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

Prashant Loyalka, *Stanford University*

Sean Sylvia, *University of North Carolina at Chapel Hill*

Chengfang Liu, *Peking University*

James Chu, *Stanford University*

Yaojiang Shi, *Shaanxi Normal University*

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 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 paying close attention to performance pay design.

We are grateful to Grant Miller, Karthik Muralidharan, Derek Neal, Scott Rozelle, Marcos Vera-Hernández, Justin Trogdon, and Rob Fairlie for helpful comments on the manuscript and to Jingchun Nie for research assistance. We also 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 the Xu Family Foundation for financing the project. Contact the corresponding author,

[*Journal of Labor Economics*, 2019, vol. 37, no. 3]

Submitted February 9, 2017; Accepted February 21, 2018; Electronically published May 2, 2019

I. Introduction

Performance pay schemes linking teacher pay directly to student achievement are now a common approach to better align teacher incentives with student learning (OECD 2009; Bruns, Filmer, and Patrinos 2011; Hanushek and Woessmann 2011; Woessmann 2011). Whether performance pay schemes can improve student outcomes, however, is likely to depend critically on their design (Bruns, Filmer, and Patrinos 2011; Neal 2011; Pham, Nguyen, and Springer 2017). Schemes that fail to closely link rewards to productive teacher effort may be ineffective (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). While studies have highlighted weaknesses in specific design features of performance pay schemes, many important aspects of design have yet to be explored empirically.[1]

We study incentive design directly by comparing performance pay schemes that vary in how student achievement is used to measure teacher performance. How student achievement scores are used to measure teacher performance can, independently of the underlying contract structure 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, Filmer, and Patrinos 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.

Sean Sylvia, at sean_sylvia@unc.edu. Information concerning access to the data used in this paper is available as supplemental material online.

[1] Important exceptions are Fryer et al. (2012), who compare incentives designed to exploit loss aversion with a more traditional incentive scheme, and Imberman and Lovenheim (2014), who examine the impact of incentive strength as proxied by the share of students a teacher instructs. 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 (2017), 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. Finally, Neal (2011) considers theory in incentive design while reviewing the effectiveness of teacher performance pay programs in the United States.

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). To isolate the effect of different ways student achievement is used to rank teachers and to compare these as budget-neutral alternatives, the distribution of rank-order tournament payouts within the small- and large-reward treatments was common across the incentive design schemes.

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 that pay-for-percentile incentives significantly outperformed both gains and levels incentives, while levels incentives outperformed gains incentives. Achievement gains under pay-for-percentile incentives were mirrored by meaningful increases in the intensity of teaching, as evidenced by teachers covering more material, teachers 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 1 month of salary to 2 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 suggest that in our context, how teacher performance is measured has a larger effect on student achievement 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 dis-

tribution. Pay-for-percentile incentives, by contrast, led to broad-based gains along the distribution.

Beyond providing more evidence on the effectiveness of incentives generally, we contribute to the teacher performance pay literature in three ways.[2] 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.[3] Only two studies provide direct 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 highlight how the achievement scores of students are combined into a measure of teacher performance matters—independent of other design features. Second, we provide evidence sug-

[2] 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, 2009; Glewwe, Ilias, and Kremer 2010; Muralidharan and Sundararaman 2011; Duflo, Hanna, and Ryan 2012; Fryer et al. 2012; Dee and Wyckoff 2015; Lavy 2015). For instance, impressive evidence comes from a large-scale experiment in India that 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 2017).

[3] Muralidharan and Sundararaman (2011) study a piece-rate scheme tied to average gains in student achievement. The scheme studied in Behrman et al. (2015) rewarded and penalized teachers based on the progression (or regression) of their students (individually) through proficiency levels. The scheme studied in Springer et al. (2010) rewarded 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 (2017) 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.

gesting that incentive schemes can be designed to reduce 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 2010; 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.[4]

Our findings also contribute to literatures outside education. Our results add to a growing number of studies that use field experiments to evaluate performance incentives in organizations (Bandiera, Barankay, and Rasul 2005, 2007; Cadsby, Song, and Tapon 2007; Bardach et al. 2013; Luo et al. 2015). We also contribute to the literature on tournaments, particularly by testing the effects of different-sized rewards. Although there is evidence from the laboratory (see Freeman and Gelber 2010), we know 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 whether teachers responded optimally to the incentives they were given, we find that they did respond more to pay-for-percentile incentives than simpler 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.

II. Experimental Design and Data

A. 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 gross domestic product per capita in China), and the sec-

[4] This adds to three recent experimental studies that test the impacts of incentive reward size in alternative contexts: Ashraf, Bandiera, and Jack (2014), Luo et al. (2015), and Barrera-Osario and Raju (2017). 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 10 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 (2017) 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.

ond 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.[5] Typical of rural China, the sampled primary schools were public schools, composed of grades 1–6, and had an average of close to 400 students.

B. 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).[6] Across these three incentive groups, we orthogonally assigned schools to reward size groups: a small-reward group (78 schools; 3,465 students) and a large-reward group (86 schools; 3,353 students). All sixth-grade math teachers in a school were assigned to the same treatment.

To improve power, we randomized within counties (16 counties or strata) and controlled for stratum fixed effects in our estimates (Bruhn and McKenzie 2009). Our sample gives us enough power to test between (*a*) the different incentive design arms (control, levels, gains, and pay-for-percentile) and (*b*) the different reward size arms (control, small, and large). We did not power the study to test for differences in effects between the individual cells in table 1 (e.g., large pay-for-percentile rewards vs. small pay-for-percentile rewards). For this reason, we prespecified that the tests of differences between incentive design arms and the tests of differences between reward size arms are primary hypotheses tests, whereas the tests for interaction effects and differences between individual cells are exploratory.

C. Incentive Design and Conceptual Framework

Our primary goal is to evaluate designs that use alternative ways of defining teacher performance as a function of student achievement. To do so, we

[5] 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 (fifth and sixth grades) and send eligible students to neighboring schools. We excluded these "incomplete" elementary schools. Third, we excluded elementary schools that had enrollments smaller than 120 (i.e., enrolling an average of fewer than 20 students per grade). Because the prefecture departments of education informed us that these schools would likely be merged or closed down in following years, we decided to exclude these schools from our sample.

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

Table 1
Experimental Design

	Number of Schools (Students)		Total
Control group	52 (2,254)		52 (2,254)
	Reward Size Groups		
	Large Reward	Small Reward	
Incentive design groups:			
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)	

NOTE.—The 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 (stratum) not being evenly divisible.

compare three alternative ways of combining the achievement scores of individual students in each teacher's class into a single measure of teacher performance (incentive design treatments), which are then used to rank teachers in tournaments with a common structure and common budget. We also compare tournaments with a common structure but with two different reward sizes.

1. Incentive Design Treatments

The three incentive design treatments that we evaluate are as follows.

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 (Ganimian and Murnane 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.[7] Within each of these comparison groups students were 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 of the 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.[8] This percentile performance index can be interpreted as the fraction of contests that students of a given teacher won compared with students who were taught by other teachers yet began the school year at similar achievement levels (Barlevy and Neal 2012).

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 underlying rank-order tournament structure. Using a common underlying rank-order tournament scheme allows us to directly compare the effects of varying how achievement scores are used to rank teachers independent of changes to payouts. This also keeps the total costs constant across these schemes within the small- and large-reward tournaments, so more effective schemes are also more cost-effective. Direct comparison would not have been possible with a piece-rate incentive scheme, as the rewarded units would have necessarily differed.

When informed of their incentive, teachers were told that they would compete with sixth-grade math teachers in other schools in their prefecture,[9] and the competition would be based on their students' performance on a common math exam.[10] According to their percentile ranking among other teachers in the program, teachers were told they would be given a cash reward within 2 months after the end of the school year.

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

$$Reward = R_{\text{top}} - (99 - Teacher'sPercentileRank) \times b,$$

where R_{top} was the reward for teachers ranking in the top percentile and b was the incremental reward for each increase in his or her percentile rank. In

[7] Teachers were not told the baseline achievement scores of individual students in any of the designs.

[8] We used the average as per Neal (2011).

[9] The two prefectures in the study each have hundreds of primary schools (751 in the prefecture in Shaanxi and 1,200 in the prefecture in Gansu). Teachers were not told the total number of teachers who would be competing in the tournament.

[10] 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.

the small-reward treatment, teachers ranking in the top percentile received 3,500 yuan ($547), and the incremental reward per percentile rank was 35 yuan.[11] In the large-reward treatment, teachers ranking in the top percentile received 7,000 yuan ($1,094), and the incremental reward per percentile rank was 70 yuan. Reward amounts were calibrated so that the top reward was equal to approximately 1 month's salary in the small-reward treatment and 2 months' salary in the large-reward treatment.[12]

Note that even though the underlying reward structure and distribution of payouts is the same, a teacher's effective "competitors" differ under levels, gains, and pay-for-percentile. Under levels or gains, teachers are given a percentile rank (between 0 and 99) based on how they perform against all other teachers (regardless of the initial achievement level of the teacher's student[s]). By contrast, under pay-for-percentile, teachers are given a percentile rank (between 0 and 99) based on how they perform against teachers in their comparison group (i.e., teachers who have students with the same initial level of achievement). Regardless of the incentive scheme, teacher percentile rank is used to calculate teacher payouts according to the linear in percentile rank formula given above.

Our rewards scheme departs from traditional schemes that have a less differentiated reward structure. Specifically, tournament schemes typically have fewer reward levels and only reward top performers (see, e.g., Lavy 2009). By setting rewards to be linearly increasing in percentile rank, our scheme is similar to the linear relative performance evaluation scheme studied in Knoeber and Thurman (1994),[13] which minimizes distortions in incentive strength due to nonlinearities in rewards.[14]

[11] 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.

[12] 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 for exams at the end of the year.

[13] 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, Barankay, and Rasul (2005) compare an LRPE scheme with piece rates in a study of fruit pickers in the United Kingdom.

[14] Tournament theory suggests a trade-off 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 laboratory 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.

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). Neal (2011) notes that tournaments may also be less subject to political pressures that flatten rewards. Importantly for risk-averse agents, tournaments are also more robust to common shocks across all participants.[15] 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; Briggs and Weeks 2009).[16]

3. Implementation

Following a baseline survey, 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 Sciences and were presented with government officials. Before signing the contract, teachers were provided with materials explaining the contract and how rewards would be calculated.[17] To better ensure that teachers understood the incentive structure and contract terms, they were also given a 2-hour training session. A short quiz was also given to teachers to check misunderstanding of the contract terms and reward determination. Correct responses were reviewed with teachers.

4. Conceptual Framework

Our goal is to evaluate how each of the three ways of measuring and ranking teacher performance using student 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. Second, we aim to understand how each scheme affects how teachers allocate effort across

[15] 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).

[16] Bandiera, Barankay, and Rasul (2005) find that piece-rate incentives outperform relative incentives in a study of fruit pickers in the United Kingdom. Their findings suggest, however, that this is due to workers' desire to not impose externalities on coworkers 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.

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

students in their classes—that is, 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 we study should depend on teachers' beliefs about the mapping between their effort and expected changes in their performance rank. The more symmetry there is in the contest—or the more a teacher's relative performance rank is attributable to effort rather than other factors—the better the reward scheme will be in eliciting effort (Lazear and Rosen 1981; Green and Stokey 1983; Nalebuff and Stiglitz 1983; Barlevy and Neal 2012). The reward schemes that we compare (levels, gains, and pay-for-percentile) differ only in how student scores are combined into a performance index for each teacher, which is then used to rank and reward teachers in the same way. Differences in strength are due to how well performance indices control for asymmetry arising from differences in class composition. The relative strength of the reward schemes will vary due to asymmetry arising from (*a*) variation in baseline student ability, (*b*) perceived variation in achievement gains (teacher returns to effort) as a function of baseline student ability, (*c*) measurement error in test scores, and (*d*) teacher uncertainty related to seeding.

With levels incentives—in which teachers are ranked and rewarded based on the average performance of their students at the end of the school year—each of these factors may contribute to asymmetry. Incentives will be weaker for teachers who teach classes that are, on average, low- or high-achieving because endline rank is largely determined by differences in baseline student ability. Less directly, how teachers perceive returns to effort will depend on (i) whether the performance of initially low-achieving students responds more or less to a given level of teaching effort than middle- or high-achieving students and (ii) 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).[18] Asymmetry may further increase, for instance, if teachers believe that returns to baseline ability and teaching effort are positively correlated. Teachers of a less able class not only would be at a disadvantage due to initial differences in ability but would also need to invest more effort to realize an equivalent gain. Asymmetry may be reduced on net if this correlation is perceived to be negative, although this may be dominated by differences in initial ability.[19]

Compared with levels, ranking and rewarding teachers according to gains may increase contest symmetry by partially adjusting for average baseline

[18] Note that there was no top coding in the exams used to assess student performance.

[19] We show evidence below (in Sec. III.D.1) that teachers do indeed believe that returns to effort (in terms of a hypothetical assessment scale) are higher for students toward the bottom of the distribution.

ability. Asymmetry will nevertheless arise if teachers believe that improving student achievement requires more or less effort for students at different levels of baseline achievement. With gains, either a positive or a negative correlation between baseline achievement and perceived returns to teaching effort will increase asymmetry. If they are positively (negatively) correlated, teachers with a low-baseline-ability (high-baseline-ability) class will be at a perceived disadvantage. The strength of gains incentives may also be weakened relative to levels if teachers recognize that gains indices are more subject to statistical noise (Ganimian and Murnane 2014).

As discussed in Barlevy and Neal (2012), pay-for-percentile is designed to "elicit efficient effort from all teachers in all classrooms" (p. 1807). Pay-for-percentile will likely produce a more symmetric contest than both levels and gains incentives because pay-for-percentile, by construction, places teachers in contests based on their students' performance relative to other students with the same baseline performance. Although asymmetry between teachers may still be present due to differences in class size, peer composition, and teacher ability (assuming that these are not addressed by seeding the contest), pay-for-percentile increases symmetry by matching a teacher's students with similar peers in other classes. Moreover, pay-for-percentile incentives may outperform levels and gains incentives because symmetry under pay-for-percentile depends less on teacher beliefs about the relationship between returns to teaching effort and baseline student ability. Under levels and gains, teachers may be reluctant to increase effort due to beliefs (and uncertainty) about this relationship.[20]

That the marginal reward for teachers is higher under pay-for-percentile than under levels or gains holds for the linear in percentile rank reward structure that we study and for rank-order tournament reward structures more generally. As an illustration, first consider an extreme example with the following assumptions: (*a*) each teacher has a single student; (*b*) there are two equally sized ex ante student achievement levels (low achieving and high achieving); and (*c*) low-achieving students are never observed to make as much progress as high-achieving students (due, for instance, to sharply decreasing marginal returns to teacher effort).

Under pay-for-percentile, teachers whose student is in the low-achieving or high-achieving group can obtain a percentile rank between 0 and 99. Students in the low-achieving group obtain a percentile rank of 99 if their student outperforms all other low-achieving students on the end-of-year exam and a percentile rank of 0 if this student ranks last. Similarly, teachers of

[20] 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 in testing.

high-achieving students receive a percentile rank of 99 if their student outperforms all other ex ante high-achieving students on the end-of-year exam and 0 if their student does not perform as well as all other ex ante high-achieving students.

By contrast, under levels or gains teachers in the low-achieving group can obtain only a percentile rank between 0 and 50, while teachers in the high-achieving group can obtain only a percentile rank between 51 and 99. Thus, according to the linear in percentile rank rewards formula, whereas teachers with students of the same ex ante achievement level (low or high) can receive anywhere from 0 to 7,000 RMB under pay-for-percentile, they can receive only from 0 to 3,500 RMB (if the teacher is in the low-achieving group) or 3,570 to 7,000 RMB (if the teacher is in the high-achieving group) under levels or gains.[21] In terms of marginal rewards, teachers potentially have twice as much to gain or lose from "beating" one more teacher (70 RMB vs. 35 RMB with 100 teachers in each group, for instance) at the same achievement level under pay-for-percentile than under levels or gains, and equilibrium effort would be higher as a result.

If we were to relax assumption *b* and assume that there are *N* equally sized ex ante achievement groups (instead of just two) that are unable to compete with each other, pay-for-percentile would offer teachers up to *N* times as much reward for beating a teacher at the same achievement level compared with levels or gains.[22] In other words, the greater the asymmetry attributable to differences in ex ante achievement levels, the greater potential marginal rewards under pay-for-percentile compared with levels and gains.[23] Assuming that contests within each ex ante achievement group are symmetric, the exact level of effort that teachers choose depends on the potential marginal reward, which will always be weakly higher under pay-for-percentile. This holds under the linear in percentile rank tournament (and in rank-order tournaments

[21] Amounts refer to the "large-payout" formula. The same arguments hold regardless of the size of the incremental payout.

[22] When there are 100 teachers in each of four equally sized groups, e.g., teachers in any of the groups still receive 70 RMB more from beating an additional teacher under pay-for-percentile but only 17.5 RMB under levels or gains. As ex ante achievement groups become more unequal in size, marginal rewards under pay-for-percentile converge to levels but always remain higher.

[23] In practice, ex ante achievement groups, while fixed by design under pay-for-percentile, are determined by the nature of the achievement production function under levels and gains. Teachers' "competitors" under these schemes could also be influenced by how measurement error in test scores varies with ex ante achievement levels. Generally, competitiveness (symmetry) in the levels and gains schemes will predominantly be a function of how quickly marginal returns to effort decrease in terms of test score gains at each point in the ex ante distribution. The faster marginal returns to effort decrease in terms of test score gains, the higher the marginal reward under pay-for-percentile relative to levels- and gains-based incentives.

with less differentiated reward structures) and even when there is only one student per teacher.

Although this framework 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. 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.—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 allocate effort across students in the class to maximize the class-average score on the final exam.[24] Assuming that 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

[24] 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.

to score close to full marks even without extra instruction). Whether and how triage occurs depends on how teacher perception of returns to effort vary across students with different baseline achievement levels.[25]

Compared with levels and gains incentives, pay-for-percentile incentives may or may not limit the potential for triage. On the one hand, triage may be reduced because pay-for-percentile rewards teachers according to each student's performance in ordinal, equally weighted contests. A teacher essentially competes in as many contests as there are students in her class that have comparison students in other schools and is rewarded based on each student's rank in these contests, independent of the assessment scale. As a result, returns to effort may be more equal across students than under levels or gains incentives. On the other hand, differences in the variance of measurement error across the baseline ability distribution of students may lead to greater triage under pay-for-percentile relative to levels or gains. Presume, for instance, that low-ability students respond more on average to teacher effort, yet tests measure their performance with a larger amount of error than for high-ability students. While under levels and gains teachers would direct more effort to low-ability students, under pay-for-percentile the relative return to effort toward low-ability students would be reduced by greater measurement error, and teachers would devote less effort to low-ability students.

D. Data Collection

Student surveys.—We conducted two baseline surveys of students, one at the beginning (September 2012) and one at the end (May 2012) of fifth grade. The surveys collected information on basic student and household characteristics (such as age, gender, parental education, parental occupation, family assets, and number of siblings).

We also conducted an endline survey of students in May 2014 (at the end of sixth grade). In the endline, students were asked detailed questions about 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); the time students spent on math and other subjects each week; perceptions of teaching practices, teacher care, teacher management of the classroom, and teacher communication; and parent involvement in schoolwork.[26]

[25] 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; $p < .001$).

[26] Measures of students' perceptions of teacher behavior were drawn from contextual questionnaires used in the 2012 Programme on International Student Assessment (PISA). These measures are discussed in detail in the PISA technical report (OECD 2013). These measures were chosen precisely because, as discussed extensively in the

Teacher surveys.—We conducted a baseline survey of all sixth-grade math teachers at the start of sixth grade (in September 2013, before the intervention). The survey collected information on teacher gender, ethnicity, age, teaching experience, teaching credentials, attitudes toward performance pay, and current performance pay. We also elicited teachers' perceived returns to teaching effort for individual students within the class (the survey is described in detail below). We administered a nearly identical survey to teachers in May 2014 after the conclusion of the experiment.

Standardized math exams.—Our primary outcome is student math achievement. Math achievement was measured during the endline and two 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), and the content validity of these test items was checked by multiple experts. The psychometric properties of the tests were then validated using data from extensive pilot testing to ensure good distributional properties (no bottom or top coding, for instance).[27] 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.

E. Balance and Attrition

Table A1 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 differences than would be expected by chance.[28] Additionally, omnibus tests across all baseline characteristics in table A1 do not reject balance across the student arms.[29] Characteristics are also balanced across the incentive design arms within the small- and large-reward 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 sam-

educational literature, they have been found to capture real information on effective classroom teaching (Tschannen-Moran and Hoy 2007; Hattie 2009; Klieme, Pauli, and Reusser 2009; Pianta and Hamre 2009; Baumert et al. 2010).

[27] In the endline exam, only 23 students (0.27%) received a full score, and no students received a zero score.

[28] Note that teacher-level characteristics in this table differ from those in our preanalysis 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.

[29] These tests were conducted by regressing treatment assignment on all of the baseline characteristics in table A1 using ordered probit regressions and testing that coefficients on all characteristics were jointly zero. The p-value of this test is .758 for the incentive design treatments and .678 for the reward size treatments.

ple.[30] Table A2 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 with the control group (by 2.6 percentage points; row 3, col. 3).

F. Empirical Strategy

Given the random assignment of schools to treatments, comparisons of mean outcomes across treatment groups provide unbiased estimates of the effect of each experimental treatment. However, to increase precision we condition our estimates on additional covariates. With few exceptions, all of the analyses presented were prespecified in a preanalysis plan written and filed before endline data were available for analysis.[31] In reporting the results below, we explicitly note analyses that deviate from the preanalysis plan.

As prespecified, we use ordinary least squares regression to estimate the effect of incentive treatments on student outcomes with the following specification:

$$Y_{ijc} = \alpha + T'_{jc}\beta + X_{ijc}\gamma + \tau_c + \varepsilon_{ijc}, \quad (1)$$

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. To increase precision, X_{ijc} includes the two waves of baseline achievement scores in all specifications. We also estimate treatment effects with an expanded set of controls. For student-level outcomes, this includes student age, 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 Liang-Zeger standard errors. 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).

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)

[30] 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.

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

and Imbens and Rubin (2015).[32] 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 $\sup(F^A(y) - F^B(y))$, 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 $\sup|F^A(y) - F^B(y)|$ 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 math scores), we use equation (1) to estimate effects on secondary outcomes that may explain underlying changes in math scores. As prespecified, the secondary outcomes are frequently summary indices constructed using groups of closely related outcome variables.[33] Specifically, we used a generalized least squares (GLS) weighting procedure to construct the weighted average of k normalized outcome variables in each group (y_{ijk}; Anderson 2008). The weight placed on each outcome variable is the sum of its row entries in the inverted covariance matrix for group j such that

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

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

III. Results

A. Average Impacts of Incentives on Achievement

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 standard deviations (table 2, panel A, row 1, col. 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

[32] This analysis was not prespecified.

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

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)
	A. Impacts Relative to Control Group									
1. Any incentive	.063	.074*								
	(.043)	(.044)								
2. Levels incentive			.056	.084			.046	.080	.064	.081
			(.048)	(.052)			(.059)	(.067)	(.059)	(.061)
3. Gains incentive			.012	.001			.049	.037	−.033	−.033
			(.051)	(.050)			(.064)	(.063)	(.060)	(.061)
4. Pay-for-percentile incentive			.128*	.148**			.089	.131	.163**	.165**
			(.064)	(.064)			(.094)	(.100)	(.059)	(.060)
5. Small reward					.063	.081				
					(.053)	(.055)				
6. Large reward					.064	.067				
					(.045)	(.046)				
7. Additional controls		X		X		X		X		X
8. Observations	7,454	7,373	7,454	7,373	7,454	7,373	4,655	4,609	4,678	4,628
	B. Comparisons between Incentive Treatments									
9. Gains − levels			−.044	−.083			.003	−.043	−.096	−.114
10. *p*-value: gains − levels			.390	.114			.974	.605	.153	.100
11. P4P − levels			.072	.064			.043	.051	.099	.085
12. *p*-value: P4P − levels			.236	.292			.648	.602	.157	.237
13. P4P − gains			.116	.147**			.041	.094	.195**	.199**
14. *p*-value: P4P − gains			.078	.023			.698	.406	.005	.004
15. Large - small					.001	−.014				
16. *p*-value: large − small					.989	.778				

NOTE.—Rows 1–6 (panel A) show estimated coefficients and standard errors (in parentheses) obtained by estimating eq. (1). Standard errors account for clustering within schools. The dependent variable in each regression is student endline standardized math exam scores 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 along with corresponding (unadjusted) *p*-values. Asterisks indicate significance after adjusting for multiple hypotheses using the step-down procedure of Romano and Wolf (2005), which controls for the family-wise error rate. P4P = pay-for-percentile.

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

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

incentives raise student scores by 0.128 standard deviations (in the basic regression specification) to 0.148 standard deviations (in the specification with additional controls; panel A, row 4, cols. 3 and 4).[34] By contrast, we find no significant effects from offering teachers levels or gains incentives based on regression estimates (panel A, rows 2 and 3, cols. 3 and 4).

Comparing across the incentive design treatment point estimates, pay-for-percentile significantly outperforms gains (by 0.147 standard deviations; panel B, row 13, col. 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 standard deviations). A joint test of equality shows that the three coefficients on the incentive design treatments differ significantly from one another ($p = .065$).

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, cols. 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 standard deviations; panel A, row 4, cols. 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 ($p = .268$).[35]

B. Distributional Treatment Effects of Incentive Designs

The separate incentive designs are hypothesized to affect not only average performance but also performance across the distribution of ability. In this section, following Abadie (2002), we therefore examine differences in the full distribution of scores across the incentive design groups. Figure 1 shows the cumulative distributions of student test performance across the experimental groups. For the full sample (fig. 1*A*), the small-reward group only (fig. 1*B*), and the large-reward group only (fig. 1*C*), we plot the distributions of student scores adjusted for the set of prespecified 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, although differences appear larger with large rewards.

[34] In addition to the student-level regressions, which were prespecified, we also estimated school-level regressions using data averaged at the school level (see table A3).

[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 prespecified.

[36] These are adjusted by estimating eq. (1) without treatment dummies and saving predicted residuals. Figure A1 shows cumulative distributions using unadjusted student scores.

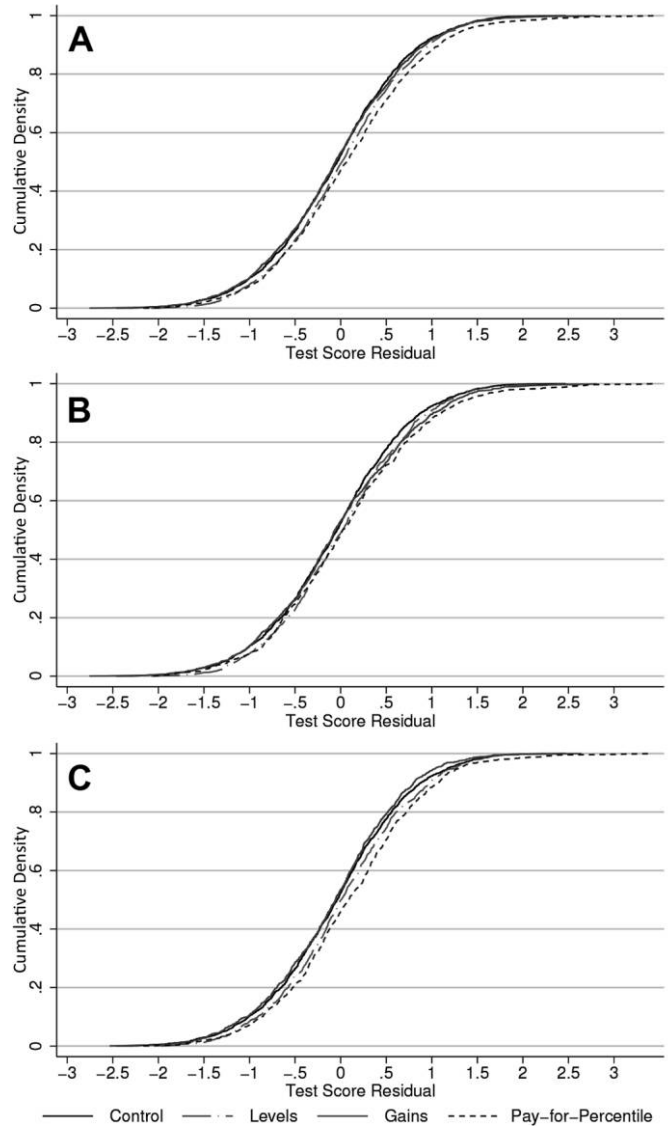

FIG. 1.—Distribution of test scores across groups. The figure shows estimated cumulative density functions of adjusted student scores across incentive treatment arms for the full sample (*A*), small-reward schools only (*B*), and large-reward schools only (*C*).

Table 3 presents results for Kolmogorov-Smirnov-type tests between each distribution pair using the full sample. Panel A presents tests comparing each incentive design with the control group, and panel B shows comparisons between each treatment pair. For each comparison we show results

Table 3
Tests for Distributional Treatment Effects

Test	Test Statistic (1)	p-Value (2)
	A. Relative to Control Group	
1. Levels incentive:		
Unidirectional: $\mathbb{F}^{\text{Levels}} - \mathbb{F}^{\text{Control}}$	.036	.077
Unidirectional: $\mathbb{F}^{\text{Control}} - \mathbb{F}^{\text{Levels}}$	.000	.976
Equality of distributions	.036	.045
2. Gains incentive:		
Unidirectional: $\mathbb{F}^{\text{Gains}} - \mathbb{F}^{\text{Control}}$	.024	.258
Unidirectional: $\mathbb{F}^{\text{Control}} - \mathbb{F}^{\text{Gains}}$	.024	.188
Equality of distributions	.024	.131
3. Pay-for-percentile incentive:		
Unidirectional: $\mathbb{F}^{\text{P4P}} - \mathbb{F}^{\text{Control}}$	.071	.018
Unidirectional: $\mathbb{F}^{\text{Control}} - \mathbb{F}^{\text{P4P}}$	.000	1.000
Equality of distributions	.071	.013
	B. Between Incentive Treatments	
4. Levels − gains:		
Unidirectional: $\mathbb{F}^{\text{Levels}} - \mathbb{F}^{\text{Gains}}$	.042	.037
Unidirectional: $\mathbb{F}^{\text{Gains}} - \mathbb{F}^{\text{Levels}}$	.008	.622
Equality of distributions	.042	.013
5. P4P − levels:		
Unidirectional: $\mathbb{F}^{\text{P4P}} - \mathbb{F}^{\text{Levels}}$	.048	.068
Unidirectional: $\mathbb{F}^{\text{Levels}} - \mathbb{F}^{\text{P4P}}$	.008	.499
Equality of distributions	.048	.043
6. P4P − gains:		
Unidirectional: $\mathbb{F}^{\text{P4P}} - \mathbb{F}^{\text{Gains}}$	.056	.033
Unidirectional: $\mathbb{F}^{\text{Gains}} - \mathbb{F}^{\text{P4P}}$	.000	1.000
Equality of distributions	.056	.023

NOTE.—Panel A shows test statistics and p-values from Kolmogorov-Smirnov tests between the distribution of adjusted endline exam scores in each treatment group and the control group following Abadie (2002). The endline exam scores were adjusted for baseline exam scores and strata fixed effects. 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. P4P = pay-for-percentile.

for three tests discussed in Section II.F: the two unidirectional tests and the nondirectional 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 .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 .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 that pay-for-percentile incentives outperform both gains and levels incentives. The p-value for the difference between levels and gains is .037 (table 3, row 4). The p-values for the difference between pay-for-percentile and levels and gains are .068 (table 3, row 5) and .033 (table 3, row 6), respectively. In all three comparisons test statics show first-order stochastic dominance or very near first-order stochastic dominance.

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. Taken together with the comparison between small and large rewards, these results suggest that how teacher performance is measured has a larger effect on student performance than doubling the size of potential rewards.

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

To estimate the effects of incentives on secondary student outcomes and teacher behavior that may explain effects on student achievement, we run regressions analogous to equation (1) but substitute endline achievement with secondary student outcomes and measures of teacher behavior.[37]

[37] The measures of secondary outcomes that we use were specified in our preanalysis plan. Most of these measures (math self-concept, math anxiety, math intrinsic and instrumental motivation, student time on math, student perception of teaching practices, teacher care, teacher management of the classroom, teacher communication, 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 Sec. II.F). These each have a mean of 0 and a standard deviation 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. The survey questions regarding curricular coverage were given at the end of the school year, at the end of sixth grade. Curricular coverage (or "opportunity to learn") is commonly measured in the education research literature (see Schmidt et al. 2015). Students were given three such examples of curricular material from the last semester of grade 5 ("easy" material), three from the first semester of grade 6 ("medium" material), and three from the second semester of

Table 4
Impacts on Question Difficulty Subscores and Curricular Coverage

	Curricular Coverage				Difficulty Subscores		
	Overall (1)	Easy (2)	Medium (3)	Hard (4)	Easy (5)	Medium (6)	Hard (7)
1. Levels incentive	.015	.019	.020	.005	.029	.094	.075
	(.010)	(.012)	(.010)	(.015)	(.044)	(.050)	(.052)
2. Gains incentive	.008	.012	.022*	−.009	−.006	−.010	.019
	(.009)	(.012)	(.010)	(.014)	(.036)	(.050)	(.053)
3. Pay-for-percentile incentive	.027**	.016	.025*	.040**	.105**	.092	.160**
	(.011)	(.012)	(.011)	(.014)	(.043)	(.062)	(.067)
4. Observations	7,363	7,373	7,370	7,366	7,373	7,373	7,373

NOTE.—Rows 1–3 show estimated coefficients and standard errors (in parentheses) obtained by estimating regressions analogous to eq. (1). Standard errors account for clustering at the school level. The dependent variables in cols. 1–4 are measures of curricular coverage (for all, easy, medium, and hard items), as reported by students. The dependent variables in cols. 5–7 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. Asterisks indicate significance after adjusting for multiple hypotheses using the step-down procedure of Romano and Wolf (2005), which controls for the family-wise error rate.

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

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

We find that the different incentive design treatments had significant effects on teaching practice as measured by curricular coverage (table 4, cols. 1–4). Pay-for-percentile also had a significant effect on curricular coverage overall (row 3, col. 1), and this effect is larger than that of gains incentives ($p = .074$) and levels incentives (although not statistically significant; $p = .238$).[38] Compared with the control group, students in the gains group report being taught more curricula at the medium level (row 2, col. 3), and students in the pay-for-percentile group report being taught more medium and hard curricula (row 3, cols. 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 (for levels, $p = .022$; for gains, $p = .001$).

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

grade 6 ("hard" material). According to national and regional standards, even the hard material should be taught before the end of sixth grade (before the endline survey). Students' binary responses to each example of curricular material were averaged for all three categories together and the easy, medium, and hard categories separately.

[38] Testing effects on overall curricular coverage (combining easy, medium, and hard) was not included in the preanalysis plan.

as possible, leaving less time for students to master the subject matter. To test this, we estimate treatment effects on subsets of test items categorized into easy, medium, and hard questions (table 4, cols. 5–7).[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 with the control group, students in classes where teachers had pay-for-percentile incentives had significantly higher scores in the easy and hard difficulty categories. Pay-for-percentile incentives increased the easy question subscore by 0.105 standard deviations (row 3, col. 5) and the hard question subscore by 0.16 standard deviations (row 3, col. 7). 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 the 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 (table A4). There are no statistically significant impacts from any of the incentive arms on time on math, perceptions of 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, Ilias, and Kremer (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 the classroom. While we do find impacts of all types of incentives on student-reported times being tutored outside class (col. 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 (table A4, cols. 1–5). There is also no evidence that any type of incentives led to increased substitution of time away from subjects other than math (col. 13).

D. Effects on the Within-Class Distribution of Student Achievement

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 elic-

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

ited as part of the baseline survey.[40] To elicit a measure of teachers' 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 1 hour of extra personal instruction from the teacher per week.[42] A teacher's perception of his or her own value added for each student is measured as the difference between these scores, normalized by the distribution of the teacher's reported expectation of gains across students. The perceived value-added measure intends to measure how much teachers perceive their effort contributes to achievement gains for different students. While the question does not capture other dimensions of teacher effort, we assume that the contribution of additional time is a good general proxy for the marginal contribution of teacher effort.[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,

[40] The analyses in this subsection were not prespecified 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. Limiting the exercise to only 12 students per class reduces the statistical power of the subsequent analyses but was necessary to ensure a higher quality of responses from teachers.

[42] Precisely, for each student teachers were asked (*a*) to rank the math achievement of the student compared with other students on the list; (*b*) to estimate by how much they would expect this student's score to change (in terms of percentage of correct answers) if this student were given curriculum-appropriate exams at the beginning and the end of sixth grade; and (*c*) to estimate by how much they would expect this student's score to change (in terms of percentage of correct answers) if the student received one extra hour of personal instruction from you per week. 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 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 above results 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.

Table 5
Correlation between Teacher Perception of Own Value Added and Student Characteristics

	Dependent Variable: Teacher Perceived Value Added							
	Teacher's Own Ranking of Students at Baseline				Ranking of Students by Baseline Exam Score			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Within-class student ranking used:								
1. Student within-class percentile rank	−.329***	−.317***			−.171*	−.186**		
	(.103)	(.104)			(.091)	(.094)		
2. Student in middle tercile of class (0/1)			−.065	−.053			−.034	−.045
			(.052)	(.053)			(.046)	(.047)
3. Student in top tercile of class (0/1)			−.206***	−.193***			−.106*	−.117*
			(.071)	(.071)			(.062)	(.064)
4. Female (0/1)		−.032		−.033		−.044		−.042
		(.045)		(.045)		(.047)		(.046)
5. Age (Years)		−.026		−.020		−.019		−.016
		(.025)		(.025)		(.026)		(.025)
6. Father attended secondary school (0/1)		−.054		−.058		−.061		−.062
		(.049)		(.049)		(.049)		(.050)
7. Mother attended secondary school (0/1)		−.025		−.027		−.029		−.030
		(.039)		(.039)		(.039)		(.038)
8. Household asset index		−.019		−.019		−.019		−.020
		(.018)		(.018)		(.018)		(.018)
9. Observations	2,444	2,347	2,444	2,347	2,444	2,347	2,444	2,347

NOTE.—Rows 1–8 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. Four students were randomly selected within each tercile of the within-class baseline achievement distribution to ensure coverage across achievement levels. For each student on the list, teachers were asked (*a*) to rank the math achievement of the student compared with other students on the list; (*b*) to estimate by how much they would expect this student's score to change (in terms of percentage of correct answers) if this student were given curriculum-appropriate exams at the beginning and the end of sixth grade; and (*c*) to estimate by how much they would expect this student's score to change (in terms of percentage of correct answers) if the student received one extra hour of personal instruction from you per week. 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 cols. 1–4, 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 cols. 5–8, rows 1–3 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 10% level.

** Significant at the 5% level.

*** Significant at the 1% level.

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 [cols. 1 and 2] or the ranking based on the baseline exam [cols. 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 (cols. 3–4, 7–8). Teachers' perceived value added is approximately 0.2 standard deviations lower for students in the top third of the distribution compared with 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% report the lowest returns for the bottom tercile, and 17% report the lowest returns for the middle tercile. Teachers were nearly evenly split in reporting the highest returns for the bottom, middle, and top of the distribution.

Second, teachers' perceived value added is not significantly related to student background 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.

2. *Within-Class Distributional Effects of Incentives*

We estimate heterogeneous effects along the within-class distribution using 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. Table 6 compares the effects of levels, gains, and pay-for-percentile incentives between the bottom, middle, and top tercile of the within-class distributions.[44]

We find that the effects of levels and gains incentives are significantly larger 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. For students in the top tercile of teachers' perceived value added, levels incentives had an approximately 0.2 standard deviations larger effect

[44] We estimate effects by tercile of the distribution for each of these variables by estimating eq. (1) but including dummy variables for the middle and top terciles and interactions with indicators for the levels, gains, and pay-for-percentile incentive arms. See table A5 for regression coefficients used to compute values in table 6.

Table 6
Within-Class Distributional Effects

	Effect on Middle Tercile versus Effect on Bottom Tercile (1)	Effect on Top Tercile versus Effect on Bottom Tercile (2)	Effect on Top Tercile versus Effect on Middle Tercile (3)
	A. Teacher Perception of Own Value Added for Student		
1. Levels incentive	.053	.213*	.160
	(.111)	(.122)	(.119)
2. Gains incentive	.163	.333**	.170
	(.146)	(.152)	(.122)
3. Pay-for-percentile incentive	.056	.056	−.001
	(.139)	(.151)	(.127)
4. *N*	2,238	2,238	2,238
	B. Teacher Ranking of Students at Baseline		
5. Levels incentive	−.050	−.091	−.040
	(.100)	(.107)	(.102)
6. Gains incentive	.051	−.090	−.140
	(.107)	(.113)	(.115)
7. Pay-for-percentile incentive	−.022	−.069	−.048
	(.108)	(.115)	(.114)
8. *N*	2,415	2,415	2,415
	C. Ranking of Students by Baseline Exam Score		
9. Levels incentive	−.026	−.071	−.045
	(.059)	(.060)	(.062)
10. Gains incentive	−.031	−.041	−.010
	(.059)	(.064)	(.063)
11. Pay-for-percentile incentive	−.055	−.063	−.008
	(.065)	(.082)	(.072)
12. *N*	7,454	7,454	7,454

NOTE.—The table shows differences in estimated treatment effects for students in different terciles of within-class distributions at baseline. Estimates were obtained using eq. (1) supplemented with indicators for student baseline tercile and interactions with treatment arm indicators. Full regressions are shown in table A5. Each regression controls for two waves of baseline standardized math exam scores and strata (county) fixed effects. All standard errors account for clustering at the school level. Panel A shows differences in estimated effects by teacher perception of her own value added for each student as reported at baseline. See the note to table 5 and text for a description of how teacher perceptions of value added were measured. Panel B shows differences in estimated effects by teachers' ability ranking of students at baseline. Panel C shows differences in estimated treatment effects by student within-class ranking by baseline exam score.

* Significant at the 10% level.

** Significant at the 5% level.

than on students in the bottom tercile and gains incentives had an approximately 0.3 standard deviations larger effect than on students in the bottom tercile (although total effects of incentives on these students is not significantly positive in either case).[45] These results should be interpreted somewhat

[45] The coefficient on the interaction term between the top tercile of perceived value added and pay-for-percentile incentives in these regressions (table A5), however,

cautiously, as our power for detecting effects on exam scores is reduced using the random 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 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 leads 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 (table 6, panels B and C). Levels and gains incentives do not have significant effects for students at any part of the baseline distribution (table A5). Columns 5 and 6 of table A5 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 would anticipate levels and gains incentives having a positive effect on students at the bottom of the distribution. This effect may have been muted on average in the sample due to the large amount of heterogeneity in teachers' perceived returns.

IV. Discussion and Conclusion

This paper provides evidence on the relative effectiveness of different designs of teacher performance pay. 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. There are 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. Because the distribution of payouts and costs of measurement were constant across the different incentive design schemes, pay-for-percentile was also the most cost-effective. Pay-for-percentile incentives, but not the other two designs, increased both the coverage and the 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 re-

is not statistically different from the coefficients on the interactions terms between the top tercile and levels incentives ($p = .224$) or gains incentives ($p = .121$).

wards, 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 score 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 1 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 initially 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 lead teachers to teach to a particular test.

A second caveat is that our study was not powered ex ante to study the interaction between different incentive designs and reward size. 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.

Appendix

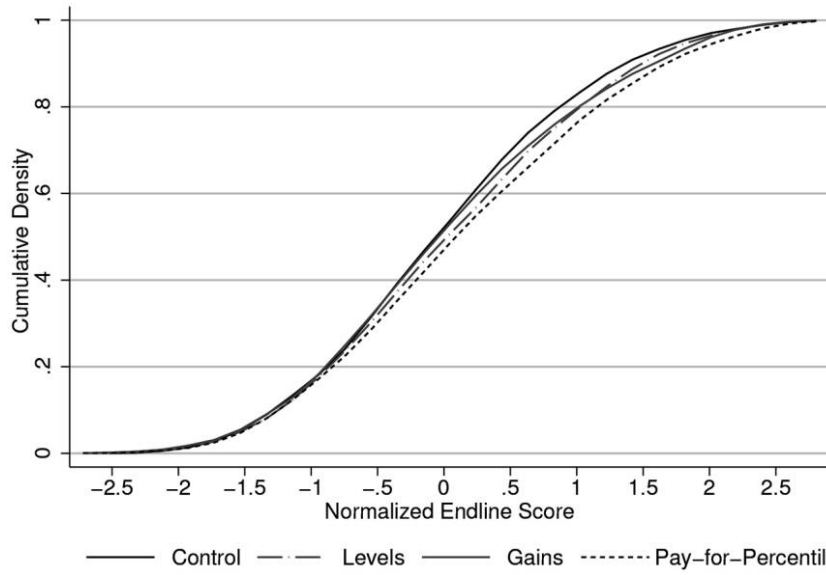

FIG. A1.—Distribution of test scores across groups (unadjusted). The figure shows estimated cumulative density functions of unadjusted student scores across incentive treatment arms for the full sample.

Table A1
Descriptive Statistics and Balance Check

	Control Mean (1)	Coefficient (Standard Error) On: Levels Incentive (2)	Gains Incentive (3)	Pay-for-Percentile Incentive (4)	Joint Test p-Value: All = 0 (5)	Coefficient (Standard Error) On: Small Incentive (6)	Large Incentive (7)	Joint Test p-Value (8)	Obs. (9)
A. Student Characteristics									
1. Standardized math exam score, beginning of previous school year	.00	−.045	−.015	−.094	.739	−.040	−.061	.751	7,996
		(.084)	(.082)	(.093)		(.079)	(.080)		
2. Standardized math exam score, end of previous school year	.00	−.005	.028	−.038	.894	.015	−.023	.848	8,136
		(.082)	(.091)	(.088)		(.080)	(.081)		
3. Female (0/1)	.49	−.010	−.002	−.011	.893	−.005	−.010	.816	7,996
		(.017)	(.015)	(.018)		(.015)	(.015)		
4. Age (Years)	11.99	.088	.137**	.082	.225	.104*	.103*	.176	7,992
		(.063)	(.066)	(.072)		(.062)	(.061)		
5. Father attended secondary school (0/1)	.52	.005	.028	.005	.686	.007	.019	.700	7,965
		(.024)	(.026)	(.026)		(.023)	(.023)		
6. Mother attended secondary school (0/1)	.31	.010	.019	.011	.900	.021	.007	.660	7,929
		(.026)	(.026)	(.026)		(.024)	(.023)		
7. Household asset index	−.64	.025	.014	.041	.865	−.001	.054	.348	7,996
		(.046)	(.048)	(.050)		(.042)	(.042)		
B. Teacher and Class Characteristics									
8. Age (Years)	32.62	1.065	−.142	.146	.863	−.225	1.021	.623	243
		(1.633)	(1.650)	(1.492)		(1.366)	(1.578)		
9. Female	.42	−.012	.104	−.021	.430	.015	.033	.928	243
		(.090)	(.089)	(.092)		(.082)	(.086)		

Table A1 (*Continued*)

	Control Mean (1)	Coefficient (Standard Error) On: Levels Incentive (2)	Gains Incentive (3)	Pay-for-Percentile Incentive (4)	Joint Test *p*-Value: All = 0 (5)	Coefficient (Standard Error) On: Small Incentive (6)	Large Incentive (7)	Joint Test *p*-Value (8)	Obs. (9)
10. Han (0/1)	.95	.012	−.056	−.002	.311	−.036	.008	.274	243
		(.031)	(.040)	(.030)		(.030)	(.034)		
11. Teaching experience (years)	11.61	1.250	..077	−.583	.734	−.218	.744	.806	243
		(1.758)	(1.899)	(1.637)		(1.497)	(1.760)		
12. Monthly base salary (yuan)	2,852.77	281.648*	−138.954	167.102	.029	125.499	75.148	.735	243
		(148.743)	(184.543)	(167.329)		(160.090)	(154.256)		
		C. School Characteristics							
13. Number of students in grade 6	43.35	−.218	2.693	−2.906	.323	−1.141	.941	.709	216
		(2.894)	(2.951)	(2.740)		(2.631)	(2.515)		
14. Number of students in school	437.83	−57.517	−33.369	−43.711	.812	−69.823	−16.254	.285	216
		(59.891)	(60.460)	(65.535)		(57.941)	(58.849)		
15. Number of teachers	29.75	−1.268	−3.310	−.423	.761	−3.757	.693	.248	216
		(4.069)	(3.618)	(4.150)		(3.424)	(3.883)		
16. Number of contract teachers	1.69	.330	−.025	.089	.941	.070	.193	.929	216
		(.621)	(.396)	(.422)		(.382)	(.502)		

SOURCE.—Baseline survey.

NOTE.—Panel A shows student-level characteristics, panel B shows teacher and class characteristics, and panel C shows school-level characteristics. The first column shows the mean in the control group. Exam scores are normalized using the 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. All significance tests account for clustering at the school level.

* Significant at the 10% level.

** Significant at the 5% level.

Table A2
Attrition

	Full Sample		Small-Reward Groups	Large-Reward Groups
	(1)	(2)	(3)	(4)
1. Levels incentive	.008		.028	−.007
	(.019)		(.033)	(.013)
2. Gains incentive	−.015	−.014	−.018	
	(.010)		(.013)	(.013)
3. Pay-for-percentile incentive	−.008		−.026*	.009
	(.017)		(.013)	(.030)
4. Small incentive		−.004		
		(.014)		
5. Large incentive		−.007		
		(.014)		
6. Observations	9,072	9,072	5,719	5,607
7. Mean in control			.064	

NOTE.—The table shows estimated coefficients and standard errors (in parentheses) 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 account for clustering at the school level.

* Significant at the 10% level.

Table A3
Impact of Incentives on Test Scores (School-Level Regressions)

	Full Sample						Small-Reward Groups Only		Large-Reward Groups Only	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	A. Impacts Relative to Control Group									
1. Any incentive	.037	.046								
	(.049)	(.050)								
2. Levels incentive			.045	.071			.019	.046	.075	.057
			(.059)	(.060)			(.077)	(.080)	(.064)	(.066)
3. Gains incentive			−.018	−.037			.008	−.002	−.055	−.075
			(.059)	(.060)			(.075)	(.078)	(.064)	(.067)
4. Pay-for-percentile incentive			.088	.110*			.037	.080	.148**	.140*
			(.060)	(.060)			(.076)	(.081)	(.066)	(.066)
5. Small reward					.023	.035				
					(.054)	(.056)				
6. Large reward					.054	.057				
					(.055)	(.056)				
7. Additional controls		X		X		X		X		X
8. Observations	214	214	214	214	214	214	137	137	129	129
	B. Comparisons between Incentive Treatments									
9. Gains − levels			−.063	−.108			−.011	−.048	−.130	−.133
10. *p*-value: gains − levels			.277	.072			.900	.598	.078	.084
11. P4P − levels			.043	.039			.018	.034	.072	.083
12. *p*-value: P4P − levels			.467	.508			.835	.707	.332	.270
13. P4P − gains			.106	.147**			.029	.082	.203**	.218**
14. *p*-value: P4P − gains			.070	.015			.734	.347	.007	.005
15. Large − small					.031	.022				
16. *p*-value: large − small					.521	.662				

NOTE.—Rows 1–6 (panel A) show estimated coefficients and standard errors (in parentheses) obtained by estimating eq. (1) but at the school level. The dependent variable in each regression is the school-level average of 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. Asterisks indicate significance after adjusting for multiple hypotheses using the step-down procedure of Romano and Wolf (2005), which controls for the family-wise error rate. P4P = pay-for-percentile.

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

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

Table A4
Impacts on Secondary Outcomes

	Dependent Variable												
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
1. Levels incentive	.023	.009	.029	−.042	.031	.014	.034	−.004	−.029	−.059	.036	.149*	−.010
	(.040)	(.039)	(.056)	(.046)	(.056)	(.040)	(.063)	(.049)	(.055)	(.049)	(.084)	(.076)	(.030)
2. Gains incentive	.012	.024	.093*	.022	.008	.022	−.003	.001	.043	.062	.008	.136*	−.014
	(.039)	(.034)	(.054)	(.039)	(.055)	(.036)	(.066)	(.052)	(.048)	(.046)	(.082)	(.070)	(.033)
3. Pay-for-percentile incentive	−.011	−.009	.083	.065	−.001	.040	−.005	.036	.071	.024	−.027	.118*	−.032
	(.043)	(.040)	(.063)	(.047)	(.054)	(.045)	(.073)	(.055)	(.067)	(.048)	(.082)	(.070)	(.034)
4. Observations	7,373	7,373	7,373	7,373	7,373	7,373	7,372	7,373	7,373	7,371	235	7,368	7,373

NOTE.—Rows 1–3 show estimated coefficients and standard errors (in parentheses) obtained by estimating regressions analogous to eq. (1). Standard errors account for clustering at the school level. Outcome variables in cols. 1–11 are summary indices. Summary indices were constructed using the generalized least squares weighting procedure in Anderson (2008). Each regression controls for two waves of baseline standardized math exam scores, strata (county) fixed effects, and student gender, age, parent educational attainment, a household asset index, class size, teacher experience, and teacher base salary. The regression reported in col. 11 is at the teacher level. Asterisks indicate significance after adjusting for multiple hypotheses using the step-down procedure of Romano and Wolf (2005), which controls for the family-wise error rate. Column headings are as follows: (1) math self-concept; (2) math anxiety; (3) math intrinsic motivation; (4) math instrumental motivation; (5) student time on math; (6) student perception of teaching practices; (7) teacher care; (8) teacher classroom management; (9) teacher communication; (10) parent involvement; (11) teacher self-reported effort; (12) out-of-class tutoring; and (13) time spent studying other subjects.

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

Table A5
Within-Class Distributional Effects (Full Regressions)

	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	−.124	−.133	.051	.053	.092	.091
	(.087)	(.087)	(.082)	(.083)	(.056)	(.058)
2. Gains incentive	−.185	−.185	.010	.017	.036	.055
	(.114)	(.114)	(.091)	(.093)	(.059)	(.061)
3. Pay-for-percentile incentive	−.020	−.031	.070	.083	.171**	.174**
	(.112)	(.118)	(.090)	(.093)	(.084)	(.083)
4. VAR (middle tercile)	−.077	−.088	.148*	.136*	−.179***	−.176***
	(.082)	(.081)	(.079)	(.082)	(.050)	(.050)
5. VAR (top tercile)	−.213**	−.237**	.424***	.411***	−.056	−.056
	(.096)	(.096)	(.079)	(.081)	(.068)	(.068)
6. Levels × VAR (middle tercile)	.053	.066	−.050	−.042	−.026	−.009
	(.111)	(.110)	(.100)	(.102)	(.059)	(.060)
7. Levels × VAR (top tercile)	.213*	.262**	−.091	−.062	−.071	−.067
	(.122)	(.122)	(.107)	(.107)	(.060)	(.062)
8. Gains × VAR (middle tercile)	.163	.158	.051	.055	−.031	−.045
	(.146)	(.143)	(.107)	(.109)	(.059)	(.060)
9. Gains × VAR (top tercile)	.333**	.354**	−.090	−.091	−.041	−.060
	(.152)	(.151)	(.113)	(.113)	(.064)	(.065)
10. Pay-for-percentile × VAR (middle tercile)	.056	.078	−.022	−.026	−.055	−.047
	(.139)	(.144)	(.108)	(.108)	(.065)	(.065)
11. Pay-for-percentile × VAR (top tercile)	.056	.086	−.069	−.081	−.063	−.066
	(.151)	(.155)	(.115)	(.114)	(.082)	(.083)
12. Additional controls		X		X		X
13. *N*	2,238	2,217	2,415	2,392	7,454	7,373

NOTE.—Rows 1–11 show estimated coefficients and standard errors (in parentheses) obtained by estimating regressions analogous to eq. (1) 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 the note to table 5 and the main 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 10% level.
** Significant at the 5% level.
*** Significant at the 1% level.

References

Abadie, Alberto. 2002. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association* 97, no. 457:284–92.

Abeler, Johannes, and Simon Jäger. 2015. Complex tax incentives. *American Economic Journal: Economic Policy* 7, no. 3:1–28.

Anderson, Michael 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, no. 484:1481–95.

Ashraf, Nava, Oriana Bandiera, and B. Kelsey Jack. 2014. No margin, no mission? A field experiment on incentives for public service delivery. *Journal of Public Economics* 120:1–17.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2005. Social preferences and the response to incentives: Evidence from personnel data. *Quarterly Journal of Economics* 120, no. 3:917–62.

———. 2007. Incentives for managers and inequality among workers: Evidence from a firm-level experiment. *Quarterly Journal of Economics* 122, no. 2:729–73.

Banerjee, Abhijit V., and Esther Duflo. 2009. The experimental approach to development economics. *Annual Review of Economics* 1, no. 1:151–78.

Bardach, Naomi S., Jason J. Wang, Samantha F. De Leon, Sarah C. Shih, W. John Boscardin, L. Elizabeth Goldman, and R. Adams Dudley. 2013. Effect of pay-for-performance incentives on quality of care in small practices with electronic health records: A randomized trial. *JAMA* 310, no. 10:1051–59.

Barlevy, Gadi, and Derek Neal. 2012. Pay for percentile. *American Economic Review* 102, no. 5:1805–31.

Barrera-Osorio, Felipe, and Dhushyanth Raju. 2017. Teacher performance pay: Experimental evidence from Pakistan. *Journal of Public Economics* 148:75–91.

Baumert, Jürgen, Mareike Kunter, Werner Blum, Martin Brunner, Thamar Voss, Alexander Jordan, Uta Klusmann, Stefan Krauss, Michael Neubrand, and Yi-Miau Tsai. 2010. Teachers' mathematical knowledge, cognitive activation in the classroom, and student progress. *American Educational Research Journal* 47, no. 1:133–80.

Behrman, Jere R., Susan W. Parker, Petra E. Todd, and Kenneth I. Wolpin. 2015. Aligning learning incentives of students and teachers: Results from a social experiment in Mexican high schools. *Journal of Political Economy* 123, no. 2:325–64.

Briggs, Derek C., and Jonathan P. Weeks. 2009. The sensitivity of value-added modeling to the creation of a vertical score scale. *Education Finance and Policy* 4:384–414.

Bruhn, Miriam, and David McKenzie. 2009. In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics* 1, no. 4:200–232.

Bruns, Barbara, Deon Filmer, and Harry Anthony Patrinos. 2011. *Making schools work: New evidence on accountability reforms*. Washington, DC: World Bank Publications.

Cadsby, C. Bram, Fei Song, and Francis Tapon. 2007. Sorting and incentive effects of pay for performance: An experimental investigation. *Academy of Management Journal* 50, no. 2:387–405.

Contreras, Dante, and Tomás Rau. 2012. Tournament incentives for teachers: Evidence from a scaled-up intervention in Chile. *Economic Development and Cultural Change* 61, no. 1:219–46.

Dee, Thomas S., and James Wyckoff. 2015. Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management* 34, no. 2:267–97.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2011. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *American Economic Review* 101, no. 5:1739–74.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. Incentives work: Getting teachers to come to school. *American Economic Review* 102, no. 4:1241–78.

Dynarski, Susan M., and Judith E. Scott-Clayton. 2006. The cost of complexity in federal student aid: Lessons from optimal tax theory and behavioral economics. *National Tax Journal* 59, no. 2:319–56.

Freeman, Richard B., and Alexander M. Gelber. 2010. Prize structure and information in tournaments: Experimental evidence. *American Economic Journal: Applied Economics* 2, no. 1:149–64.

Fryer, Roland G. 2013. Teacher incentives and student achievement: Evidence from New York City public schools. *Journal of Labor Economics* 31, no. 2:373–407.

Fryer, Roland G., Steven D. Levitt, John List, and Sally Sadoff. 2012. Enhancing the efficacy of teacher incentives through loss aversion: A field experiment. NBER Working Paper no. 18237, National Bureau of Economic Research, Cambridge, MA.

Ganimian, Alejandro J., and Richard J. Murnane. 2014. Improving educational outcomes in developing countries: Lessons from rigorous impact evaluations. Working Paper no. 20284, National Bureau of Economic Research, Cambridge, MA.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. Teacher incentives. *American Economic Journal: Applied Economics* 2, no. 3:205–27.

Green, Jerry R., and Nancy L. Stokey. 1983. A comparison of tournaments and contracts. *Journal of Political Economy* 91, no. 3:349–64.

Hanushek, Eric, and Ludger Woessmann. 2011. Overview of the symposium on performance pay for teachers. *Economics of Education Review* 30, no. 3:391–93.

Hattie, John. 2009. *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge: Cambridge University Press.

Imberman, Scott A., and Michael F. Lovenheim. 2014. Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay system. *Review of Economics and Statistics* 97, no. 2:364–86.

Ito, Koichiro. 2014. Do consumers respond to marginal or average price? Evidence from nonlinear electricity pricing. *American Economic Review* 104, no. 2:537–63.

Klieme, Eckhard, Christine Pauli, and Kurt Reusser. 2009. The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In *The power of video studies in investigating teaching and learning in the classroom*, 137– 160. Münster: Waxmann.

Knoeber, Charles R., and Walter N. Thurman. 1994. Testing the theory of tournaments: An empirical analysis of broiler production. *Journal of Labor Economics* 12, no. 2:155–79.

Kolenikov, Stanislav, and Gustavo Angeles. 2009. Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer? *Review of Income and Wealth* 55, no. 1:128–65.

Lavy, Victor. 2002. Evaluating the effect of teachers' group performance incentives on pupil achievement. *Journal of Political Economy* 110, no. 6:1286–317.

———. 2009. Performance pay and teachers' effort, productivity, and grading ethics. *American Economic Review* 99, no. 5:1979–2011.

———. 2015. Teachers' pay for performance in the long-run: Effects on students' educational and labor market outcomes in adulthood. Working Paper no. 20983, National Bureau of Economic Research, Cambridge, MA.

Lazear, Edward P., and Sherwin Rosen. 1981. Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89, no. 5:841–64.

Liebman, Jeffrey B., and Richard J. Zeckhauser. 2004. Schmeduling. Working paper, Harvard University.

Luo, Renfu, Grant Miller, Scott Rozelle, Sean Sylvia, and Marcos Vera-Hernández. 2015. Can bureaucrats really be paid like CEOs? School administrator incentives for anemia reduction in rural China. Working Paper no. 21302, National Bureau of Economic Research, Cambridge, MA.

Moldovanu, Benny, and Aner Sela. 2001. The optimal allocation of prizes in contests. *American Economic Review* 91, no. 3:542–58.

Muralidharan, Karthik. 2012. Long-term effects of teacher performance pay: Experimental evidence from India. Working paper, University of California, San Diego.
Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. Teacher performance pay: Experimental evidence from India. *Journal of Political Economy* 119, no. 1:39–77.
Nalebuff, Barry J., and Joseph E. Stiglitz. 1983. Prizes and incentives: Towards a general theory of compensation and competition. *Bell Journal of Economics* 14, no. 1:21–43.
NBS (National Bureau of Statistics of China). 2014. China statistical yearbook.
Neal, Derek. 2011. The design of performance pay in education. In *Handbook of the economics of education*, vol. 4, chap. 6, ed. Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, 495–550. Amsterdam: Elsevier.
Neal, Derek, and Diane Whitmore Schanzenbach. 2010. Left behind by design: Proficiency counts and test-based accountability. *Review of Economics and Statistics* 92, no. 2:263–83.
OECD (Organization for Economic Cooperation and Development). 2009. Evaluating and rewarding the quality of teachers: International practices. Technical report, OECD, Paris.
———. 2013. PISA 2012 technical report. Technical report, OECD, Paris.
Pham, Lam, Tuan Nguyen, and Matthew Springer. 2017. Teacher merit pay and student test scores: A meta-analysis. Working paper, Vanderbilt University.
Pianta, Robert C., and Bridget K. Hamre. 2009. Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher* 38, no. 2:109–19.
Romano, Joseph P., and Michael Wolf. 2005. Stepwise multiple testing as formalized data snooping. *Econometrica* 73, no. 4:1237–82.
Schmidt, William H., Nathan A. Burroughs, Pablo Zoido, and Richard T. Houang. 2015. The role of schooling in perpetuating educational inequality: An international perspective. *Educational Researcher* 44, no. 7:371–86.
Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J. R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2010. Teacher pay for performance: Experimental evidence from the Project on Incentives in Teaching. Nashville, TN: National Center on Performance Incentives at Vanderbilt University.
Tschannen-Moran, Megan, and Anita Woolfolk Hoy. 2007. The differential antecedents of self-efficacy beliefs of novice and experienced teachers. *Teaching and Teacher Education* 23, no. 6:944–56.
Woessmann, Ludger. 2011. Cross-country evidence on teacher performance pay. *Economics of Education Review* 30, no. 3:404–18.