Teacher performance pay:
Experimental evidence from Pakistan

Felipe Barrera-Osorio
Harvard University
felipe_barrera-osorio@gse.harvard.edu

Dhushyanth Raju
World Bank
draju2@worldbank.org

Abstract: We present findings from a randomized controlled program introduced by the provincial government of Punjab, Pakistan in 2010 that offers monetary incentives to teachers in 450 government primary schools (and 150 schools in the control group) in the form of annual cash bonuses linked to school performance. The bonus amount is set as a linear function of a composite score of school performance based on school level gains in enrollment and academic test, and the student participation rate in the test (as a deterrent against potential test-pool selection by schools.) To test whether leveraging head teachers helps, the receipt and amount of the bonus for a given level of school performance were experimentally varied by whether the teacher was a head teacher or not among three groups of schools. We do not find that the program improved average PEC test scores or raised primary enrollment. We find positive impacts on test participation rates from all treatment groups.

JEL codes: C93, I21, I28, M52
Keywords: teacher performance pay, teacher incentives, Pakistan, Punjab, government schools, teachers, field experiment

We thank the Government of Punjab's School Education Department and the Program Monitoring and Implementation Unit and, in particular, Iqbal Awan, Farah Masud, Naseem Nawaz, Sohail Raza, and Fiaz Ahmed Sajid for extensive collaboration with and assistance to the study team; Amit Dar, Sofia Shakil, and Huma Waheed for their encouragement and support for the study; and Alejandro Ganimian for outstanding research assistance. The Government of Punjab's Education Sector Reform Program, including the improvers' bonus program which is the subject of interest in this study, received financial and technical assistance from the Canadian Department of Foreign Affairs, Trade, and Development (DFATD), the United Kingdom Department for International Development (DfID), the Weiss Family Program Fund at Harvard University and the World Bank. We gratefully acknowledge financial support for the study from the Australian Department of Foreign Affairs and Trade (DFAT) and the World Bank. The findings, interpretations, and conclusions expressed herein are our own and do not necessarily represent the views of the Government of Punjab, UK DfID, Canadian DFATD, Australian DFAT, Harvard University, or the World Bank. All remaining errors are our own.

1. Introduction

Teacher performance matters for student achievement (Hanushek and Rivkin 2010; Rivkin, Hanushek, and Kain 2005; Aaronson, Barrow, and Sander 2007; Staiger and Rockoff 2010). Teacher performance depends on both teacher capability and teacher motivation and effort. While ensuring minimum teacher capability has traditionally been an important policy goal, ensuring minimum teacher effort has received greater policy focus relatively recently, motivated mainly by credible information revealing the prevalence and severity of low teacher effort. For example, rates at which teachers are absent from school are found to be high across many developing countries. High teacher absence rates also appear to be due to many teachers being occasionally absent rather than a few teachers being frequently absent. When teachers are present in school, large shares of them are found to be off task (Kremer, Ilias, and Glewwe, 2010; Chaudhury, Hammer, Kremer, Muralidharan, and Rogers 2006).

These patterns in teacher behavior are often posited to be linked to weak incentives for teachers to perform in government school systems, the main providers of schooling in most developing countries. Weak incentives are, in turn, often posited to be linked to shortcomings in the design and implementation of governance and accountability arrangements in these systems (Kremer, Brannen, and Glennerster 2013; Hanushek 2003 and 2006; Glewwe and Kremer 2006; World Bank 2011).

In this study, we present findings from a field experiment of a program introduced by the provincial government of Punjab, Pakistan in 2010 that offers monetary incentives in the form of annual cash bonuses linked to school performance to teachers in government primary schools. Called the improvers' bonus program, it was designed to incentivize incumbent teachers to raise school performance by raising teacher effort. Apart from some basic teaching and test preparation tips delivered at orientation meetings with program teachers as part of program administration, it was not designed to raise school performance by raising teacher capability.

The bonuses are offered to teachers on top of their salary (and allowances), the latter based on the traditional single-salary schedule. The bonus amount is set as a linear function of a composite score of school performance. The composite score is obtained from a weighted sum of three school-level indicators: (1) the gain in average score in a standardized grade-5 academic test, (2) the gain in total school enrollment, and (3) the student participation rate in the grade-5 academic test. The first two indicators are consistent with the government's main education

goals: to raise child school participation and student academic achievement, both of which are worryingly low in the country. The third indicator was added in as a deterrent against potential test-pool selection by schools. The data for the indicators come from an annual standardized academic achievement test—The Punjab Examination Commission (PEC) test—which all grade-5 students are required to take and an annual field survey of government schools—the Annual School Census (ASC)—carried out by the public education department.

The program was more or less targeted at schools was the lowest student achievement levels in the province. The evaluation school sample, from which program schools were drawn, consists of the 600 primary schools with the lowest school PEC test scores in the three districts with the lowest district PEC test scores in the province. The targeting resulted in an evaluation school sample with an average PEC test score which was roughly four standard deviations lower than that of the universe of government primary schools in the province.

Studies find that head teacher performance—as the leader of the school—is an important determinant of school performance (Water et al, 2003; Robison et al 2008; Mulford 2003; Murphy 1988 is an early review of the conceptual framework). Plausible pathways behind the positive relationship between head teacher and school performance include, among others, "instructional leadership" (e.g. the ability to create an environment conducive to learning), "transformational leadership (e.g. the ability to motivate and power up teachers) and "managerial skills" (the ability to manage the school as a typical bureaucracy) (Boyd at al 2009; Ingersoll 2011; Grissom and Loeb 2011). From a policy point of view, several high performing systems are nurturing, retaining and promoting leaders among their teacher body as a strategy to increase learning (Darling-Hammond & Rothman, R. 2011; World Bank 2011).

Based on this rich body of literature, it is possible to conjecture that raising the effort of incumbent head teachers may then be an important way to improve school performance, and (large) bonuses to head teachers may help elicit higher effort from them. (Large) bonuses to head teachers may also be a more efficient way to raise school performance than providing bonuses to all teachers in a school. To test whether leveraging head teachers helps, the receipt and amount of the bonus for a given level of school performance were experimentally varied by whether the teacher was a head teacher or not. In the first treatment, all teachers—whether head teacher or not—were eligible for "level-1" bonuses. In the second treatment, head teachers were eligible for level-1 bonuses, while other teachers were not eligible for any bonuses. In the third and last

treatment, head teachers were eligible for higher "level-2" bonuses, while other teachers were eligible for level-1 bonuses. Level-2 bonuses were fixed to be twice as large as level-1 bonuses, for a given level of school performance.

We use the same administrative data used by the public education department in calculating the bonuses (PEC test score and ASC student enrollment data) to evaluate the average causal impacts of the program on the three indicators in the school performance composite score. We evaluate one-, two-, and three-year program impacts between the 2009/10 school year (the baseline year) to the 2012/13 school year (the final follow-up year), with the program launched in the 2010/11 school year. Progressively enlarging the evaluation window allows us to describe how program impacts evolved over the three years.

Regardless of the specific treatment or treatment period examined, we do not find that the program improved average PEC test scores or raised primary enrollment. In contrast, we find that all three treatments equally raised the PEC test participation rate among grade-5 students by an average of three percentage points by the end of the third year of the program, which maxes out the average test participation rate from the baseline average of roughly 97%. While the program impact on the test participation rate is statistically significant, it does not provide a large payoff: given the small weight in the school performance composite score assigned to the test participation rate, the estimated impact translates into a nominal increase of 4,500 Pakistani rupees (US$43, at the current exchange rate) in the bonus amount, or a X% increase in the average bonus amount in any given year.[1]

We cannot rigorously identify what were the barriers to positive program impacts. Using our administrative data, which are basic and limited, we find that the average levels in school and teacher characteristics were not statistically different between the experimental groups three years after program inception.

In the final part of the paper we discuss several potential explanations for the results. We argue that the results are not due to problems in implementation or low bonus sizes. If something, the program was very well executed, and bonus sizes are large. Three potential

[1] The PEC and ASC data are used by the public education authorities for a wide range of decisions which potentially affect schools. The bonus program has made these data more high stakes for beneficiary teachers, amplifying incentives for teachers to behave in ways that may produce artificial gains in the measures. Indeed, high-stake programs are frequently documented to (only) result in significant gaming (see, for example, Jacob 2005). The evaluation of program impacts being limited only to these high-stake outcome measures is arguably less of a concern in the absence of program impacts.

explanations emerge as the most prominent culprits: first, teacher effort may not have been one of the primary constraints to higher outcomes, especially in schools targeted by the program. Second, teachers may not have known what exact strategies to pursue to raise student achievement and school enrollment. In other words, they may not know the production function of test scores (Fryer, 2013). Third, even if the teachers know the production function of test and students attendance, the actions they take depend as well on the cost of performing these tasks. Arguably, teachers choose the less costly action or the action they know how to perform: increase test participation.

The rest of the paper is organized as follows. Section 2 briefly reviews the evaluative research on the impacts of teacher incentive programs. Section 3 provides background information and describes the program. Section 4 discusses the data, sample, and identification and estimation strategies. Section 5 discusses the impact findings. Section 6 summarizes the main findings and discusses implications.

2. Related literature

Performance pay initiatives for teachers are relatively rare in public school systems. Examples exist in the form of small pilots as well as at significant scale (e.g., school district wide), in both developed and developing countries. The number of examples is growing at a fast pace, as governments and other actors increasingly consider teacher performance pay as a promising instrument for raising student academic achievement and explore its use (Fryer 2013).

The impacts of some of teacher performance pay initiatives have been rigorously evaluated, although the number of such evaluations is small. The evidence on program impacts from these evaluations is inconclusive: some studies consistently find success, others find mixed success, and yet others consistently find no success across examined measures. Findings from selected studies that use experimental data designs are discussed next, organized by our categories of program success.

Successful programs

Duflo, Hanna, and Ryan (2012) experimentally evaluate a program that combined monitoring of teacher presence by students and teacher salary linked to teacher presence in nonformal education schools run by an NGO and staffed by contract teachers in tribal villages in Rajasthan,

India. They find that the program increased teacher presence, and this impact lasted beyond the life of the program. The program also increased the amount of time students were taught, student test scores, and the likelihood that students would transfer to formal education in government schools.

Muralidharan and Sundararaman (2011) and Muralidharan (2013) experimentally evaluate a Foundation-run program that provided group and individual cash bonuses to teachers in rural government schools in Andhra Pradesh, India. The bonuses were linked to average student score gains in Foundation-administered tests. The studies find that both types of bonuses increased average student scores in both bonus-linked tested subjects (math and language) and unlinked tested subjects (science and social studies) in every program year over a five-year period. Starting from the second program year, the impacts of the individual bonuses were larger than that for group bonuses. By the end of five program years, group bonuses had increased average student test scores between 0.3–0.5 standard deviations, depending on the tested subject.[2]

Mixed/conditional success

Some studies provide a picture that is more mixed and less sanguine picture on the impacts of performance pay, or one where impacts hinge on certain program design or agent behavioral conditions and developments. Kremer, Glewwe and Illias (2010) experimentally evaluate an NGO-run program which provided group in-kind rewards to teachers for average student performance in district standardized exams, with penalties for students who did not participate in the exams. The program was introduced in a set of rural government schools with below-average student exam performance assisted by the same NGO in western Kenya. They find that the program increased student exam participation and exam preparation. The program also increased average student exam scores by about 0.1 standard deviations in the second program year, but the impact appears to have faded out after the program ended. The program did not change teacher presence or instruction, and did not raise average scores in tests not directly linked to the rewards.

[2] Lavy (2002) evaluates group cash rewards to teachers awarded to the top-third best schools measured by the average number of credits per student, the share of students receiving matriculate certificates, and the school dropout rate in a set of secondary schools in Israel. Using alternative quasi-experimental data designs, the study finds that the program had positive impacts on the three reward-linked outcomes as well as average test scores.

Behrman, Parker, Todd, and Wolpin (2012) experimentally evaluate a program that offered cash rewards for levels of student performance in standardized math exams in a set of federal high schools in Mexico. Three program variants were tested. The first variant only offered rewards to students linked to their individual exam performance. The second variant only offered rewards to math teachers linked to the aggregate exam performance of their students. The third variant offered rewards to students linked to their individual exam performance as well as the average exam performance of co-students; rewards to math teachers linked to the aggregate exam performance of theirs students as well as the performance of students of math teachers for other grades; and rewards, fixed as a fraction of the average reward to math teachers, to other subject teachers and school administrators. The two variants that offered rewards to students were found to have positive impacts on average student math exam scores. The estimated impacts were largest (0.3–0.6 standard deviations) for the variant that rewarded both students and teachers. The researchers argue that the findings suggest that teachers exert higher effort when students exert higher effort (such as in response to rewards for students), and that adding on rewards to teachers induces even higher effort from both teachers and students.

Fryer et al (2012) experimentally evaluate a program that offered individual or group cash bonuses to teachers in public schools that cater to low-income students in Chicago Heights, US. The bonuses were linked to average student score gains in a standardized assessment. The timing and framing of the bonuses were also experimentally crossed in: bonuses were offered at the beginning or end of the school year. Teachers offered bonuses at the beginning of the year were expected to be returned in proportion to how far below average student performance fell relative to the broader average test score gain over the school year ("loss" framing). Bonuses offered at the end of the year were in proportion to how far above average student performance rose relative to the broader average test score gains over the school year ("gain" framing). The loss-framed bonuses are found to raise average student scores by 0.2–0.4 standard deviations in bonus-linked and -unlinked assessments while the gain-framed bonuses did not. Whether framed as losses or gains, the impacts of the individual and group bonuses were found to be similar. The experiment demonstrates that power of loss aversion for teachers in inducing higher effort, and thereby, yielding positive impacts on student academic achievement.

Unsuccessful programs

Fryer (2013) experimentally evaluates a program that offered per-teacher cash rewards to a set of public schools that cater to low-income students in New York City, US. The school rewards were typically distributed to teachers as group bonuses; they were linked to performance targets on school report cards. Report card scores were based how well the school performed—measured by, as relevant, quality of school learning environment measures, student attendance, the level and gain in student performance in standardized state assessments/examinations, course credits, and the graduation rate—relative to the aggregated performance of comparable public schools and public schools in the city. The study did not find positive impacts on any of the measures on which the report card scores were based, nor on student performance in other unlinked assessments, student disciplinary incidents, or multiple, diverse measures of teacher behavior.

Springer et al (2011) experimentally evaluate a program that offered math teachers in a set of public middle schools in Nashville, US the option to receive cash rewards. The rewards were linked to average student score gains in a state standardized assessment. The study finds that the program did not raise average student math assessment scores. The program also does not appear to have altered teacher behavior, measured variously.

It is difficult to draw general lessons from the available evidence. It is unclear, for example, what specific conditions and pathways (or the lack thereof) are behind (the absence of) positive impacts. As the description of the various performance pay experiments show, the design of programs differs in many fundamentally ways. Program contexts vary. Notwithstanding, evaluated programs tend to have targeted schools where student academic achievement is relatively low. Documented program impacts vary widely in size, but can be large.

3. Background and program details

A. Background

Like the rest of the Pakistan, Punjab province, where the pilot improvers' bonus program is being implemented, has low levels of child school participation and student academic achievement. In 2012/13, the net enrollment rate at the primary level was estimated at 72% in (Pakistan Bureau of Statistics, Federal Government of Pakistan). In 2007, students in grade 3 in rural communities scored 30%, 29%, and 38% in independently-administered, competency-based

tests in English, Urdu, and mathematics, respectively (Andrabi et al 2007). There has been modest improvement, at best, in school participation and student academic achievement over the last decade.

The government school system is the main provider of education in Punjab, accounting for 65% of primary school students in 2012/13. It is a massive system, composed of 54,392 functional schools. Out of this number, roughly 70% are primary schools (i.e., have preschool to grade 5 only) with 103,955 assigned teachers and about 2.6 million enrolled children (October 2012 Annual School Census, School Education Department, Government of Punjab).

Teacher performance in the government school system appears to be poor. A comparison of teacher performance in government versus private schools points to differences in incentives as driving differences in performance. Andrabi et al (2009) find that government school teachers tend to have higher levels of academic qualifications, teaching experience, and professional teacher training than private school teachers. Despite the higher credentials of government school teachers, the levels of teacher presence in school and student test scores are lower in government than private schools. In addition, while the salary level for government school teachers is about three to four times higher than that for private school teachers after adjusting for differences in credentials and other characteristics between teachers, government school teacher salaries do not vary with teacher competency, measured through teacher test scores, and student competency, measured through student test scores. Higher government school teacher salaries are also associated with *lower* teacher presence. In contrast, higher private school teacher salaries are associated with higher teacher presence, teacher test scores, and student test scores.

B. The improvers' bonus program

Against this backdrop, the improvers' bonus program was designed and approved by the provincial government in FY2009/10 for implementation on a pilot basis starting in FY2010/11 (the 2010–11 school year). The main aim of the program was to encourage greater teacher effort by conditioning cash bonuses on improving school participation and student achievement. In designing the program, we review carefully the literature on pay per performance (Neal 2011 provides an overview).

Measuring school performance

The main considerations of the provincial government in selecting indicators to link bonuses to were that the indicators were consistent with its main education goals (raising school participation and student achievement); they could be measured regularly, widely, and in a standardized way across schools; and the data used to measure them were arguably least susceptible to manipulation by schools. The government possessed two available data sources that satisfied these criteria: the Annual School Census (ASC) surveys and Punjab Examination Commission (PEC) tests. These instruments however only allow the measurement of performance at the school level, not at the teacher level.

The ASC is an annual interviewer-administered field survey of government schools which captures basic school and teacher level information, including student enrollment by grade and gender. The PEC test is an annual, standardized academic test that all students in grades 5 and 8 are required to take.[3] The tests are curriculum linked, cover multiple subjects, and use selected and constructed response question types. The provincial education department is responsible for managing all stages of the ASC survey and PEC test activities. It is solely responsible for the design of the data collection instruments. It shares the responsibility of administering the instruments in the field with district education departments.

Three indicators were selected for measuring school performance using the ASC survey and PEC test score data: (1) the gain in school-level average test scores (*GTS*), (2) the gain in school enrollment (*GSE*), and (3) the school test participation rate (*TPR*). *GTS* is defined as the year-over-year gain in the school-level percent score in the grade-5 PEC test, measured in percentage point terms. Instead of targeting an absolute value, which will induce higher effort from teachers towards students close to the cutoff point, we decided to target changes in score, incentivizing teachers to exercise effort towards all students, and perhaps, more effort towards weakest students with more room to improve (Barlevy and Neal 2012) The test scores were restricted to raw scores—the number of test questions answered correctly—for the multiple choice question modules in the subjects of Urdu, English, Islamic studies, mathematics, science, and social studies.[4] We used five subjects, in contrast to math and language, to avoid reallocation

[3] In practice, all government schools comply, while only a portion of private schools do.

[4] Raw scores were used instead of the scores scaled by PEC because of concerns related to the conversion formula that was applied. The scoring of answers to multiple-choice questions was considered more reliable than the scoring of answers to open-ended questions. Filled-in multiple choice scoring sheets were delivered to PEC headquarters to be read by Optical Mark Readers while answers to open-ended questions were scored by government school teachers at grading centers in the districts. A wider set of subjects were tested; several of the other subjects were

of time to only those subjects rewarded (in order to avoid the problem arisen in multitask settings; Holmstrom and Milgrom 1991). *GSE* is defined as the year-over-year gain in enrollment in grades one through five, transformed into percent terms. The *TPR* was included as an indicator mainly to protect against schools gaming performance by positively selecting academically stronger students to register and take the PEC test (this is one of many unintended reactions from part of teachers in pay per performance; see Koretz 2002). The *TPR* is defined as the number of students who took the grade-5 PEC test divided by the number of enrolled students in grade 5, transformed into percent terms. Note that, while *GTS* and *GSE* measure change, *TPR* measures level. The values for each of the three indicators are either naturally or artificially bounded at 0% and 100%.

A composite school performance score (*CS*) was formulated as follows:

$$CS = 0.65XGTS + 0.25XGSE + 0.15XTPR \qquad (1)$$

In (1), *GTS* is weighted highest, followed, in turn, by *GSE* and *TPR*. The sizes of the assigned weights roughly reflect the provincial government's relative valuation of the three indicators in assessing school performance.

Program treatments

In principle, teachers officially designated head teacher status (including those who hold this status on an acting basis) are authorized to manage their schools and teachers. Head teachers can manage teachers through alternative means, including supervising, rewarding, sanctioning, and supporting teachers. If head teachers are a sufficiently effective instrument for managing teachers and their performance, offering bonuses to head teachers may be a more efficient way to raise school performance than offering bonuses to all teachers in a school.

These views served as the basis for the design of three program treatments, which varied by bonus eligibility and size along head teacher/teacher lines.[5] In the first treatment variant (*HT only*), only head teachers were eligible for level-1 bonuses. In the second treatment variant (*T*), both head teachers and teachers were eligible for level-1 bonuses. In the third treatment variant

assessed through practicum tests. The selected subjects were viewed as core subjects; they were assessed through pencil-and-paper tests.

[5] A teacher appointed as head teacher on an acting basis is considered eligible for the head teacher bonus. A (head) teacher whose assignment to a program school is effective at the start of the school year, even if the (head) teacher is then transferred at any point during the school year to any another school, is considered eligible for the (head) teacher bonus.

(*HT+*), head teachers were eligible for higher level-2 bonuses, while other teachers were eligible for level-1 bonuses. The size of the level-2 bonus was set as twice the size of the level-1 bonus, for a given level of school performance. In treatments *T* and *HT+* we decided to rewards all the teachers in the school in order to avoid problems unfair competition, at the risk of free rider problems (Holmstrom 1982). The evidence of India points towards positive effects of both individual and group bonuses (Muralidharan and Sundararaman 2011).

A rich literature on leadership and educational outcomes motivated the decision to provide different incentives to headteachers (Robison et al 2008 provides an overview). Particularly important is the idea of "instructional leadership": headmaesters can perform actions that are conducive to higher learning. For instance, headmasters can monitor classes and help teachers; headmasters can reallocate strategically teachers; headmasters can retain better teachers in the schools (Grissom and Loeb 2011; Murphy 1988).

The specific bonus formula, by treatment type (*HT* only, *T*, and *HT+*) and teacher type (head teacher or teacher), are present in Table 1. The slope parameters in the bonus formula were calibrated so that teachers in the median school in the target sample would face high-powered incentives, while, at the same, forecasted program expenditures for a wild-card scenario of maxed scores by all program schools fell within the approved budget. For the median school, maxing the composite score in one year would mean maximum bonus awards of X and Y Pakistani rupees for head teachers and teachers, respectively. These bonus awards represent X% and Y% of the starting basic annual salary at program inception of a BPS–14 teacher (the mean pay grade for a head teacher in a primary school) and BPS–9 teacher (the mean pay grade for primary school teacher), respectively.

Program targeting

The provincial government targeted the program at government primary schools with the lowest average academic achievement following a two-step process. In the first step, the three districts (out of the 36 districts in the province) with the lowest district-level average scores in the 2010 PEC grade-5 test were identified. In the second step, functional, primary schools in these districts, based on the October 2009 ASC, with at least ten students who took the 2010 PEC grade-5 test were identified. There were 1,963 such schools. These schools were ranked from highest to lowest in school-level average scores in the 2010 PEC grade-5 test, and the bottom

600 schools were selected to form the evaluation sample, out of which schools for the program were drawn.

Program implementation

Field visits and score card distribution

In the first and second program years (FY2010/11 and FY2011/12), staff from the provincial education department and partnering organizations conducted two field visits, roughly in the middle of the 2010/11 and 2011/12 school years. During each visit, teachers from program schools were invited to attend group meetings—organized separately by treatment-specific, spatially-clustered school groupings—to become acquainted with the relevant treatment and to register for the program, which entailed filling out a form that captured, among other things, employment details and the bank account information of the teacher.

Roughly two months after the end of the 2010/11 and 2011/12 school years, school performance score cards, which reported the school's *GTS*, *GSE*, *TPR*, and *CS* values for the school year that just ended, were sent from the provincial education department to the district education departments with instructions to send the cards to program schools and to have them displayed in a prominent, open location on school premises. These cards were printed on large, durable (synthetic flex) paper.

Administration of the registration and orientation meetings and the distribution of school performance score cards differed somewhat in the third program year (FY2012/13). Staff from the provincial education department and partnering organizations conducted two visits, one in the middle of the 2012/13 school year and the second two months after the end of that school year. The first visit was structured similarly to the visits in the previous years. There were however three key additions in the contents: program teachers (1) received guidance on what they could do to raise school participation and student achievement, (2) participated in an experience-sharing session, and (3) were provided leaflets with basic teaching and test preparation tips. In the second visit, only head teachers were invited to group meetings. The group meetings included the same three additions, in relation to basic school leadership. Program head teachers were also directly handed school performance score cards (with a new format) to take with them and display at school.

Bonus calculation, fund flow, and award transfer

In each of the program years, calculations of the school performance scores and bonus awards were performed by the provincial education department, with technical support and quality control from the partnering organizations. The bonus awards, aggregated up to the district level, were communicated to the provincial finance department. The provincial finance department transferred the requested total bonus award for a district to an account dedicated to the program operated by the district education department. Concurrently, the provincial education department sent program school and teacher lists with the bonus award for and personal bank account information of the program teacher to the district education department. The district education department used the funds in the dedicated account and followed the instructions provided by the provincial government to electronically transfer the bonus awards directly into the personal bank accounts of intended beneficiaries. The last step was to be carried out after the school performance score cards were distributed to program schools, but, in practice, the sequencing was not strictly followed, mainly because the government was racing against the end-of-fiscal-year clock.

Implementation performance record

Provincial government and third-party reviews find that the record of program implementation performance has been generally satisfactory starting from the first year of implementation. School performance score and teacher bonus award calculations and the distribution of school performance score cards have by and large been carried out without any major issue. Bonus fund flow from the province to the district has been smooth and swift. The full bonus awards reached intended beneficiaries.

Although it did not turn out to be critical, there was some slippage in implementation. Some stages of implementation were rushed because they were commenced late and had to be completed before binding deadlines. The field visits for program teacher orientation and registration were conducted two-three months later in each school year than considered optimal.

4. Data and empirical strategy

A. Data and sample

Four years of annual rounds of the ASC survey and PEC test score data were collapsed to the school level and linked together using the government's unique school identification codes (the Education Management Information System [EMIS] code). The structure and contents of the two data sources were discussed earlier in Section 3. The October 2009 ASC survey and January/February 2010 PEC test rounds serve as our pre-program data (the program came into effect in July 2011). The October 2013 ASC survey and January/February 2013 PEC test rounds serve as our third program year data.

Evaluation sample schools were largely linked without issue across the various databases. Very few schools could not be linked because the schools were merged into other schools as part of the public education department's province-wide school merger program. As a result, 583 schools is the study sample in 2013. The merger resulted in the schools losing their status as a separate school; merged schools were provided the same EMIS code as the school they were merged into. The percentage of schools lost from the evaluation sample across the treatment groups was similar: 2.7 and 3.3 for treatment and control groups respectively.

Column 1 in table 2 presents pre-program mean characteristics of our evaluation sample schools which survived linking across data rounds. Panel A in the table contains basic input indicators and panel B contains outcome indicators. Ninety-two percent of the evaluation sample schools were located in urban areas. They had, on average, (close to) 3 teachers and classrooms. Eighty-six percent of them had buildings rated as in satisfactory condition or needing at most basic repairs, and, on average, 3.19 out of four basic amenities (drinking water, electrical connection, toilets, and sewer connection). Evaluation sample schools had, on average, an enrollment of 99.98 children in grades one through five, student-teacher and student-classrooms ratios slightly over 50:1, a PEC test participation rate of 97%, and a PEC test, standardized to the mean, of 0.

Pre-program characteristics of evaluation sample schools are compared to government primary schools in the province at large. Column 2 reports pre-program mean characteristics of 42,438 functional, primary schools, while column 3 reports the size and statistical significance of differences in means between evaluation school schools and the rest of schools. Evaluation sample schools are older but have larger and better school infrastructure. They have, on average, more classrooms and basic amenities. They have a higher proportion of safe buildings. Evaluation sample schools are similarly distributed across urban/rural areas as all schools but are

more likely to be coeducational. Evaluation sample schools have larger student bodies and are more crowded. They have, on average, higher enrollment in grades one through five, enrollment in grade five, and student-teacher and student-classroom ratios. Finally, evaluation sample schools have, on average, a higher PEC test participation rate and substantially lower PEC test scores. The mean PEC test score for evaluation sample schools is roughly four standard deviations lower than the corresponding statistic for all schools. The low student achievement level in the evaluation sample schools is a natural outcome of how the sample was constructed as discussed in Section 3.

We present summary statistics of the main four outcome variables in Table 3 for the study sample by year. Enrollment size has declined since 2010. On average, schools had 100 students and 86 students in 2010 and 2013, respectively. The percent of test takers is very high and without a clear time trend; it fluctuates between 92 and 97 percent. The test scores show a steady increment from a standardized mean (with respect to the control group at baseline) of 0.0 sd to 2.49 sd. The increment applies to the whole distribution of test scores (Figure 2). The most probable explanations to this trend are twofold. First, the change can obey changes in the format and difficulty of the test. Second, PEC test has become more salient with the pass of time; schools may be teaching in a systematic way the format and questions from past years.

B. Empirical strategy

The evaluation is based on a randomized block design. The 600 evaluation sample schools were randomly assigned into our four treatment statuses (*HT only*, *T*, and *HT+*, and control status) of 150 schools each, after blocking by school average 2010 PEC test scores. This design ensures that any differences in average outcomes of interest between our treatment statuses are attributable to the treatment and not due to systematic differences in characteristics between the statuses prior to treatment.

We estimate intention-to-treat (ITT) effects by regressing the school (S) mean value in an outcome variable (Y_s), against treatment indicators ($T_{i,s}$, for i=1, 2 and 3, each of them representing treatments *HT only, T* and *HT+* respectively) and a set of covariates measured at baseline ($X_{s,2010}$):

$$Y_{s,t} = \beta_0 + \sum_{i=1}^{3} \beta_{i,1} T_{i,s} + BX_{s,2010} + \gamma_d + \gamma_b + \varepsilon_{s,t} \qquad (1)$$

where γ_b and γ_d denote fixed effects by randomization block and district, respectively. Standard errors are clustered at the tehsil level. We estimate several variations of the basic specification: first, we present results from pooling together treatment (T_s equal one if treated, zero otherwise) and effects by treatment ($T_{i,s}$, i equal 1, 2, or 3); second, we present results with and without baseline controls[6]; finally, we present results of estimating equation (1) for each year separate and also results of estimating the panel data at the school level, all years. For the last specification, we included time fixed effects. For test scores we used student level variation, in contrast to school average.

Table 2 also presents pre-program mean characteristics for evaluation sample schools, separately by pooled treatment assignment status (treated schools in column 4 and untreated schools in column 5), and the size and significance of differences in means between the two treatment assignment statuses in column 6.[7] Pre-program mean characteristics of schools in the two statuses appear to be balanced. We do not find statistically significant differences in any of the pair-wise comparisons of means. We also test whether the differences in means for the two panel-grouped sets of characteristics between the two statuses are jointly zero. The test statistics for outcomes, reported in the table, are not statistically significant as well. All in all, the table confirm the balance between treatment and control schools.

Table 2 present an additional validation of the experiment. Columns (7) and (8) present the mean and SD of baseline characteristics for the set of treatment and control schools that we observe in 2013; in column (9) the differences between them. While not exactly as an attrition test, since the loss of schools respond to merges in contrast to real attrition, the exercise allows to test balance between treatment and control groups in the 2013 school sample. We don't find any systematic differences between the final schools sample in baseline characteristics.

5. Findings

A. Impacts

[6] School controls, constructed using pre-program (October 2010) Annual School Census data, comprise of (1) year the school was established, (2) indicator for whether school is rural, (3) indicator for school building is in satisfactory condition/requires minor repairs only, (4) count of selected basic amenities (drinking water, electricity, toilets, and sewerage), (6) number of assigned teachers, (7) student-teacher ratio, and (8) student-classroom ratio.
[7] We tested balance between treatment and control groups for each experimental group; results were similar.

Tables 4 to 7 show the main result in each of the outcome variables for years 2011, 2012 and 2013: enrollment (Table 4), test-takers rate (Table 5), standardized test scores (Table 6), and the composite index (Table 7). The structure of these four tables is the same: columns (1) and (3) show results without baseline controls; columns (2) and (4) show results with controls. Columns (1) and (2) present the average impact of the program by comparing all treated schools versus control schools. Columns (3) and (4) show the effects for each treatment scheme.

Table 4 presents the results for enrollment. As shown in Table 3, all schools in the sample are losing students. The constants term of all Columns (1)—the mean for the control group—shows a negative time trend: mean control enrollment for years 2011, 2012 and 2013 are, respectively 93.6, 87.6 and 83.5 students. This negative trend is similar between treatment and control groups. All the coefficients of impact of the program on the pooled specification are not statistically significantly different from zero. Apart from negative effects from Treatment 3 in the first year (estimate of -4.65 with controls), the other estimates by treatment are not statistically different from zero.

Table 5 shows the results of the program on the test-taker ratio. In the first year of the program implementation, treatment schools presented a slightly lower ratio (however, not statistically significantly different from zero). At the end of the third year of implementation, there is an increment of (close to) 3 percentage points in test takers in the treatment schools. These effects are present across all three treatments. Treatment 3 shows effects since the second year of implementation (the difference in coefficient between treat 1 and 3 is statistically significant). In short, there is a positive trend and by year 2013, all treatment schools increased the ratio of test-takers.

Table 6 present the results for students' test scores, all of them standardized to the control group at baseline.[8] In contrast to the two previous outcomes, in this case there is not clear pattern in the effects of the program. For the most part, the program has not had statistically significant positive effects on test scores. Treatment 3 is the only one that shows consistent positive point estimates, and only in the year 2012 is the effect statistically significant (0.44 sd). For completeness, we present impact in the index, *CS* (Table 7). Again, there are not clear trends in the impact of the program across years and across treatments.

[8] Test scores is the only variable for which we exploit student variation. We also ran a regression using schools' average test (not shown here); as expected, the results are similar.

We confirm these results by using the panel nature of the data. Table 8 presents the estimation by pooling across all years, using schools' average for all four outcomes. The model controls for time fixed effects and baseline characteristics. The only statistically significant estimate is the one associated with the test-taker ratio, driven by treatment 3 effects. Overall, treatments 1 and 2 show negative point estimates for changes in test scores. In contrast, treatment 3 presents positive results (0.271 sd), although not statistically significant.

Three main messages emerge from these four tables. First, the only variable that shows systematic results is the test taker rate in 2013. The original intent in including the test taker rate in the bonus formula was to reduce the likelihood of test score manipulation (e.g., teachers changing the pool of test takers with the aim of increasing average test scores). Among the three variables that the program was based on, the test taker ratio was the factor most easily controlled by teachers and headmasters. The coefficients of the three treatments are close to 3 percentage points; the mean test taker rate in Punjab was 93%.

Second, there seems to be a positive trend in the impact on two outcome measures: enrollment and test taker ratio. By 2013, the point estimates for those two variables are all positive—although most are not statistically significant—, while the point estimates on average were negative in 2011. Third, is the results suggest that there is a differential impact from treatment 3 (bonuses for all teachers with the head teacher receiving more) in contrast with the impacts from treatments 1 and 2. The point estimates for treatment 3 with respect to test scores are always positive, and in one year significant. In contrast, the point estimates for treatments 1 and 2 are usually negative and not statistically significant. We read this result as merely suggestive.

B. Heterogeneous impacts

One potential explanation for the (lack) of results may reside in heterogeneity responses to the program. In theory, a better performing school at baseline may know better the production function of test scores than a lower performing school. In light of this, we hypothesize that better performing schools at baseline would respond to the incentives with improvements in the test scores. Nevertheless, the production function knowledge may be attenuated by the fact that the program rewards changes in tests scores, and a school that starts with higher test scores has less

room for improvement. In short, the idea of testing for heterogeneous effects emanates from the margins that schools can affect in response to the program.

The original randomization (e.g. using blocks ordered by baseline test results) was done with two objectives: to maximize the probability of balance across treatments and to explore differential responses of schools by level of test performance. Schools were organized by block randomization, and an indicator variable was created with one for schools with block randomization above 75 and zero otherwise. Given the lottery, this indicator variable is one for schools with high average baseline test score, and zero for schools with low average baseline test score.

Table 9 presents results from potential heterogeneous effects. We find no substantial differences in effects between high and low baseline schools. The estimation uses the values at 2013 of the four main outcomes variables. The table shows the coefficient of impact for the estimation with all controls. The most consistent result of the program—e.g. changes in test taker rates—is present for both types of schools. Furthermore we do not find any differences in the other outcomes (enrollment and test scores) or by treatment scheme.

C. Retention and inputs changes

Even in the absent of results in the outcome variables of interest, schools may take (ineffective) actions in response to the program. Our data allow testing two potential responses of the schools to the program: changes in school inputs and student retention before grade 5.

An important discussion in the literature of economics of education is the tension between "more resources" versus "better incentives" (Hanushek, 2003). In short, analyses of large datasets for several countries show that, after certain minimum level of investment in education systems, more resources are not necessarily translated to better educational outcomes due to the lack of correct incentives, especially for teachers. Our results can be compatible with schools changing the input mix without improvements in learning outcomes.

The data allow us to investigate changes in key school inputs. As a response, schools may have invested in some inputs such as infrastructure of the school and number of teachers, but these changes were not once to improve learning. Table 10 presents the results from an estimation of key inputs of the school, including infrastructure variables—such as presence of playgrounds, toilets, and libraries—; changes in student-teacher ratios, changes in the number of

teachers, and changes in the gender of student body. The estimation is based on the regression of each input (levels) versus the indicator of treatment, individually and pooled, for the last year of data available (2013), under the assumption that effects would be more salient after three years of implementation.

The results are very clear: there are no differences in the inputs (that we measure) between control schools and treatment schools. This is true when the three treatments are considered jointly and when each individual treatment is considered separately. All in all, the table supports the null hypothesis of no change in inputs as a result of the program.

Potentially the school could also keep back weaker students so that only the best students are admitted to 5^{th} grade in an effort to boost test scores. Table 11 presents results of, first, aggregate enrollment in grades 1 to 4, and, second, enrollment in grade 5. When all the treatment schools are pooled together there are no-statistically significant differences in the average enrollments. When separating effects by treatment does not give any consistent results. All in all, there is no evidence of more students in Grades 1 to 4 in any of the treatments arms

6. Discussion

All in all, we only show modest effects of the program on the test-taker ratio, and no effects on enrollment or tests. Also, there is some (weak) evidence of larger responses to treatment 3 than the other two treatments. Impact heterogeneity across baseline test scores is not an explanation for the (lack of) aggregate results. Furthermore we reject two hypotheses regarding school responses to the program: there is no evidence of input changes in the treatment schools, and there is no evidence of student retention in lower grades.

There are several factors that can explain these results. We discuss the most predominant ones based on accounts gathered from program administrators and beneficiaries and information from program implementation reports.

We start off by touching on usual suspects that we believe we can rule out as being behind the absence of positive program impacts. A first suspect is that implementation may have fallen short of design in some way. For example, a new program may experience teething problems (see, e.g., Bouguen, Filmer, Macours, and Naudeau 2013). In such cases, positive program impacts may only materialize well after program inception, when the problems are effectively resolved. Separately, implementation performance may be sensitive to implementer

characteristics such as the implementer's interest, commitment, capability, and effort (Allcott and Mullainathan 2012), and governments may be more likely or have larger shortfalls in these characteristics relative to, for example, (certain types of) NGOs (Bold, Kimenyi, Mwabu, Ng'ang'a, and Sandefur 2013).

While we evaluate a government program over the first three years of implementation, we discount that the absence of positive program impacts was due to program implementation failures. In many respects, the program appears to have been implemented satisfactorily, and markedly better than many other programs administered by the public education department. Program registration and orientation, the distribution of school performance score cards, and the transfer of bonus awards to teacher bank accounts were carried out in a timely, efficient, and full way. The PEC test and ASC survey activities were also carried out as scheduled. Credibility in the program was reportedly low in the first year of program implementation: program teachers conveyed that they were initially skeptical that they would receive bonuses (consistent with program rules). This skepticism appears to have died away after program teachers saw how the government performed in running the program and they received their bonus awards.

We also discount that the absence of positive program impacts was due to bonus awards being too low, the bonus formula being too complex for teachers to comprehend, or the group-based nature of the bonuses inducing significant freeriding. The average bonus award was roughly X of the pre-program average annual basic salary of teachers at the typical pay scale in government schools. This relative size is several percentage points higher than, for example, the relative bonus sizes in teacher incentive field experiments in India (Muralidharan and Sundararaman 2011) and Kenya (Glewwe, Illias, and Kremer 2010), both of which found positive program impacts. Knowledge of elementary arithmetic is sufficient to understand how the bonus, school performance composite score, and indicators are determined. All definitions and sources of data for the bonus calculations were communicated by the program administrators to program teachers during school visits and at orientation meetings in every bonus round. Teacher bonuses were also linked to own school performance only, and not, for example, to own school performance compared to other schools, the latter formulation potentially introducing added complexity and uncertainty for teachers (Fryer 2012). The average number of teachers in program schools at baseline was three. It is likely easier to effectively counteract any incentives

for freeridership in such small schools compared to larger schools with an identical bonus design (Goodman and Turner 2012).

In contrast, we posit two hypotheses that may explain the lack of results. First, teacher effort may not have been one of the primary constraints to higher outcomes, especially in schools targeted by the program, which were those with the lowest average test scores. While program schools clearly have ample room for improvement, it may be particularly difficult to raise outcomes in these schools. Low teacher effort may be one of multiple problems program schools face, and these other problems may have been the binding constraints to improving outcomes. Consequently, teachers may have chosen to withhold higher effort or seen the effort they expended go to waste.

The second way the program may have been (very) limited results is internal to the teacher effort-outcome relationship. Even if higher effort was a binding constraint to higher outcomes, teachers may have lacked the knowhow or agency to effect change. Teachers may not have known what exact strategies to pursue to raise student achievement and school enrollment, and whether what they intended to pursue was the most efficient way to raise outcomes. In the face of this uncertainty, teachers may have withheld higher effort or misdirected their effort (Fryer 2013, Fryer 2010). Even if teachers had identified strategies to pursue, they may not have had the needed discretion to follow through, discouraging higher effort. Arguably, teachers may be (more) impotent or ignorant on how to raise student achievement and school enrollment relative to the test participation rate, resulting in the differing impacts detected across these outcomes. Finally, the reaction of teachers depends both in knowing the production function of test and attendance and in the cost of implementing these actions (Neal, 2011). It is possible that it is highly costly to attract new students to the school, or to change test scores. In contrast, it may be less costly to sit all students in the test.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95-135.

Allcott, H., & Mullainathan, S. (2012). *External validity and partner selection bias* (No. w18373). National Bureau of Economic Research.

Andrabi, T., Das, J., & Khwaja, A. I. (2009). Report cards: The impact of providing school and child test scores on educational markets. *Unpublished Working Paper*.

Andrabi, T., Das, J., Khwaja, A. I., Vishwanath, T., & Zajonc, T. (2007). Learning and Educational Achievements in Punjab Schools (LEAPS): Insights to inform the education policy debate. *World Bank, Washington, DC*.

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

Behrman, J., Parker, S., Todd, P., & Wolpin, K. I. (2012). Aligning Learning Incentives of Students and Teachers: Results from a Social Experiment in Mexican High Schools. *Unpublished Manuscript, CIDE and University of Pennsylvania*.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2013). Scaling-up what works: experimental evidence on external validity in Kenyan education. *Unpublished Manuscript*.

Bouguen, A., Filmer, D., Macours, K., & Naudeau, S. (2013). Impact evaluation of three types of early childhood development interventions in Cambodia.

Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., & Wyckoff, J. (2009). "The Influence of School Administrators on Teacher Retention Decisions." Retrieved July 26, 2010, fromhttp://www.teacherpolicyresearch.org/portals/1/pdfs/TeacherRetentionAdministrators22May2009.pdf

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

Darling-Hammond, L.; & Rothman, R. (2011) Teacher and leader effectiveness in high-performing education systems. Washington, DC: Alliance for Excellent Education.

Duflo, E., Hanna, R., & Rya, S. P. (2012). Incentives work: Getting teachers to come to school. *The American Economic Review*, *102*(4), 1241-1278.

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

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

Glewwe, P., & Kremer, M. (2006). Schools, teachers, and education outcomes in developing countries. *Handbook of the Economics of Education*, *2*, 945-1017.

Goodman, S., & Turner, L. (2010). Teacher Incentive Pay and Educational Outcomes: Evidence from the NYC Bonus Program. Program on Education Policy and Governance Working Papers Series. PEPG 10-07. *Program on Education Policy and Governance, Harvard University*.

Grissom, J. A. & Loeb, S. (2011). "Triangulating Principal Effectiveness: How Perspectives of Parents,Teachers and Assistant Principals Identify the Central Importance of Managerial Skills." American Educational Research Journal, XX(X), 1-33

Hanushek, E. A. (2003). "The Failure of Input-based Schooling Policies". *The economic journal*, *113*(485), F64-F98.

Hanushek, E. A. (2006). School resources. *Handbook of the Economics of Education*, *2*, 865-908.

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

Holmstrom, B. 1982. Moral hazard in teams. *Bell Journal of Economics* 13, no. 2: 324–40.

Holmstrom, B. and P. Milgrom. 1991. Multitask principal-agent analyses: incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization* 7 (January): 24–52.

Ingersoll (2001). "Teacher Turnover, Teacher Shortages and the Organization of Schools." Research Report. Seattle, WA: Center for the Study of Teaching and Policy, University of Washington.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago Public Schools. *Journal of public Economics*, *89*(5), 761-796.

Koretz, D.M. (2002). "Limitations "in the Use of Achievement Tests as Measures of Educators' Productivity". Journal of Human Resourses; 37(4),752–777.

Kremer, M., Brannen, C., & Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, *340*(6130), 297-300.

Kremer, M., Glewwe, P., & Illias, N. (2010). Teacher incentives. *American Economic Journal: Applied Economics*, *2*(3).

Mulford, B. (2003). "School Leaders: Changing Roles and Impact on Teacher and School Effectiveness." Paper commissioned for the 'Attracting, Developing and Retaining Effective Teachers' Activity. Paris, France: Organization for Economic Co-Operation and Development (OECD), Directorate for Education.

Muralidharan, K. (2013). Long-Term Effects of Teacher Performance Pay: Experimental Evidence from India. Manuscript, Poverty Action Lab, MIT

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

Murphy, J. (1988). Methodological, measurement, and conceptual problems in the study of instructional leadership. *Educational evaluation and policy analysis*, *10*(2), 117-139.

Neal, D. (2011). "The design of performance pay in education." In *Handbook of economics of education, volume 4*, ed. Eric Hanushek, Steve Machin, and Ludger Woessmann. Amsterdam: Elsevier.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

Robinson, V. M., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on student outcomes: An analysis of the differential effects of leadership types. Educational administration quarterly, 44(5), 635-674.

Springer, M. G., Ballou, D., Hamilton, L., Le, V. N., Lockwood, J. R., McCaffrey, D. F., ... & Stecher, B. M. (2011). Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT). *Society for Research on Educational Effectiveness*.

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

Waters, T., Marzano, R. J., & McNulty, B. (2003). Balanced leadership: What 30 years of research tells us about the effect of leadership on student achievement (pp. 1-19). Aurora, CO: Mid-continent Research for Education and Learning.

World Bank (2011) "What Matter Most for Teacher Policies: A Framework Paper" World Bank SABER, Washington.

Figure 1. Standarized test, 2010-2013

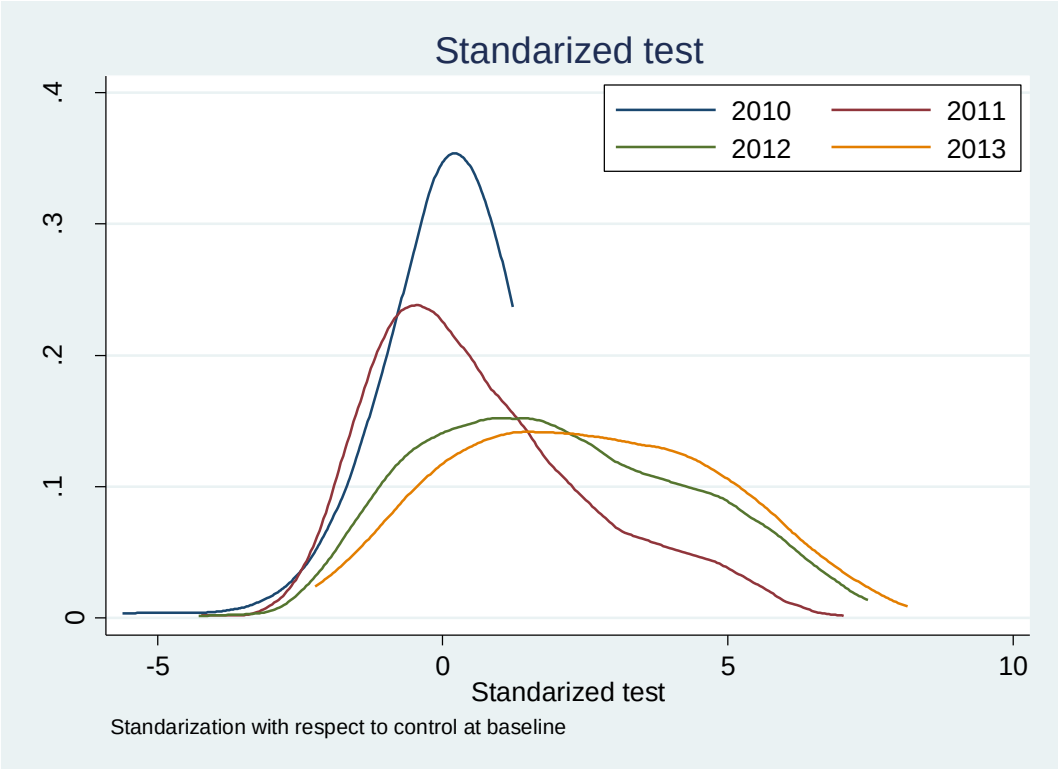


Table 1. Bonus formula, by treatment and teacher type

Bonus treatment type	Teacher type	
	Headteachers (*HT*)	Teachers (*T*)
HT only	1,000 × CS	0
T	1,000 × CS	1,000 × CS
HT+	2,000 × CS	1,000 × CS

Table 2. Baseline characteristics and Balance

				Lottery sample			2013 sample		
	Study Sample	Primary Schools	Difference	Treatment (T)	Control (C)	Difference =(T)-(C)	Treatment (T)	Control (C)	Difference =(T)-(C)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Characteristics of School									
School Age	42.67	39.56	3.112***	42.65	42.75	-0.1	42.54	42.67	-0.466
	(17.90)	(29.67)	(1.10)	(18.05)	(17.50)	(1.66)	(17.87)	(17.59)	(1.83)
Location (1=rural)	0.08	0.08	-0.002	0.09	0.06	0.027	0.09	0.06	0.027
	(0.27)	(0.27)	(0.02)	(0.28)	(0.24)	(0.02)	(0.29)	(0.24)	(0.02)
Girls school only	0.11	0.26	-0.150***	0.1	0.13	-0.024	0.1	0.13	-0.029
	(0.31)	(0.44)	(0.02)	(0.30)	(0.33)	(0.04)	(0.30)	(0.34)	(0.04)
Coed school	0.51	0.43	0.084*	0.49	0.58	-0.091	0.49	0.57	-0.08
	(0.50)	(0.49)	(0.05)	(0.50)	(0.50)	(0.06)	(0.50)	(0.50)	(0.06)
Number of teachers	2.96	2.77	0.188	2.98	2.91	0.064	3	2.91	0.078
	(1.66)	(1.81)	(0.12)	(1.70)	(1.53)	(0.10)	(1.72)	(1.53)	(0.10)
Number of classrooms	3.19	2.75	0.432***	3.19	3.16	0.034	3.21	3.17	0.042
	(1.63)	(1.68)	(0.17)	(1.64)	(1.60)	(0.14)	(1.64)	(1.62)	(0.15)
Building Condition (1= Satisfactory)	0.86	0.84	0.027**	0.86	0.86	-0.002	0.87	0.87	-0.008
	(0.34)	(0.37)	(0.01)	(0.34)	(0.34)	(0.02)	(0.34)	(0.34)	(0.02)
Utilities index (0 to 4)	3.09	2.79	0.302**	3.09	3.11	-0.018	3.09	3.09	-0.004
	(1.05)	(1.17)	(0.13)	(1.02)	(1.14)	(0.09)	(1.02)	(1.14)	(0.09)
Joint test of all coefficients equal 0						15.951			10.161
Probability						0.043			0.254
Panel B. Enrollment and test scores									
Enrollment, grade 1-5	99.98	77.8	22.187***	100.16	99.47	0.691	100.71	100.17	0.794
	(51.69)	(59.06)	(4.31)	(51.59)	(52.17)	(5.08)	(51.56)	(52.20)	(5.42)
Enrollment, grade 5	18.25	11.16	7.084***	18.4	17.79	0.604	18.43	17.86	0.609
	(9.81)	(10.82)	(0.77)	(10.01)	(9.20)	(1.04)	(9.87)	(9.21)	(1.11)
Test taker rate, grade 5	0.97	0.93	0.044***	0.97	0.98	-0.005	0.97	0.97	-0.004
	(0.06)	(0.14)	(0.01)	(0.06)	(0.06)	(0.01)	(0.06)	(0.07)	(0.01)
Pupil-Classroom ratio	52.48	44.16	8.323**	52.34	52.89	0.046	52.39	53.35	0.162
	(36.99)	(29.59)	(3.79)	(38.26)	(32.94)	(3.00)	(38.56)	(33.05)	(3.18)
Pupil-teacher ratio	56.07	44.15	11.924***	56.22	55.62	0.641	56.23	56.26	0.338
	(36.72)	(28.36)	(3.90)	(36.96)	(36.13)	(4.05)	(37.20)	(36.53)	(4.29)
Test score mean (SD)	0	4.18	-4.186***	-0.01	0	-0.005	-0.01	0.01	-0.006
	(1.01)	(3.09)	(0.17)	(1.02)	(1.00)	(0.01)	(1.03)	(0.99)	(0.01)
Joint test of all coefficients equal 0						2.079			0.862
Probability						0.912			0.99
Number of Observations (max)	600	42438	43038	450	150	600	438	145	583

Note: Columns (1), (2), (4), (5), (7) and (8): means and standard deviations. Columns (3), (6) and (9): difference and standard errors. Column (2) Census of primary, functional schools in Punjab.

Table 3. Outcome variables, 2010-2013

		2010	2011	2012	2013
Enrollment	Mean	99.98	92.58	87.04	86.01
	SD	(51.69)	(51.37)	(49.86)	(52.08)
	No. of Obs.	600	598	593	583
% Test-takers	Mean	0.97	0.95	0.92	0.96
	SD	(0.06)	(0.11)	(0.12)	(0.09)
	No. of Obs.	599	597	591	575
Test	Mean	0	0.67	2	2.49
	SD	(1.01)	(1.91)	(2.29)	(2.21)
	No. of Obs.	600	598	593	583
Index	Mean		5.02	11.92	4.43
	SD		(21.36)	(21.20)	(23.00)
	No. of Obs.		597	590	571

Note: data for the 600 schools in the study.

Table 4. Impact on Enrollment
Dependent variable: Yearly enrollment

	2011				2012				2013			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Pooled Treatment	-1.395	-1.696			-0.704	-0.944			3.336	2.798		
	(4.44)	(1.21)			(3.80)	(1.28)			(4.76)	(2.81)		
Treat. 1: HT bonus>0, T bonus=0			-6.404	-1.459			-6.745	-2.676			-1.337	2.398
			(5.98)	(1.73)			(5.83)	(1.79)			(6.10)	(2.65)
Treat. 2: HT bonus=T bonus>0			-2.463	1.329			-0.525	2.53			1.087	4.339
			(4.62)	(1.88)			(5.36)	(2.20)			(6.00)	(3.67)
Treat. 3: HT bonus>T bonus>0			4.637	-4.927***			5.161	-2.602			9.946	1.699
			(6.60)	(1.74)			(5.90)	(2.28)			(8.02)	(4.32)
Constant	93.628***	36.862	93.630***	52.33	87.568***	43.893	87.568***	64.286	83.499***	114.23	83.519***	123.003
	(4.85)	(50.70)	(4.87)	(56.91)	(4.52)	(49.58)	(4.54)	(46.30)	(4.65)	(213.58)	(4.68)	(206.33)
Treat. 1 = Treat. 2: F stat				1.506				3.401				0.264
Prob				0.243				0.09				0.616
Treat. 1 = Treat. 3: F stat				2.054				0.001				0.044
Prob				0.177				0.977				0.837
Treat. 2= Treat. 3: F stat				11.174				2.656				0.339
Prob				0.006				0.129				0.571
Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
No of Obs	598	598	598	598	593	593	593	593	583	583	583	583
R-Square	0.269	0.914	0.275	0.916	0.283	0.86	0.29	0.862	0.293	0.783	0.3	0.784

Note: Columns (1) and (3) include fixed effects by block randomization. Columns (2) and (4) include, in addition, the following baseline (2010) covariates: school's age; location (1=rural); number of teachers; number of classrooms; building condition (1=satisfactory or minor reparis); school utility index (from 0 to 4: drinking water, electricity, toilet, sewer); enrollment grades 1-5; enrollment grade 5; test takers rate grade 5; pupil to class ratio; pupil to teacher ratio and school mean test score. Treatment 1: Bonuses for head teachers only. Treatment 2: Equal bonuses for all teachers. Treatment 3: Bonuses for all teachers; higher bonus for head teacher.

Table 5. Impact on Test takers rate, grade 5
Dependent variable: Yearly test takers rate

	2011				2012				2013			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Pooled Treatment	-0.004	-0.006			0.007	0.006			0.034***	0.033***		
	(0.01)	(0.01)			(0.01)	(0.01)			(0.01)	(0.01)		
Treat. 1: HT bonus>0, T bonus=0			-0.012	-0.012			-0.012	-0.014			0.035***	0.034***
			(0.01)	(0.01)			(0.02)	(0.02)			(0.01)	(0.01)
Treat. 2: HT bonus=T bonus>0			-0.003	-0.008			0.011	0.007			0.029**	0.028*
			(0.01)	(0.01)			(0.01)	(0.01)			(0.01)	(0.02)
Treat. 3: HT bonus>T bonus>0			0.003	0.003			0.022**	0.024**			0.036***	0.036***
			(0.01)	(0.01)			(0.01)	(0.01)			(0.01)	(0.01)
Constant	0.949***	1.046*	0.949***	1.044*	0.918***	0.574	0.918***	0.619	0.936***	0.950**	0.936***	0.925**
	(0.01)	(0.54)	(0.01)	(0.56)	(0.01)	(0.63)	(0.01)	(0.56)	(0.01)	(0.39)	(0.01)	(0.38)
Treat. 1 = Treat. 2: F stat				0.074				2.214				0.242
Prob				0.79				0.163				0.632
Treat. 1 = Treat. 3: F stat				2.161				3.934				0.041
Prob				0.167				0.071				0.844
Treat. 2= Treat. 3: F stat				0.695				2.056				0.521
Prob				0.421				0.177				0.484
Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
No of Obs	597	597	597	597	591	591	591	591	575	575	575	575
R-Square	0.339	0.372	0.341	0.374	0.262	0.293	0.273	0.305	0.276	0.305	0.277	0.306

Note: Columns (1) and (3) include fixed effects by block randomization. Columns (2) and (4) include, in addition, the following baseline (2010) covariates: school's age; location (1=rural); number of teachers; number of classrooms; building condition (1=satisfactory or minor reparis); school utility index (from 0 to 4: drinking water, electricity, toilet, sewer); enrollment grades 1-5; enrollment grade 5; test takers rate grade 5; pupil to class ratio; pupil to teacher ratio and school mean test score. Treatment 1: Bonuses for head teachers only. Treatment 2: Equal bonuses for all teachers. Treatment 3: Bonuses for all teachers; higher bonus for head teacher.

Table 6. Impact on Tests
Dependent variable: Yearly standarized test, Student Level Data

	2011				2012				2013			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Pooled Treatment	-0.007	-0.015			0.179	0.196			-0.046	-0.005		
	(0.23)	(0.20)			(0.32)	(0.25)			(0.36)	(0.29)		
Treat. 1: HT bonus>0, T bonus=0			-0.125	-0.089			0.044	0.097			-0.156	-0.025
			(0.29)	(0.26)			(0.41)	(0.33)			(0.46)	(0.42)
Treat. 2: HT bonus=T bonus>0			-0.047	-0.1			0.017	0.023			-0.163	-0.2
			(0.24)	(0.21)			(0.32)	(0.29)			(0.33)	(0.26)
Treat. 3: HT bonus>T bonus>0			0.143	0.135			0.443	0.438*			0.135	0.174
			(0.21)	(0.18)			(0.36)	(0.24)			(0.38)	(0.30)
Constant	0.537	23.175**	0.536	22.866**	1.934***	-18.017	1.932***	-19.022*	2.812***	5.24	2.814***	4.103
	(0.37)	(11.62)	(0.38)	(11.14)	(0.54)	(11.53)	(0.54)	(10.92)	(0.48)	(8.43)	(0.47)	(8.01)
Treat. 1 = Treat. 2: F stat				0.011				0.07				0.323
Prob				0.919				0.796				0.58
Treat. 1 = Treat. 3: F stat				1.108				2.085				0.565
Prob				0.313				0.174				0.467
Treat. 2= Treat. 3: F stat				2.201				3.717				2.85
Prob				0.164				0.078				0.117
Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
No of Students	9030	9030	9030	9030	8085	8085	8085	8085	8211	8211	8211	8211
No of Schools	598	598	598	598	593	593	593	593	583	583	583	583
R-Square	0.169	0.203	0.17	0.204	0.224	0.268	0.226	0.27	0.235	0.289	0.237	0.291

Note: Standarization of the test with respect to control at baseline. Columns (1) and (3) include fixed effects by block randomization. Columns (2) and (4) include, in addition, the following baseline (2010) covariates: school's age; location (1=rural); number of teachers; number of classrooms; building condition (1=satisfactory or minor reparis); school utility index (from 0 to 4: drinking water, electricity, toilet, sewer); enrollment grades 1-5; enrollment grade 5; test takers rate grade 5; pupil to class ratio; pupil to teacher ratio and school mean test score. Treatment 1: Bonuses for head teachers only. Treatment 2: Equal bonuses for all teachers. Treatment 3: Bonuses for all teachers; higher bonus for head teacher.

Table 7. Impact on Index
Dependent variable: Combined Index

	2011				2012				2013			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Pooled Treatment	-0.943	-0.863			1.679	2.049			-2.947*	-2.66		
	(2.27)	(1.92)			(2.47)	(2.52)			(1.73)	(2.12)		
Treat. 1: HT bonus>0, T bonus=0			-1.841	-1.886			1.529	1.551			-2.632	-2.021
			(2.59)	(2.55)			(2.55)	(2.59)			(2.00)	(2.03)
Treat. 2: HT bonus=T bonus>0			-1.207	-1.162			2.117	2.598			-4.662*	-4.701
			(2.65)	(2.28)			(3.00)	(3.23)			(2.42)	(2.99)
Treat. 3: HT bonus>T bonus>0			0.208	0.458			1.391	2.005			-1.531	-1.298
			(2.09)	(1.59)			(2.92)	(2.78)			(2.25)	(2.74)
Constant	5.725	294.375**	5.726	294.355***	10.656***	-174.028**	10.656***	-170.362**	6.642***	-59.183	6.641***	-70.123
	(4.28)	(117.51)	(4.29)	(110.59)	(2.84)	(77.20)	(2.84)	(76.88)	(2.56)	(77.78)	(2.56)	(72.07)
Treat. 1 = Treat. 2: F stat				0.252				0.323				0.91
Prob				0.625				0.58				0.359
Treat. 1 = Treat. 3: F stat				1.208				0.05				0.06
Prob				0.293				0.827				0.811
Treat. 2= Treat. 3: F stat				0.879				0.036				1.845
Prob				0.367				0.853				0.199
Covariates	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
No of Obs	597	597	597	597	590	590	590	590	571	571	571	571
R-Square	0.352	0.429	0.353	0.431	0.251	0.28	0.252	0.281	0.275	0.305	0.277	0.307

Note: Columns (1) and (3) include fixed effects by block randomization. Columns (2) and (4) include, in addition, the following baseline (2010) covariates: school's age; location (1=rural); number of teachers; number of classrooms; building condition (1=satisfactory or minor repairs); school utility index (from 0 to 4: drinking water, electricity, toilet, sewer); enrollment grades 1-5; enrollment grade 5; test takers rate grade 5; pupil to class ratio; pupil to teacher ratio and school mean test score. Treatment 1: Bonuses for head teachers only. Treatment 2: Equal bonuses for all teachers. Treatment 3: Bonuses for all teachers; higher bonus for head teacher.

Table 8. Panel Estimation

	Dependent variable:							
	Enrollment		Test takers rate		Standarized Test scores		Index	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Pooled Treatment	0.016		0.011*		-0.013		-0.644	
	(1.39)		(0.01)		(0.17)		(0.80)	
Treat. 1: HT bonus>0, T bonus=0		-0.591		0.002		-0.115		-1.038
		(1.55)		(0.01)		(0.25)		(1.02)
Treat. 2: HT bonus=T bonus>0		2.646		0.008		-0.145		-1.198
		(1.98)		(0.01)		(0.16)		(0.77)
Treat. 3: HT bonus>T bonus>0		-1.959		0.021***		0.217		0.287
		(2.27)		(0.01)		(0.15)		(0.89)
Constant	69.971	84.566	0.858***	0.862***	7.782	7.279	25.770**	23.808**
	(82.23)	(79.03)	(0.28)	(0.24)	(5.29)	(4.84)	(11.68)	(10.19)
Treat. 1 = Treat. 2: F stat		1.933		0.566		0.033		0.039
Prob		0.19		0.466		0.86		0.846
Treat. 1 = Treat. 3: F stat		0.445		3.915		3.965		2.256
Prob		0.517		0.071		0.07		0.159
Treat. 2= Treat. 3: F stat		3.209		3.075		18.749		9.43
Prob		0.098		0.105		0.001		0.01
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No of Obs	1774	1774	1763	1763	1774	1774	1758	1758
R-Square	0.83	0.831	0.147	0.151	0.285	0.288	0.085	0.086

Note: All estimations include fixed effects by year and block randomization. Additionally, the following baseline (2010) covariates: school's age; location (1=rural); number of teachers; number of classrooms; building condition (1=satisfactory or minor reparis); school utility index (from 0 to 4: drinking water, electricity, toilet, sewer); enrollment grades 1-5; enrollment grade 5; test takers rate grade 5; pupil to class ratio; pupil to teacher ratio and school mean test score

Table 9. Heterogeneous effects

	Dependent variable: Value in year 2013							
	Enrollment		Test takers rate		Test scores		Index	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Pooled Treatment (T)	5.057		0.033**		-0.476		-5.023	
	(3.09)		(0.01)		(0.46)		(3.69)	
Above mean baseline test (A)	4.737	4.872	0.007	0.006	-0.28	-0.3	-1.305	-1.485
	(6.05)	(6.08)	(0.02)	(0.02)	(0.47)	(0.47)	(3.37)	-3.52
T X A	-3.375		0		0.512		3.777	
	(3.88)		(0.02)		(0.47)		(3.33)	
Treatment 1 (T1)		6.677*		0.034***		-0.461		-2.967
		(3.86)		(0.01)		(0.45)		(3.84)
T1 X A		-7.315		-0.001		0.269		0.792
		(4.46)		(0.01)		(0.42)		(4.95)
Treatment 2 (T2)		7.228		0.026		-0.807**		-9.197**
		(4.90)		(0.02)		(0.38)		(4.18)
T2 X A		-3.635		0.003		0.779**		8.266**
		(6.74)		(0.02)		(0.38)		(4.15)
Treatment 3 (T3)		1.669		0.037**		-0.199		-3.181
		(3.86)		(0.02)		(0.60)		(4.06)
T3 X A		0.272		-0.002		0.549		2.761
		(4.79)		(0.02)		(0.77)		(3.97)
Constant	-22.15	-22.809	0.839***	0.837***	0.125	0.025	-3.324	-4.924
	(18.36)	(17.24)	(0.08)	(0.08)	(2.05)	(2.11)	(20.54)	(20.82)
(T1 X A) = (T2 X A): F stat		0.388		0.039		2.469		1.933
Prob		0.545		0.847		0.142		0.19
(T1 X A) = (T3 X A): F stat		1.262		0.019		0.182		0.125
Prob		0.283		0.893		0.677		0.73
(T2 X A) = (T3 X A): F stat		0.347		0.154		0.171		2.426
Prob		0.567		0.702		0.686		0.145
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No of Obs	583	583	575	575	583	583	571	571
R-Square	0.705	0.706	0.053	0.055	0.095	0.104	0.042	0.049

Note: All estimations include fixed effects by block randomization and the following baseline (2010) covariates: school's age; location (1=rural); number of teachers; number of classrooms; building condition (1=satisfactory or minor reparis); school utility index (from 0 to 4: drinking water, electricity, toilet, sewer); enrollment grades 1-5; enrollment grade 5; test takers rate grade 5; pupil to class ratio; pupil to teacher ratio and school mean test score. Treatment 1: Bonuses for head teachers only. Treatment 2: Equal bonuses for all teachers. Treatment 3: Bonuses for all teachers; higher bonus for head teacher

Table 10. Difference in inputs, 2013

	Treatment (T)	Control (C)	T-C	Treatment 1	T1-C	Treatment 2	T2-C	Treatment 3	T3-C
	(1)	(2)	(3)	(1)	(3)	(1)	(3)	(1)	(3)
Index of Utilities	3.66	3.72	-0.066	3.68	-0.007	3.66	-0.101	3.63	-0.091
	(0.60)	(0.52)	(0.05)	(0.55)	(0.05)	(0.66)	(0.06)	(0.59)	(0.06)
Index of Infrastructure	2.7	2.69	0.01	2.73	0.051	2.7	0.005	2.68	-0.025
	(0.63)	(0.64)	(0.05)	(0.60)	(0.07)	(0.65)	(0.08)	(0.65)	(0.04)
Playground (1=yes)	1.39	1.39	0	1.47	0.068	1.33	-0.042	1.37	-0.025
	(0.49)	(0.49)	(0.04)	(0.50)	(0.05)	(0.47)	(0.04)	(0.48)	(0.04)
Library (1=yes)	1.57	1.57	-0.007	1.57	-0.013	1.6	0.017	1.55	-0.028
	(0.50)	(0.50)	(0.04)	(0.50)	(0.04)	(0.49)	(0.06)	(0.50)	(0.06)
No of Toilets	2.37	2.27	0.081	2.14	-0.107	2.57	0.293	2.41	0.062
	(2.20)	(0.98)	(0.15)	(0.92)	(0.16)	(3.50)	(0.33)	(1.17)	(0.16)
Pupil-toilet ration	69.77	66.76	3.372	68.43	5.057	67.87	3.829	73.08	1.167
	(49.41)	(39.68)	(3.52)	(52.88)	(5.07)	(44.63)	(5.46)	(50.51)	(5.62)
No of Classrooms	3.48	3.44	0.008	3.45	0.059	3.54	0.058	3.44	-0.092
	(1.94)	(1.79)	(0.12)	(1.66)	(0.14)	(1.89)	(0.18)	(2.22)	(0.15)
Pupil-classroom ration	47.16	47.41	0.066	42.41	-3.318*	44.3	-0.498	54.67	3.968
	(30.72)	(36.32)	(1.24)	(25.87)	(1.94)	(26.89)	(2.79)	(36.86)	(2.59)
Teachers	3.27	3.29	-0.023	3.22	-0.058	2.97	-0.165	3.59	0.146
	(2.12)	(1.81)	(0.22)	(1.79)	(0.25)	(1.48)	(0.24)	(2.79)	(0.20)
Pupil-teacher ration	48.1	48.76	-0.215	44.57	-0.427	50.63	1.948	49.02	-2.058
	(25.90)	(29.65)	(2.36)	(21.52)	(2.75)	(29.11)	(3.85)	(26.23)	(1.79)
Grade 5 enrollment, boys	10.17	9.78	-0.569	9.34	-0.501	10.05	-0.509	11.1	-0.696
	(11.06)	(10.72)	(0.76)	(9.24)	(0.69)	(10.65)	(1.06)	(12.95)	(0.96)
Grade 5 enrollment, girls	3.96	4.82	-0.153	3.68	-0.538*	3.85	0.08	4.34	0.002
	(6.76)	(6.61)	(0.56)	(6.01)	(0.32)	(5.30)	(0.70)	(8.55)	(0.92)
Grade 5 enrollment	14.13	14.6	-0.722	13.03	-1.039	13.91	-0.429	15.44	-0.694
	(11.15)	(9.95)	(0.71)	(8.55)	(0.63)	(9.99)	(0.90)	(14.08)	(0.94)

Note: Column (3) is the coefficient of the variable treatment agains each of the variables in the table. All estimations include fixed effects by block randomization, in addition to the following baseline (2010) covariates: school's age; location (1=rural); number of teachers; number of classrooms; building condition (1=satisfactory or minor reparis); school utility index (from 0 to 4: drinking water, electricity, toilet, sewer); enrollment grades 1-5; enrollment grade 5; test takers rate grade 5; pupil to class ratio; pupil to teacher ratio and school mean test score

Table 11. Impact on Enrollment by Grade

Dependent variable: Enrollment 1-4 grades and only 5 grade

	Enrollment, Grades 1 to 4						Enrollment, Grade 5					
	2011		2012		2013		2011		2012		2013	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Pooled Treatment	-1.839*		-0.631		3.588		0.143		-0.313		-0.79	
	(1.01)		(1.42)		(2.40)		(0.58)		(0.50)		(0.65)	
Treat. 1: HT bonus>0, T bonus=0		-2.516		-1.414		3.425		1.058**		-1.262**		-1.027*
		(1.64)		(1.94)		(2.44)		(0.53)		(0.60)		(0.55)
Treat. 2: HT bonus=T bonus>0		0.917		1.871		4.955		0.412		0.659		-0.616
		(1.86)		(2.22)		(3.07)		(0.85)		(0.55)		(0.85)
Treat. 3: HT bonus>T bonus>0		-3.889**		-2.292		2.431		-1.038*		-0.31		-0.732
		(1.56)		(2.27)		(3.71)		(0.58)		(0.60)		(0.92)
Constant	65.815*	81.259*	83.729***	97.527***	149.176	156.579	-28.953	-28.929	-39.835	-33.241	-34.946*	-33.576*
	(35.58)	(43.22)	(28.66)	(28.47)	(202.28)	(197.20)	(22.91)	(25.13)	(25.52)	(21.85)	(19.51)	(17.49)
Treat. 1 = Treat. 2: F stat		2.315		1.416		0.209		1.216		13.356		0.382
Prob		0.154		0.257		0.656		0.292		0.003		0.548
Treat. 1 = Treat. 3: F stat		0.306		0.125		0.128		24.235		2.977		0.11
Prob		0.59		0.729		0.726		0		0.11		0.746
Treat. 2= Treat. 3: F stat		4.837		2.042		0.393		4.499		2.508		0.026
Prob		0.048		0.179		0.543		0.055		0.139		0.876
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No of Obs	598	598	593	593	583	583	598	598	593	593	583	583
R-Square	0.896	0.897	0.835	0.837	0.75	0.751	0.799	0.804	0.758	0.763	0.755	0.755

Note: All regressions include fixed effects by block randomization and the following baseline (2010) covariates: school's age; location (1=rural); number of teachers; number of classrooms; building condition (1=satisfactory or minor reparis); school utility index (from 0 to 4: drinking water, electricity, toilet, sewer); enrollment grades 1-5; enrollment grade 5; test takers rate grade 5; pupil to class ratio; pupil to teacher ratio and school mean test score. Treatment 1: Bonuses for head teachers only. Treatment 2: Equal bonuses for all teachers. Treatment 3: Bonuses for all teachers; higher bonus for head teacher