	0.248 (0.501)	0.884	216

Notes. Data source: baseline survey. The first column shows the mean in the control group. Panel A shows student-level characteristics, Panel B shows teacher and class characteristics and Panel C shows school level characteristics. Exam scores are normalized usingthe distribution in the control group. Columns 2-4 and 6-7 show coefficients and standard errors (in parentheses) 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 Wald test that preceding coefficients are jointly zero. Test account for clustering at the school level.
*** Significant at the 1 percent level; ** Significant at the 5 percent level; * Significant at the 10 percent level.

Appendix Table 2: Attrition

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

Notes. Table shows estimated coefficients and standard errors from a regression of a dummy variable indicating that a student was absent from the endline survey on indicators for incentive treatments and controlling for randomization strata Standard errors in parentheses account for clustering at the school level.
***Significant at the 1 percent level.
** Significant at the 5 percent level.
* Significant at the 10 percent level.

Appendix Table 3: Impacts on Secondary Outcomes

Dependent Variable:		Math Self Concept (1)	Math Anxiety (2)	Math Intrinsic Motiva -tion (3)	Math Instrumental Motiva -tion (4)	Student Time on Math (5)	Student Perception of Teaching Practices (6)	Teacher Care (7)	Teacher Classroom Manage -ment (8)	Teacher Commu- nication (9)	Parent Involve ment (10)	Teacher Self-reported Effort (11)	Out-of- Class Tutoring (12)	Time spent Studying other subjects (13)
(1)	Levels Incentive	0.023	0.009	0.029	-0.042	0.031	0.014	0.034	-0.004	-0.029	-0.059	0.055	0.149*	-0.010
		(0.040)	(0.039)	(0.056)	(0.046)	(0.056)	(0.040)	(0.063)	(0.049)	(0.055)	(0.049)	(0.078)	(0.076)	(0.030)
(2)	Gains Incentive	0.012	0.024	0.093*	0.022	0.008	0.022	-0.003	0.001	0.043	0.062	0.003	0.136*	-0.014
		(0.039)	(0.034)	(0.054)	(0.039)	(0.055)	(0.036)	(0.066)	(0.052)	(0.048)	(0.046)	(0.075)	(0.070)	(0.033)
(3)	Pay-for-Percentile Incentive	-0.011	-0.009	0.083	0.065	-0.001	0.040	-0.005	0.036	0.071	0.024	-0.024	0.118*	-0.032
		(0.043)	(0.040)	(0.063)	(0.047)	(0.054)	(0.045)	(0.073)	(0.055)	(0.067)	(0.048)	(0.076)	(0.070)	(0.034)
(4)	Observations	7373	7373	7373	7373	7373	7373	7372	7373	7373	7371	235	7368	7373

Note. Rows (1) to (3) show estimated coefficients and standard errors (in parentheses) obtained by estimating regressions analogous Equation 5. Standard errors account for clustering at the school level. Outcome variables in columns (1) to (11) are summary indices. Summary indices 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 regression reported in column (11) is at the teacher level. Significance stars indicate significance after adjusting for multiple hypotheses using the step-down procedure of Romano and Wolf (2005) to control the familywise error rate (FWER).

** Significant at the 5 percent level after adjusting for multiple hypotheses.

* Significant at the10 percent level after adjusting for multiple hypotheses.