Policy Research Working Paper 7307

Teacher Performance Pay

Experimental Evidence from Pakistan

Felipe Barrera-Osorio
Dhushyanth Raju

WORLD BANK GROUP
South Asia Region
Office of the Chief Economist
June 2015

Abstract

This paper presents evidence from the first three years of a randomized controlled trial of a government-administered pilot teacher performance pay program in Punjab, Pakistan. The program offers yearly cash bonuses to teachers in a sample of public primary schools with the lowest mean student exam scores in the province. Bonuses are linked to three school-level indicators: the gain in student exam scores, the gain in school enrollment, and the level of student exam participation. Bonus receipt and size are also randomly assigned across schools according to whether or not the teacher is the school's head. On average, the program increases school enrollment by 4.1 percent and student exam participation rates by 3.4 percentage points, both in the third year. The analysis does not find that the program increases student exam scores in any year. Mean impacts are similar across program variants. The positive mean impact on school enrollment is mainly seen in urban schools and the positive mean impact on student exam participation rates is only seen in rural schools.

This paper is a product of the Office of the Chief Economist, South Asia Region. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The authors may be contacted at felipe_barrera-osorio@gse.harvard.edu and draju2@worldbank.org.

Produced by the Research Support Team

Teacher performance pay:
Experimental evidence from Pakistan

Felipe Barrera-Osorio
Harvard Graduate School of Education

Dhushyanth Raju
World Bank

JEL classification codes: C93, I21, I28, M52

Keywords: performance pay, incentive pay, public schools, teachers, field experiment, primary education, Punjab, Pakistan

Barrera-Osorio: felipe_barrera-osorio@gse.harvard.edu; Raju: draju2@worldbank.org. 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; Alejandro Ganimian for research assistance; and Robert Chase, Jishnu Das, Gladys Lopez-Acevedo, Richard Murnane, Quynh Nguyen, Laura Trucco, and Katja Vinha for comments and suggestions. We are also grateful for comments and suggestions by seminar participants at Colgate, Harvard, and Stanford Universities. The Government of Punjab's Education Sector Reform Program, including the teacher performance pay program under 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), and the World Bank. Financial support for the study from the Australian Department of Foreign Affairs and Trade (DFAT), the Weiss Family Program Fund for Research in Development Economics at Harvard University, and the World Bank is gratefully acknowledged.

I. Introduction

Teacher effectiveness—a key determinant of school quality—is perceived to be poor in many low-income countries (Hanushek and Rivkin 2006, Glewwe and Kremer 2006). International development organizations increasingly promote performance (or incentive) pay to governments as an option to raise teacher effort and, thereby, teacher effectiveness.

Some recent teacher performance pay ventures in low-income countries have been rigorously evaluated. These evaluated interventions serve as valuable proof-of-concept demonstrations but are "special" in character. Although largely administered in public schools, the evaluated interventions were often designed and administered by interested and capable NGOs, with significant guidance and oversight from outside researchers. In addition, data for administering the interventions were gathered by the researchers and partnering organizations. (See, for example, Glewwe, Ilias, and Kremer 2010; Duflo, Hanna, and Ryan 2012; Muralidharan and Sundararaman 2011.) Hence, documented impacts may not be generalizable to other, different arrangements, such as where the government has overall primary responsibility for implementation (Bold, Kimenyi, Mwabu, Ng'ang'a, and Sandefur 2013).

We evaluate the *Improvers Bonus Program for Government School Teachers*—a pilot public school teacher performance pay program conceived, designed, and managed by the provincial government of Punjab, Pakistan. The government is responsible for all aspects of the program and uses its own administrative data to assess teacher performance. While multiple international donor agencies provide financial and technical support to the government's education reform agenda which includes the teacher performance pay program, the degree of external influence on the program does not differ from other, supported but unevaluated programs in the government's reform agenda. Presumably, there is also less such influence than in the demonstrations we noted above.

Our study contributes evidence on the effectiveness of teacher performance pay programs, which fall under the class of incentive-based, supply-side education interventions.[1]

[1] Available rigorous evidence on the impacts of teacher performance pay is inconclusive: in some studies, there is consistent evidence of positive impacts (see, for example, Lavy 2002; Duflo, Hanna, and Ryan 2012; Muralidharan and Sundararaman 2011; and Muralidharan 2012); in others, the evidence is mixed or conditional (see, for example, Glewwe et al 2010; Behrman, Parker, Todd, and Wolpin 2012; Sojourner, Mykerezi, and West 2014; and Fryer, Levitt, List, and Sadoff 2012; and yet in others, there is no evidence of positive impacts (see, for example, Fryer 2013; Springer et al 2011; and Goodman and Turner 2013).

More generally, our study contributes evidence on the effectiveness of public service delivery reform in a low-income setting where government capacity and accountability are considered to be poor.

Initiated in mid 2010, the program offers yearly cash bonuses linked to school performance to public primary school teachers. At program registration, teachers are offered some basic school management, teaching, and exam preparation tips. The program is however designed to incentivize incumbent teachers to raise school performance by increasing their effort rather than by directly increasing their skill.[2]

The bonuses are offered to teachers on top of their standard salaries, and are set as a linear function of a composite score of school performance. The composite score is obtained from a weighted sum of three indicators:

- the gain in the school's enrollment in first through fifth grade,
- the gain in the school's mean score on Punjab's standardized fifth-grade examination, and
- the participation rate of the school's fifth-grade students in the examination.

The first two indicators are consistent with the government's main education goals: to raise school participation and student academic achievement. Participation and achievement are both acutely low in Punjab. The exam participation rate is included as a deterrent against potential exampool selection by program schools. The data for the indicators come from the Punjab Examination Commission (PEC) fifth-grade exam, which is a yearly standardized exam, and the Annual School Census (ASC) survey, which is a yearly field survey of all public schools. Both data instruments are administered by the government.

Increasing the effort of head teachers may be an important way to raise school performance, and bonuses to head teachers may help elicit greater effort from them. Studies find

In spite of these mixed results, a few general patterns can be drawn from the collective evidence. Programs tend to have targeted schools with low baseline levels of student academic achievement. Program impacts are usually modest if present, but are large in a few cases. Programs in low-income countries have a higher success rate than those in high-income countries. Discussions of program implementation, in particular the fidelity and quality of implementation, are limited, if any. There is little or no rigorous evidence on, where relevant, the conditions and pathways behind intended (and unintended) impacts or the factors behind the lack of intended impacts.

[2] Growing credible evidence suggests the prevalence and severity of low teacher effort, of which two basic measures are whether the teacher is present in school and whether the teacher is on-task. Teacher absentee rates are found to be high in many low- and middle-income countries, and appear to result from many teachers being occasionally absent rather than a few teachers being frequently absent. When teachers are present in school, a large share of them are found to be off task (Glewwe et al 2010; Chaudhury, Hammer, Kremer, Muralidharan, and Rogers 2006).

that the performance of head teachers—as school leaders—is an important determinant of school performance (see, for example, Waters, Marzano, and McNulty 2003; Robinson, Claire, Lloyd, and Rowe 2008; Mulford 2003).[3] Bonuses to head teachers may also be a more efficient way to raise school performance than providing bonuses to all teachers. To test whether leveraging head teachers is either sufficient or augments program impacts, the receipt and size of the bonus for a given level of school performance are randomly assigned across schools according to whether or not the teacher is the school's head teacher.

In the first treatment variant, head teachers are eligible for "level-1" bonuses, while other teachers are not eligible for any bonuses. In the second treatment variant, all teachers (whether head teacher or not) are eligible for level-1 bonuses. In the third and last treatment variant, head teachers are eligible for higher "level-2" bonuses, while other teachers are eligible for level-1 bonuses. Level-2 bonuses are fixed to be twice as large as level-1 bonuses for a given level of school performance. The treatment variants and untreated status are randomly assigned across the evaluation sample of 600 public primary schools with essentially the lowest school-level mean 2010 PEC exam scores, located in three districts (out of Punjab's 36 districts) with the lowest district-level mean 2010 PEC exam scores.

Using the PEC exam and ASC survey data, we evaluate program impacts on school enrollment, exam participation rates, and student exam scores in the first three years of program implementation. Evidence of significant mean impacts across the three indicators is mixed. Significant positive mean impacts are concentrated in the third year. The program increases school enrollment by, on average, 4.1 percent in the third year. Treatment variants have similarly-sized third-year mean impacts, but only the impact for the treatment variant that offers bonuses solely to head teachers is significant. The program increases exam participation rates by, on average, 3.4 percentage points (ppts) in the third year. Given a baseline mean exam participation rate of 97 percent, the mean impact maximizes the rate. Treatment variants have similarly-sized third-year mean impacts, and all of them are significant. We do not find that any of the treatment variants, and hence the program, increases student exam scores in any year.

[3] Factors that may drive a positive relationship between head teachers and school performance include "instructional leadership" (the ability to create an environment conducive for student learning), "transformational leadership" (the ability to motivate and empower teachers), and "managerial skills" (the ability to manage the school as a typical bureaucracy) (Boyd at al 2011, Ingersoll 2011, Grissom and Loeb 2011). High-performing school systems are often found to nurture, retain, and promote leaders in their teaching workforces as a core strategy for student learning (see, for example, Darling-Hammond and Rothman 2011 and World Bank 2011).

We test for two forms of strategic manipulation that could explain the mean impacts on school enrollment and exam participation rates. First, program schools may choose to reduce fifth-grade enrollment by limiting promotion to fifth grade or shedding fifth-grade students in order to raise their exam participation rates. We do not find consistent evidence across treatment variants that the program decreases fifth-grade enrollment. Second, program schools may choose to inflate reported enrollment. By constructing pseudo-panel cohorts using grade-specific enrollment information from the yearly ASC surveys, we test whether the program flattens the gradient at which enrollment declines with grade (or, more extremely, inverts it), which would be consistent with enrollment inflation. We do not find evidence suggesting this.

Post-hoc subgroup analysis reveals that the third-year program mean impacts on school enrollment and exam participation rates vary between urban and rural schools. The program mean impact on enrollment in the total sample is mainly due to the urban subsample (which represents 8 percent of the total sample). The program increases enrollment in urban schools in the third year by, on average, 23 percent (the third-year mean impact in rural schools is 2 percent and insignificant). Looking at enrollment separately by grade, we find significant positive mean impacts of similar size in percent terms across first through fourth grade, indicating higher intake as well as higher retention. The program mean impact on exam participation rates in the total sample is fully due to the rural subsample. The program increases exam participation rates in rural schools in the third year by, on average, 3.6 ppts (the third-year mean impact in urban schools is –1.4 ppts and insignificant). Program mean impacts on student exam scores differ between the urban and rural subsamples in the first year but the differences in later years are not significant.

Given that the program increases exam participation rates, we test whether the absence of significant positive mean impacts on student exam scores is due to negative selection of students into examtaking. We do not find evidence suggesting this. Apart from this test, we are left to speculate on factors that could have blocked impacts. We conclude the paper with this speculative discussion but, to summarize, we discount that the absence of positive mean impacts is due to program implementation failures. We also discount that it is due to bonus awards being too low, to the bonus formula being too complex for teachers to comprehend, or (as relevant for the treatment variant) to the group-based nature of the bonuses inducing significant freeriding.

We posit three hypotheses for the absence of positive mean impacts on student exam scores. First, nonschool or school factors outside the control of teachers may make raising student exam scores difficult even with greater teacher effort. Targeting schools with the lowest mean student exam scores in the province places the program in circumstances that may be especially handicapping. Second, teachers may lack the knowhow to raise student exam scores. Third, teachers may optimally choose to direct their effort at those incentivized margins that maximize payoffs net of effort costs. In this case, it appears that the net payoff-maximizing margin is school enrollment in urban areas, exam participation rates in rural areas, and not student exam scores anywhere.

The rest of the paper is organized as follows. Section II describes the context and the program. Section III discusses the data, sample, and empirical strategy. Section IV reports the empirical results. In concluding, Section V summarizes the main results and discusses potential explanations for the absence of program impacts on student exam scores.

II. Context and program

A. Context

Punjab, the site of the teacher performance pay program, is the largest of Pakistan's four provinces, accounting for three-fifths of the country's population and income. In 2012/13, the estimated net enrollment rate at the primary school level was 66 percent (2012–13 Pakistan Social and Living Standards Measurement [PSLM] survey report, Pakistan Bureau of Statistics, Government of Pakistan). In 2007, third-grade students in a school sample in rural Punjab scored, on average, 31 percent in English, 27 percent in Urdu, and 34 percent in mathematics on independently administered, competency-based tests. These results are significantly below official grade-level standards.[4]

The public school system is the main provider of education in Punjab. In 2012–13, 61 percent of primary school students were in public schools (2012–13 PSLM report, Pakistan Bureau of Statistics, Government of Pakistan). The system is large, composed of 51,504 functional schools. Seventy percent of these schools are primary schools (preschool to fifth grade) with 105,300 assigned teachers and 2.7 million children enrolled in first through fifth

[4] Statistics provided by Jishnu Das.

grade (October 2013 Annual School Census, School Education Department, Government of Punjab).

Available evidence indicates that teacher performance in public schools is poor. Using field survey data from rural Punjab, Andrabi, Das, Khwaja, Vishwanath, and Zajonc (2009) cast light on the performance of public school teachers, particularly when contrasted with teachers in low-cost, for-profit private schools operating in the same villages. They find that public school teachers tend to have higher levels of academic qualifications, teaching experience, and professional training than do private school teachers. Public school teacher salaries are, on average, three to four times higher than private school teacher salaries after accounting for differences in teacher credentials and other characteristics.

Despite the higher levels of credentials among public school teachers, Andrabi et al find that the levels of teacher presence and student academic achievement are lower in public schools than in private schools. It also appears that public school teacher salaries do not vary with teacher competency (measured by teacher test scores) or student competency (measured by student test scores). Higher public school teacher salaries also seem to be associated with lower teacher effort, as measured by teacher presence at school. This pattern is driven mainly by the relatively lower presence of more senior (and, therefore, better paid) teachers in public schools. In contrast, among private school teachers, higher salaries are associated with higher teacher presence, higher teacher test scores, and higher student test scores.

B. Program

The teacher performance pay program was designed and approved by the Punjab government in fiscal year (FY) 2009–10 for implementation on a pilot basis starting in FY2010–11 (corresponding to the 2010–11 school year).[5] The main expressed aim of the program is to raise school performance by encouraging greater teacher effort.

Measuring school performance

The government selected performance indicators which could be measured regularly, throughout the province, in a standardized way across schools, and for which the underlying data

[5] The government's fiscal year begins on July 1 and ends on June 30. The public school system's school year begins on April 1 and ends on March 15.

were considered least susceptible to manipulation by schools. The government possessed two data sources that fulfilled these criteria: the Annual School Census (ASC) surveys and the Punjab Examination Commission (PEC) exams.

Initiated in 2003, the ASC is an annual field survey of public schools which captures basic school and teacher information, including student enrollment by grade and gender, with a reference date of October 31. Initiated in 2006, the PEC exam is an annual, standardized academic test which is mandatory for public school students in fifth and eighth grades. The test uses selected and constructed response questions based on standard learning objectives of Pakistan's official curriculum. Core subjects are English, Urdu, Islamic studies, mathematics, science, and social studies, all assessed through pencil-and-paper tests.

The provincial education department is responsible for managing all stages of the ASC survey and PEC exam activities, and is solely responsible for the design of the data collection instruments. At the field level, Monitoring and Evaluation Assistants (MEAs) visit assigned public schools in November to collect ASC survey data. The MEAs are mostly retired members of the armed services and report directly to the provincial education department. The fifth-grade PEC exam is administered in January and/or February by district education department staff and enlisted public middle and high school teachers. Public middle and high schools serve as exam centers. Filled-in multiple choice scoring sheets are dispatched to PEC headquarters to be read by optical mark readers. Answers to open-ended questions are scored by recruited public middle and high school teachers at grading centers in the district; the scores are transmitted to PEC headquarters.

The provincial education department selected three indicators for measuring school performance using the ASC survey and PEC exam data:

- the gain in mean exam scores (*GTS*),
- the gain in enrollment in first through fifth grade (*GSE*), and
- the exam participation rate (*TPR*).

GTS is defined as the year-over-year gain in the school's percent mean score in the fifth-grade PEC exam. The indicator is measured in percentage point terms. The exam scores used by the government for the indicator were raw scores—the number of exam questions answered correctly—for the multiple choice question modules in the core subjects.[6] *GSE* is defined as the year-over-year gain in the school's enrollment in first through fifth grade captured in the ASC survey. The indicator is expressed in percent terms. *TPR* is defined as the number of students who took the fifth-grade PEC exam from a school (recorded by PEC) divided by the number of enrolled students in fifth grade in the school (captured in the ASC survey). The indicator is expressed in percent terms. The values for each of the three indicators are bounded by zero percent and 100 percent.

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

$$CS = 0.65GTS + 0.25GSE + 0.15TPR. \, (1)$$

The assigned weights reflect the provincial government's relative valuation of the three indicators in assessing school performance.

Targeting

The program is targeted at public primary schools with among the lowest mean student exam scores in the province following a three-step process. First, the three districts with the lowest district-level mean scores in the 2010 fifth-grade PEC exam in Punjab's 36 districts were selected. These districts were Attock, Mandi Bahuaddin, and Rahimyar Khan. Second, within these districts, schools at the primary level that were functional according to the October 2009 ASC survey and had at least ten students who took the 2010 fifth-grade PEC exam were identified. Third, 1,962 such schools were ranked from highest to lowest by their 2010 school-level mean fifth-grade PEC exam scores; the lowest ranked 600 schools were selected for the evaluation sample. The range of mean scores for the sample corresponds to that of 12 percent of all public primary schools in the province.

Treatment variants

In principle, the teacher assigned to be (acting) head teacher is authorized by the district education department to manage the school and its teachers. Head teachers can manage teachers by supervising, rewarding, sanctioning, and supporting teachers. If head teachers are a

[6] Raw scores are used instead of the scores scaled by PEC because of concerns related to the conversion formula that is applied. The scoring of answers to multiple-choice questions is considered more reliable than the scoring of answers to open-ended questions.

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 the school.

These views shaped the design of three alternative treatment variants to test in parallel. Bonus eligibility and size were varied according to whether or not the teacher is the school's head teacher.[7] The treatment variants are:

- ***HT only***: only head teachers were eligible for level-1 bonuses.
- ***all T***: both head teachers and teachers were eligible for level-1 bonuses.
- ***HT+***: head teachers were eligible for 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. Treatment variants and untreated status are randomly assigned across the evaluation sample by first blocking schools into quads based on 2010 school-level mean fifth-grade PEC exam scores.[8]

Bonus formula

Table 1 presents the bonus formula, by treatment variant and teacher type (head teacher or other teacher). The slope parameters in the bonus formula were calibrated so that teachers in the average school at baseline in the evaluation sample face high-powered incentives to perform. Even if the average school does not increase enrollment and mean student exam scores but maintains its baseline exam participation rate of 97 percent, the minimum teacher bonus payout would be 14,550 Pakistani rupees (US$169).[9] If the average school maxes out the composite score in the first year, the minimum teacher bonus payout would be 86,200 Pakistani rupees (US$1,002).[10] These two minimum payouts represent 10 and 57 percent respectively of the

[7] A teacher appointed as a 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 is considered eligible for the (head) teacher bonus even if the (head) teacher is then transferred at any point during the school year to any another school.

[8] The trial is unmasked: the experimental groups that schools are assigned to are known by the provincial and district education departments, program implementing partners, supporting international donor agencies, and us (the researchers). Program schools are directly made aware of their assigned treatment variant.

[9] We use the December 2010 exchange rate of 86 Pakistani rupees per US dollar.

[10] Maxing out the composite score in the first year would imply raising the exam participation rate from 97 to 100 percent, doubling enrollment from 100 students, and raising the school mean exam score from 23 to 100 percent.

yearly basic salary of an average teacher at baseline in the evaluation sample.[11] The percentages would be two-thirds to one-half their size if we instead use take-home pay (basic pay plus cash benefits) as the base.[12]

Implementation

In each of the three program years (FY2010–11, FY2011–12, and FY2012–13), staff from the provincial education department and partnering organizations conducted a field visit in the middle of the school year. During this visit, program school teachers were invited to attend meetings, separately by treatment variant-specific, proximity-based school groupings.

The main aim of the meetings was to acquaint teachers with their relevant treatment variant and register them for the program. To register, the teacher filled in a form providing employment details and bank account information. The meetings were also used to facilitate experience sharing between teachers and to offer some tips for basic school and class management, teaching, and exam preparation.

At the end of the first and second program years (two months after the end of the school year), school performance score cards were sent from the provincial education department to the district education departments. The score cards reported the school's *GTS*, *GSE*, *TPR*, and *CS* values for the school year that just ended. The district departments were instructed to dispatch 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.

At the end of the third program year, the provincial education department and partnering organizations held meetings with program school head teachers and directly handed them score cards (with a new format) to take back and display at their schools. This round of meetings was also used to facilitate experience sharing between teachers and to reiterate offered tips.

In each program year, school performance scores and bonus awards were calculated by the provincial education department, with technical support and quality control by partnering organizations. The bonus awards, aggregated up to the district level, were communicated to the

[11] The average teacher is at pay grade 9 and has 17 years of government service. The 2011 pay schedule for the Government of Pakistan indicates that an official at this pay grade and with this service length would earn a yearly basic salary of 151,920 Pakistan rupees (US$1,767).

[12] Government school teachers receive multiple monthly cash benefits (or "allowances") for, among other things, transportation, medical, and housing expenses, and cost-of-living relief.

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. At the same time, the provincial education department sent the district education department a list of program schools and teachers with the bonus award and the personal bank account information of program school teachers.

Provincial government and third-party validations find that program implementation has been generally satisfactory since the first year of implementation.

Comparison with advocated best practice

The program's design takes account of key considerations and issues discussed in the literatures on optimal incentive contracting and incentive pay (see, for example, Neal 2011 for a review of performance pay in education) in at least four ways. First, the program's piece-rate bonus structure arguably incentivizes efficient effort from teachers who may differ in their level of skill. Second, the program links bonuses to the exam participation rate to protect against teachers' selecting academically-stronger students to register and take the exam. Indeed, high-stakes accountability and performance pay programs are often documented to result in gaming (see, for example, Figlio and Winicki 2005 and Jacob and Levitt 2003). Third, in two of the three treatment variants, the program offers school-level group bonuses which can help prevent perceptions of unfairness and unproductive competition between teachers in the same school, although group bonuses can create a freerider problem (Holmstrom 1982; Cohn, Fehr, Herrmann, and Schneider 2013). Fourth, the program attempts to avoid a potential multitasking problem (Baker 1992, Holmstrom and Milgrom 1991). The program links bonuses to changes in school-level mean student exam scores to incentivize program school teachers to exert effort across all students and perhaps greater effort on relatively weaker students with more room to improve as well as incentivizes all program schools (irrespective of their baseline mean exam scores) to raise their performance (Barlevy and Neal 2012).[13] The program also links bonuses to student exam scores from the five core subjects in the country's curriculum to avoid teachers' shifting their effort to a narrow set of incentivized subjects (Jacob 2005, Koretz 2002).

[13] This structure contrasts with one that links bonuses to a minimum test score which can distort teacher effort towards students near the minimum test score or links bonuses to a minimum test pass rate which can discourage teacher effort in schools that are relatively farther away from the minimum test pass rate.

Nevertheless, the program's design is imperfect in at least three ways. These imperfections, either singly or jointly, are not unique to this program. First, the program does not use teacher value-added measures which adjust gains in student exam scores for existing differences in relevant student and school characteristics (Koedel, Mihaly, and Rockoff *forthcoming*).[14] Second, while the ASC survey and PEC exam data are used for both monitoring and accountability, the program makes these data more high stakes, increasing the risk that gaming and data manipulation undermine data reliability (even if this risk is limited to the relatively small sample of program schools). Third, gains in student exam scores may be due to teachers coaching students in examtaking strategies rather than teachers engaging in actions that produce actual student learning.[15]

III. Data, sample, and empirical strategy

A. Data and sample

Four annual rounds of the ASC survey and PEC exam data were linked together using the government's unique school identification codes (the Education Management Information System [EMIS] code). The October 2009 ASC survey and the January–February 2010 PEC exam rounds serve as our baseline data. (The program came into effect in July 2010.) The October 2012 ASC survey and January–February 2013 PEC exam rounds serve as our data for the third program year.

Evaluation sample schools were linked largely without issue across the various databases. We could not link a few schools because they were merged into other schools as part of Punjab's school merger program.[16] As a result, 583 schools (or 97 percent of the original evaluation sample) remain in the evaluation sample in the third program year. The shares of schools "attrited out" from the original evaluation sample do not differ between treated and untreated samples (2.7 percent versus 3.3 percent, respectively).

[14] The use of value-added measures for teacher performance management is however contentious in education research and practice communities (see, for example, Goldhaber 2010).

[15] Neal (2011) argues that exam preparation can be defensible if performance pay shifts teachers from ineffective or lack of teaching—which presumably applies in our case (see Chaudhury et al 2006 for evidence from low- and middle-income countries on teacher absentee and off-task rates)—to exam preparation than from effective teaching to exam preparation.

[16] Acquired schools lose their status as separate schools and are assigned the same EMIS codes as their respective acquiring schools.

Column 1 in Table 2 reports baseline mean outcomes and characteristics for schools in the original evaluation sample. Eight percent of schools are urban. Eighty-six percent of schools have buildings in satisfactory condition. On average, schools have 3.1 out of 4 basic amenities (drinking water, electrical connection, toilets, and sewer connection). On average, schools have 3 teachers, with an average of 17 years of experience in government service and 10 years of service in their currently-assigned schools. Fifty-one percent of teachers have at most a high school diploma, 35 percent have a bachelor's degree, and 14 percent have at least a master's degree. On average, schools have 100 enrolled students in first through fifth grade, student-teacher and student-classroom ratios higher than 50:1, and an exam participation rate of 97 percent. School mean exam scores are standardized using the mean and standard deviation (sd) for school mean exam scores for untreated schools.

We compare schools in the original evaluation sample to all public primary schools in the province. Column 2 in Table 2 reports baseline mean outcomes and characteristics for 42,438 functional primary schools in the province (for ease, referred to as "all schools") and Column 3 reports differences in baseline means between evaluation sample schools and all schools.

On average, evaluation sample schools were established earlier (+3 years). Evaluation sample schools are more likely to have buildings in satisfactory condition (+3 ppts); they have, on average, more classrooms (+0.4) and basic amenities (+0.3). Evaluation sample schools are just as likely to be urban as all schools, and less likely to be de-facto girls only (–15 ppts) and more likely to be de-facto coeducational or boys only. Teachers in evaluation sample schools are more likely to hold a bachelor's degree than teachers in all schools (+11 ppts), and less likely to have completed high school (–9 ppts) or hold at least a master's degree (–3 ppts).

Evaluation sample schools are, on average, larger (+22 students in first through fifth grade, +7 students in fifth grade), and more crowded (+8.3 students per classroom, +11.9 students per teacher). Finally, evaluation sample schools have, on average, higher exam participation rates (+4.4 ppts) and much lower exam scores (–4.2 sds). The last finding results from the way the evaluation sample is set.

B. *Empirical strategy*

The evaluation is based on a randomized block design. The 600 evaluation sample schools were randomly assigned into the four treatment statuses (*HT only*, *all T*, and *HT+*, and untreated) of 150 schools each, after blocking by baseline school mean exam scores. This design ensures that any differences in mean outcomes between treatment statuses are attributable to differences in treatment statuses and not due to differences in baseline mean outcomes and characteristics between treatment statuses.

We obtain mean impacts (specifically, intent-to-treat effects) by estimating via ordinary least squares an outcome regression of the form

$$Y_s = \beta_0 + \sum_{i=1}^{3} \beta_i T_{i,s} + X_{s,0}\gamma + \mu_d + \varepsilon_s, \qquad (2)$$

where Y_s denotes the outcome in school *s*, $T_{i,s}$, $i = 1,2,3,$ treatment variant indicators (corresponding to treatment variants *HT only*, *all T* and *HT+*), $X_{s,0}$ baseline characteristics, and μ_d district indicators. We estimate alternative variants of (2). We estimate regressions where we pool the three treatment variants (T_s equals one if the school is treated, and zero if untreated), as well as regressions where each treatment variant is included additively. We estimate separate regressions for each program year. Regressions for enrollment and exam participation rates are at the school level and estimated standard errors are clustered at the tehsil (sub-district) level; regressions for exam scores are at the student level and estimated standard errors are clustered at the school level.[17]

Table 2 reports baseline mean outcomes and characteristics for the original evaluation sample, separately by pooled treatment status (treated schools in Column 4 and untreated schools in Column 5), and differences in means between treated and untreated schools (Column 6). With the exception of the de-facto gender type of the school and years of government service of the teacher, results of tests of pairwise differences in baseline means between treated and untreated schools are insignificant. Results of tests of joint differences in baseline means between treated and untreated schools for three groups of variables, namely school-, teacher-, and enrollment and exam score-related (Panels A, B, and C), are also insignificant.[18] In sum, the results confirm that treated and untreated school samples are balanced at baseline.

[17] The country has four tiers of government administration: province, district, tehsil, and union council. The lowest level of public school system administration is the tehsil.

[18] We also check balance between treated and untreated schools in baseline mean outcomes and characteristics for schools that remain in the evaluation sample in the third program year. This test is important because, as we will show,

IV. Results

A. Program impacts

Table 3 reports estimated mean impacts on school enrollment (Columns 1–3), exam participation rates (Columns 4–6), and student exam scores (Columns 7–9) in each of the three program years.

First, with respect to school enrollment, mean impacts of the treatment variants either turn from negative to positive or are positive and grow in size over the three years. Third-year mean impacts are 3.9 students for *HT only*, 2.5 students for *all T*, and 5.8 students for *HT+*, and are not statistically different from each other. However, only the mean impact for *HT only* is significant. The pooled treatment has a significant mean impact of 4.1 students (baseline mean enrollment is 100 students).

Second, with respect to exam participation rates, mean impacts of the treatment variants also either turn from negative to positive or are positive and grow in size over the three years. Third-year mean impacts are 3.4 ppts for *HT only*, 2.8 ppts for *all T*, and 3.6 ppts for *HT+*, and are not statistically different from each other. The mean impacts for all treatment variants are significant. The pooled treatment has a significant mean impact of 3.3 ppts. Given a baseline exam participation rate of 97 percent, the mean impact maximizes the rate.

Third, with respect to student exam scores, mean impacts of the treatment variants do not exhibit a monotonic pattern of change over the three years. Only treatment variant *HT+* has consistent positive mean impacts of .09, .14, and .05 sds in the first, second, and third years, but they are all insignificant. Mean impacts across the years for the other two treatment variants are also all insignificant. The pooled treatment has mean impacts of .02, .08, and –.02 sds in the first, second, and third years, but they are all insignificant.

Potential strategic manipulation of enrollment data

The significant positive mean impacts on school enrollment in the third year may be due to program schools inflating enrollment. While schools do not have direct control over ASC

we find significant mean impacts mainly in the third year. Results from pairwise and joint tests suggest that baseline means do not differ between treated and untreated schools in the attrited sample.

survey data, MEAs capture enrollment figures from school enrollment registers, which are under the control of schools.[19]

Children mainly join public primary schools in preschool and first grade; re-entrants or transfers into higher grades are rare and dropouts much more frequent. Hence, we expect that enrollment in grade g in year y is equal to or less than enrollment in grade $g-1$ in year $y-1$. Enrollment inflation in year y can flatten or invert this gradient. We test for this by estimating an OLS regression of the form

$$E_{g,3,s} = \beta_0 + \beta_1 E_{g-t,3-t,s} + \beta_2 T_s E_{g-t,3-t,s} + \beta_3 1\{t=2\} + \gamma_g + \mu_s + \varepsilon_{g,3,s}, \qquad (3)$$

where $E_{g,3,s}$ denotes enrollment in grade g in the third program year in school s, T_s the pooled treatment indicator, $E_{g-t,3-t,s}$ enrollment in grade $g-t$ in program year $3-t$ in school s, where t denotes a one-year $(t=1)$ or two-year split $(t=2)$, $1\{t=2\}$ an indicator for a two-year time split, γ_g grade g indicators, and μ_s school indicators. A positive β_1 is taken to indicate inflation in untreated schools and a positive β_2 the differential inflation in treated schools. We find $\hat{\beta}_1 = -.53$ $(p=.000)$ and $\hat{\beta}_2 = -.03$ $(p=.703)$. Thus, we do not find evidence consistent with potential enrollment inflation.

Potential strategic manipulation of exam grade size

Increased exam participation rates may be due to more fifth-grade students taking the exam (an increase in the fraction’s numerator) or to fewer fifth-grade students (a decrease in the fraction’s denominator), or to both. Exam grade size can decline if teachers hold back students in earlier grades or if teachers expel fifth-grade students (although the bonus’s link to school enrollment would deter such behavior).[20] Evidence of a negative mean impact on fifth-grade enrollment would be consistent with strategic manipulation of exam grade size by program school teachers.

[19] The provincial education department reports that MEAs crosscheck school enrollment figures against the actual presence of students but specific written guidelines are not provided.

[20] Holding back or expelling academically-weaker students can provide the added benefit of higher student exam scores.

Table 4 reports estimated mean impacts on fifth-grade enrollment in each of the program years. Mean impacts are negative for each of the treatment variants and the pooled treatment in the third year. However, only the mean impact for *HT only* of –1.1 students (or –6 percent given a baseline mean fifth-grade enrollment of 18.3 students) is significant. The pooled treatment's mean impact of –.8 students is insignificant. Thus, we find inconsistent evidence of strategic manipulation of exam grade size by program schools.

Potential negative selection into examtaking

We find that the program increases exam participation rates in the third year. If the program induces weaker students to participate in the exam, the mean impact on student exam scores will be biased downwards.

Adopting the approach by Angrist, Bettinger, and Kremer (2006), we re-estimate mean impacts on student exam scores by fitting a Tobit model to the data, artificially censoring the student exam score distribution at selected alternative low percentiles. All exam takers with scores below the censoring percentile are assigned the score at the censoring percentile. Assuming that (the few) exam nontakers would have scores below the artificial censoring point if they had taken the exam, exam nontakers are also assigned the score at the censoring percentile. Recall that mean impacts on student exam scores estimated via ordinary least squares (reported in Table 3) are insignificant. If mean impacts estimated via Tobit are positive and significant, it suggests that negative selection into examtaking may be behind the least squares results.

Table 5 reports third-year mean impacts on student exam scores estimated via Tobit at four different censoring percentiles: 5th, 10th, 15th, and 20th. Mean impacts for the treatment variants, and the pooled treatment, are largely stable across the four censoring percentiles and are all insignificant. Thus, we do not find evidence of potential negative selection into examtaking.

B. Program impacts by location

In post-hoc subgroup analysis, we examine whether mean impacts vary by school location.[21] Location is defined as whether the school is urban or rural. Only 48 schools (or 8

[21] We also examine whether mean impacts vary by baseline mean exam scores, and find that they generally do not. Results are available upon request.

percent) are in the urban subsample, and, hence, the analysis of their mean impacts may be underpowered.

Table 6 reports baseline mean outcomes and characteristics for urban and rural schools (Columns 1 and 2), and pairwise differences in means between the two subsamples (Column 3). Compared to rural schools, urban schools have a higher mean number of school amenities (+.3 out of 4 amenities), mean number of teachers (+2.3 teachers), mean number of students in first through fifth grade (+40.2 students), mean number of students in fifth grade (+6.7 students), mean years of government service for teachers (+3.4 years), and mean exam scores (+.4 sds). Urban schools also have a lower mean student-teacher ratio (–18 students per teacher) than rural schools.

Table 6 also reports differences in baseline mean outcomes and characteristics between treated and untreated schools in the urban and rural subsamples (Columns 4 and 5). Treated and untreated schools differ in two dimensions in the urban subsample and one dimension in the rural subsample. The differences in means, when present, are large: the share of schools with buildings in satisfactory condition is 18.4 ppts lower in treated than in untreated schools in the urban subsample; the share of coeducational schools is 11.2 ppts lower in treated than in untreated schools in the rural subsample. Nevertheless, we cannot reject that the differences in means are jointly equal to zero for our school-, teacher-, and enrollment and exam score-related groupings of dimensions (Panels A, B, and C) in the urban and rural subsamples.

We test for differential mean impacts between urban and rural schools by estimating an OLS regression of the form

$$Y_s = \beta_0 + \beta_1 T_s + \beta_2 U_s + \beta_3 T_s U_s + X_{s,0}\gamma + \mu_d + \varepsilon_s, \qquad (4)$$

where U_s denotes whether or not the school is urban, and all other variables are as defined previously in (2). The mean impact for rural schools is given by β_1, the differential mean impact for urban schools by β_3, and the net mean impact for urban schools by $\beta_1 + \beta_3$.

As before, we estimate alternative variants of (3). We estimate regressions (a) where the three treatment variants are pooled, (b) where the treatment variants are included additively, and (c) separately for each of the three program years. Regressions for enrollment and exam participation rates are at the school level and standard errors are clustered at the tehsil level;

regressions for exam scores are at the student level and standard errors are clustered at the school level.

Table 7 reports estimated differential mean impacts by school location. Mirroring the findings for the overall evaluation sample, significant mean impacts are concentrated in the third year. First, with respect to school enrollment, third-year mean impacts for the urban subsample relative to the rural subsample are an additional 34 students for *HT only*, 27 students for *all T*, and 29 students for *HT+*, but are significant only for *HT only* and *all T*. The pooled treatment has a significant third-year mean impact on enrollment of an additional 30 students in the urban subsample. The pooled treatment's *net* third-year mean impact in the urban subsample is 32 students (or 23 percent given a baseline mean enrollment of 137 students in urban schools).

Second, with respect to exam participation rates, third-year mean impacts in the rural subsample are 3.7 ppts for *HT only*, 3.0 ppts for *all T*, and 4.2 ppts for *HT+*, and are all significant. The third-year net mean impacts of the treatment variants in the urban subsample are all negative and insignificant. The pooled treatment has a significant third-year mean impact of 3.6 ppts in the rural subsample (and an insignificant third-year net mean impact of –1.4 ppts in the urban subsample).

Third, with respect to student exam scores, third-year (differential) mean impacts of the treatment variants and the pooled treatment are insignificant for the rural (urban) subsamples. The pooled treatment however has a significant first-year differential mean impact of .34 sds in the urban subsample. While the corresponding mean impacts are .37 sds for *HT only*, .29 sds for *all T*, and .34 sds for *HT+*, only the impact for *HT only* is significant.

We do not find that the differential mean impact on school enrollment for the urban subsample is due to enrollment inflation. Nor does it appear that the mean impact on exam participation rates for the rural subsample is due to manipulation of exam grade size. We check third-year means of basic school and teacher characteristics for treated and untreated schools separately for the urban and rural subsamples. In each subsample, the means are largely similar, thus failing to suggest potential pathways behind detected impacts.[22]

Lastly, we examine how the third-year mean impact on school enrollment in the urban subsample is spread across grades. Table 8 reports estimated differential mean impacts of the

[22] Results for the two tests of potential strategic behavior and the estimated third-year means of basic school and teacher characteristics are available upon request.

pooled treatment on grade-specific enrollment by school location. Depending on the grade, differential mean impacts for the urban subsample turn from negative to positive or are positive and grow in size over the three years. The third-year differential mean impact in the urban subsample is 12 students in first grade, and declines monotonically by grade to 1.9 students in fifth grade. As a *percentage* of baseline mean grade-specific enrollment in urban schools, the differential mean impacts across grades are similar in size. The third-year differential mean impacts are significant for all but fifth grade.

V. Conclusion

Most evaluations of teacher incentives in low- and middle-income countries can be seen as efficacy trials. This study reports findings from an effectiveness trial in a low-income setting, where, in contrast to the efficacy trials, the government was responsible for the design and implementation of a teacher performance pay program in public primary schools, including using their own administrative data for measuring performance. Monetary bonuses provided to teachers were linked to gains in school enrollment; to gains in school mean scores in a fifth-grade standardized exam; and to the level of participation in the exam by fifth-grade students. To test whether leveraging head teachers either suffices or further augments program impacts, the receipt and size of the bonus were also randomly assigned across schools according to whether or not the teacher was the school's head teacher.

Looking at the first three years of implementation, and using the government's program administrative data for the evaluation, we find mixed evidence of significant mean impacts on school enrollment, exam participation rates, and student exam scores. Significant positive mean impacts are concentrated in the third year. The program raises school enrollment and exam participation rates in the third year. The trends in impacts over the three years however suggest possible learning by teachers on how to respond to the program or possible growing confidence among teachers in the administration and continuation of the program. We do not find that the program raises student exam scores in any year. Inference results for the program as a whole are mirrored by each of the treatment variants in the case of exam participation rates and mean student exam scores and by the treatment variant that offers bonuses solely to head teachers in the case of school enrollment. Post-hoc subgroup analysis reveals that the third-year program mean impacts on school enrollment and exam participation rates in the total sample are driven by the urban and rural subsamples, respectively. Program mean impacts on student exam scores differ between the urban and rural subsamples in the first year but the differences in later years are not significant.

Given that the program increases exam participation rates, we test whether the absence of significant positive mean impacts on student exam scores is due to negative selection of students into examtaking. We do not find evidence supporting this hypothesis. Aside from this test, we are left to speculate on factors that could have blocked impacts. Some studies argue or show that implementation failures can explain the lack of program impacts.[23] We discount that the absence of positive mean impacts on student exam scores is due to this problem. Third-party validations indicate that the program has been implemented satisfactorily, and markedly better than many other programs administered by the provincial 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 fully and in a timely and efficient way. The PEC exam and ASC survey activities were also carried out as scheduled. Credibility in the program was reportedly low in the first year of implementation: program school teachers conveyed that they were initially skeptical that they would receive bonuses (consistent with program rules). This skepticism appears to have faded away after program school teachers experienced how the government performed in running the program and they received their bonus awards.

We also discount that the absence of positive mean impacts on student exam scores is due to bonus awards being too low, the bonus formula being too complex for teachers to understand, or the group-based nature of the bonuses (in the case of the two treatment variants that offer bonuses to all teachers) inducing significant freeriding. The mean bonus payout was 16 percent of the yearly basic salary for a teacher at the baseline mean pay grade and with baseline mean years of government service.[24] All definitions and data sources for bonus calculations were communicated by program administrators to program school teachers during school visits and at

[23] For example, a new program may experience teething problems (see Bouguen, Filmer, Macours, and Naudeau 2013). In such cases, positive impacts may only materialize well after program inception, when problems are effectively resolved. Separately, implementation performance may be sensitive to program implementer characteristics such as the implementer’s interest, commitment, capability, and effort (Allcott 2015), and governments may be more likely or have larger shortfalls in these characteristics relative to, for example, (certain types of) NGOs (Bold et al 2013).

[24] This relative size is several percentage points higher than, for example, the bonuses in teacher-incentive field experiments in India (Muralidharan and Sundararaman 2011) and Kenya (Glewwe et al 2010); both experimental evaluations find positive mean program impacts.

orientation meetings for every bonus round. Teacher bonuses are also linked to own-school performance only, and not, for example, to own-school performance compared to other schools—the latter formulation could introduce added complexity and uncertainty for teachers (Fryer 2013). The mean 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 than in larger schools with an identical bonus design (Goodman and Turner 2013).

We posit three hypotheses for the absence of positive mean impacts on student exam scores. First, the binding constraint may be external to the link between teacher effort and student exam scores. Nonschool and school factors outside the control of teachers may make improving student exam scores difficult even with greater teacher effort. Although the public schools targeted by the program were those with the lowest mean student exam scores in the province, and thus have ample room for improvement, the targeting places the program under circumstances that may be especially handicapping. In the face of these external constraints, teachers may choose to withhold greater effort or see expended effort go to waste.

Second, the binding constraint may be that teachers lack the knowhow to improve student exam scores, essentially blocking the link between teacher effort and student exam scores. Teachers may not know what exact strategies to pursue to raise student exam scores, and whether what they intend to pursue is the most efficient way to raise student exam scores. In the face of this uncertainty, teachers may withhold greater effort or misdirect their effort (Fryer 2013).

Third, teachers may optimally choose to direct their effort at those incentivized margins—school enrollment, exam participation rates, and student exam scores—that maximize payoffs net of effort costs (Neal 2011). In this case, it appears that the net payoff-maximizing margin is school enrollment in urban areas, exam participation rates in rural areas, and not student exam scores anywhere. The extent to which movement on a given margin pays off would depend on how accurately movement is measured, and teacher effort would be sensitive to this (Baker 2002). Measurement error may be higher for student scores in PEC exams than enrollment information in the ASC surveys. Notwithstanding, if measurement error is large, PEC exam scores would be a highly imperfect measure of latent student academic achievement. The latter is one of the main targets of the program.

References

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." Forthcoming *Quarterly Journal of Economics*.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, Tara Vishwanath, and Tristan Zajonc. 2009. *Learning and Educational Achievements in Punjab Schools (LEAPS) Report*. Washington, DC: World Bank.

Angrist, Joshua, Eric Bettinger, and Michael Kremer. 2006. "Long-term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review* 96 (3): 847–62.

Baker, George P., 2002. "Distortion and Risk in Optimal Incentive Contracts." *Journal of Human Resources* 37 (4): 728–52.

Baker, George P. 1992. "Incentive Contracts and Performance Measurement." *Journal of Political Economy* 100 (3): 598–614.

Barlevy, Gadi, and Derek Neal. 2012. "Pay for Percentile." *American Economic Review* 102 (5): 1805–31.

Behrman, Jere H., 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." Forthcoming *Journal of Political Economy*.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2013. "Scaling-up What Works: Experimental Evidence on External Validity in Kenyan Education." Centre for the Study of African Economies Working Paper Series 2013-04.

Bouguen, Adrien, Deon Filmer, Karen Macours, and Sophie Naudeau. 2013. "Impact Evaluation of Three Types of Early Childhood Development Interventions in Cambodia." World Bank Policy Research Working Paper Series No. 6540.

Boyd, Donald, Pamela Grossman, Marsha Ing, Hamilton Lankford, Susanna Loeb, and James Wyckoff. 2011. "The Influence of School Administrators on Teacher Retention Decisions." *American Education Research Journal* 48 (2): 303–33.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers. 2006. "Missing in Action: Teacher and Health Worker Absence in Developing Countries." *Journal of Economic Perspectives* 20 (1): 91–116.

Cohn, Alain, Ernst Fehr, Benedikt Herrmann, and Frederic Schneider. 2014. "Social Comparison and Effort Provision: Evidence from a Field Experiment." *Journal of the European Economic Association* 12 (4): 877–98.

Darling-Hammond, Linda, and Robert Rothman. 2011. *Teacher and Leader Effectiveness in High-Performing Education Systems*. Washington, DC: Alliance for Excellent Education.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102 (4): 1241–78.

Figlio, David, and Joshua Winicki. 2005. "Food for Thought: The Effects of School Accountability Plans on School Nutrition." *Journal of Public Economics* 89 (2–3): 381–94.

Fryer, Roland G. 2013. "Teacher Incentives and Student Achievement: Evidence from New York City Public Schools." *Journal of Labor Economics* 31 (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." National Bureau of Economic Research Working Paper Series 18237.

Glewwe, Paul and Michael Kremer. 2006. "School, Teachers, and Education Outcomes in Developing Countries." In Eric A. Hanushek and Finis Welch (Eds), *Handbook of the Economics of Education,* Volume 2. Amsterdam: Elsevier B.V.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2 (3): 205–27

Goldhaber, Dan. 2010. When the Stakes Are High, Can We Rely on Value-Added? Exploring the Use of Value-Added Models to Inform Teacher Workforce Decisions. Center for American Progress.
<https://cdn.americanprogress.org/wp-content/uploads/issues/2010/12/pdf/vam.pdf>.

Goodman, Sarena, and Lesley J. Turner. 2013. "The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program." *Journal of Labor Economics* 31 (2): 409–20.

Grissom, Jason A., and Susanna Loeb. 2011. "Triangulating Principal Effectiveness: How Perspectives of Parents, Teachers and Assistant Principals Identify the Central Importance of Managerial Skills." *American Educational Research Journal* 48 (5): 1091–1123.

Hanushek, Eric A., and Steven G. Rivkin. 2006. "Teacher Quality." In Eric A. Hanushek and Finis Welch (Eds), *Handbook of the Economics of Education,* Volume 2. Amsterdam: Elsevier B.V.

Holmstrom, Bengt. 1982. "Moral Hazard in Teams." *Bell Journal of Economics* 13 (2): 324–40.

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

Ingersoll, Richard M. 2001. “Teacher Turnover and Teacher Shortages: An Organizational Analysis.” *American Economic Research Journal* 38 (3): 499–534.

Jacob, Brian A. 2005. “Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools.” *Journal of Public Economics* 89 (5): 761–96.

Jacob, Brian A., and Steven D. Levitt. 2003. “Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating.” *Quarterly Journal of Economics* 118 (3): 843–77.

Koedel, Cory, Kata Mihaly, and Jonah E. Rockoff (forthcoming). Value-Added Modeling: A Review. *Economics of Education Review*.

Koretz, Daniel. 2002. “Limitations in the Use of Achievement Tests as Measures of Educators’ Productivity.” *Journal of Human Resources* 37 (4): 752–77.

Lavy, Victor. 2002. “Evaluating the Effect of Teachers’ Group Performance Incentives on Pupil Achievement.” *Journal of Political Economy* 110 (6): 1286–1317.

Mulford, Bill. 2003. “School Leaders: Changing Roles and Impact on Teacher and School Effectiveness.” Paper for OECD Education and Training Policy Division, Attracting, Developing and Retaining Effective Teachers Activity.

Muralidharan, Karthik. 2012. “Long-Term Effects of Teacher Performance Pay: Experimental Evidence from India.” http://econweb.ucsd.edu/~kamurali/papers/Working%20Papers/Long%20Term%20Effects%20of%20Teacher%20Performance%20Pay.pdf/

Muralidharan, Karthik, and Venkatesh Sundararaman. 2011. “Teacher Performance Pay: Experimental Evidence from India.” *Journal of Political Economy* 119 (1): 39–77.

Neal, Derek. 2011. “The Design of Performance Pay in Education.” In E. Hanushek, S. Machin, and L. Woessmann (Eds), *Handbook of the Economics of Education*, Volume 4, 495–550. Amsterdam: Elsevier B.V.

Pakistan Bureau of Statistics, Government of Pakistan. 2014. *Pakistan Social and Living Standards Measurement Survey (PSLM) 2012-13*. Islamabad: Federal Bureau of Statistics.

Robinson, Viviane M. J., Claire A. Lloyd, and Kenneth J. Rowe. 2008. “The Impact of Leadership on Student Outcomes: An Analysis of the Differential Effects of Leadership Types.” *Educational Administration Quarterly* 44 (5): 635–74.

Springer, Matthew G., Dale Ballou, Laura Hamilton, Vi-Nhuan Le, J.R. Lockwood, Daniel F. McCaffrey, Matthew Pepper, and Brian M. Stecher. 2011. “Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching (POINT).” *Society for Research on Educational Effectiveness*.

Sojourner, Aaron J., Elton Mykerezi, and Kristine L. West. 2011. “Teacher Pay Reform and Productivity: Panel Data Evidence from Adoptions of Q-Comp in Minnesota. *Journal of Human Resources* 49 (4): 945–81.

United Nations Development Programme (UNDP). 2013. *Human Development Report 2013: The Rise of the South: Human Progress in a Diverse World*. New York: UNDP.

United Nations Education, Scientific and Cultural Organization (UNESCO). 2014. *Education for All-Global Monitoring Report 2013/4: Teaching and Learning: Achieving Quality for All*. Paris, France: UNESCO.

Waters, Tim, Robert J. Marzano, and Brian McNulty. 2003. *Balanced leadership: What 30 years of research tells us about the effect of leadership on student achievement*. Aurora, CO: Mid-continent Research for Education and Learning.

World Bank. 2011. *What Matter Most for Teacher Policies: A Framework Paper*. Washington, DC: World Bank.

World Bank, World Development Indicators. 2012. GNI per capita, Atlas method; Population. Available from http://data.worldbank.org/data-catalog/world-development-indicators.

Table 1. Bonus formula by treatment variant and teacher type

Bonus treatment type	Number of schools	Teacher type	
		Head teachers (*HT*)	Teachers (*T*)
HT only	150	1,000 × CS	0
T	150	1,000 × CS	1,000 × CS
HT+	150	2,000 × CS	1,000 × CS
Control	150	0	0

Table 2. Baseline means

Variable	Evaluation sample	All schools	Difference (1)–(2)	Treated	Untreated	Difference (4)–(5)
	(1)	(2)	(3)	(4)	(5)	(6)
			A. School related			
School age	42.67	39.56	3.112***	42.65	42.75	–0.1
	(17.90)	(29.67)	(1.10)	(18.05)	(17.50)	(1.61)
Urban	0.08	0.08	–0.002	0.09	0.06	0.027
	(0.27)	(0.27)	(0.02)	(0.28)	(0.24)	(0.02)
Girls-only	0.11	0.26	–0.150***	0.1	0.13	–0.024
	(0.31)	(0.44)	(0.02)	(0.30)	(0.33)	(0.03)
Coeducational	0.51	0.43	0.084*	0.49	0.58	–0.091**
	(0.50)	(0.49)	(0.05)	(0.50)	(0.50)	(0.05)
Classrooms	3.19	2.75	0.432***	3.19	3.16	0.027
	(1.63)	(1.68)	(0.17)	(1.64)	(1.60)	(0.12)
Building in satisfactory condition	0.86	0.84	0.027**	0.86	0.86	–0.002
	(0.34)	(0.37)	(0.01)	(0.34)	(0.34)	(0.03)
Amenities index	3.09	2.79	0.302**	3.09	3.11	–0.018
	(1.05)	(1.17)	(0.13)	(1.02)	(1.14)	(0.11)
H_0 (Differences jointly zero), *p*-value						0.301
			B. Teacher related			
Teachers	2.96	2.77	0.188	2.98	2.91	0.064
	(1.66)	(1.81)	(0.12)	(1.70)	(1.53)	(0.12)
Years of government teaching service	17.27	17.03	0.242	17.52	16.54	0.979*
	(7.04)	(7.34)	(0.56)	(7.18)	(6.59)	(0.54)
Years of service in current school	9.8	10.07	–0.266	9.84	9.69	0.142
	(5.49)	(5.69)	(0.39)	(5.49)	(5.48)	(0.46)
Less than high school diploma	0.34	0.43	–0.094***	0.34	0.31	0.03
	(0.34)	(0.37)	(0.03)	(0.35)	(0.34)	(0.03)
High school diploma	0.17	0.16	0.013	0.17	0.17	–0.001
	(0.26)	(0.26)	(0.02)	(0.26)	(0.26)	(0.02)
Bachelor’s (or equivalent) degree	0.35	0.24	0.111***	0.34	0.36	–0.018
	(0.34)	(0.32)	(0.03)	(0.34)	(0.34)	(0.03)
At least a master’s degree	0.14	0.17	–0.030*	0.14	0.15	–0.008
	(0.25)	(0.29)	(0.02)	(0.25)	(0.24)	(0.02)
H_0 (Differences jointly zero), *p*-value						0.686
			C. Enrollment and exam scores related			
Enrollment, grades 1 to 5	99.98	77.8	22.187***	100.16	99.47	0.691
	(51.69)	(59.06)	(4.31)	(51.59)	(52.17)	(5.58)
Enrollment, grade 5	18.25	11.16	7.084***	18.4	17.79	0.604
	(9.81)	(10.82)	(0.77)	(10.01)	(9.20)	(0.98)
Exam participation rate, grade 5	0.97	0.93	0.044***	0.97	0.98	–0.005
	(0.06)	(0.14)	(0.01)	(0.06)	(0.06)	(0.01)
Student-classroom ratio	52.48	44.16	8.323**	52.34	52.89	–0.548
	(36.99)	(29.59)	(3.79)	(38.26)	(32.94)	(4.03)
Student-teacher ratio	56.07	44.15	11.924***	56.22	55.62	0.604
	(36.72)	(28.36)	(3.90)	(36.96)	(36.13)	(4.22)
Mean exam scores (sds.)	0	4.18	–4.186***	–0.01	0	–0.005
	(1.01)	(3.09)	(0.17)	(1.02)	(1.00)	(0.07)
H_0 (Differences jointly zero), *p*-value						0.936
Observations	600	42,438	43,038	450	150	600

Notes: * denotes significant at the 10% level; ** at the 5% level; and *** at the 1% level. Standard deviations are reported in parentheses in Columns (1), (2), (4) and (5). Standard errors are reported in parentheses in Columns (3) and (6). Estimated standard errors are clustered at the tehsil level. “All schools” is defined as all functional primary schools in Punjab. Baseline data are from the 2009 Annual School Census and the 2010 Punjab Examination Commission fifth-grade exam.

Table 3. Program impacts

	Enrollment (grades 1–5)			Exam participation rate (grade 5)			Student exam scores (grade 5)		
	Program year								
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Pooled T	–1.587	–0.222	4.067*	–0.006	0.005	0.033***	0.016	0.078	–0.015
	(1.40)	(1.32)	(2.39)	(0.01)	(0.01)	(0.01)	(0.07)	(0.08)	(0.09)
Untreated group mean	93.77	87.61	83.73	0.950	0.920	0.940	0.000	0.000	0.000
	(48.98)	(48.03)	(45.57)	(0.10)	(0.11)	(0.12)	(1.00)	(1.00)	(1.00)
R-squared statistic	0.882	0.812	0.711	0.093	0.072	0.065	0.111	0.165	0.149
				Treatment variants					
HT only	–1.564	–2.348	3.946*	–0.015	–0.015	0.034***	–0.004	0.047	–0.026
	(1.31)	(1.91)	(2.02)	(0.01)	(0.02)	(0.01)	(0.09)	(0.10)	(0.11)
All T	1.517	3.655*	5.794	–0.005	0.008	0.028*	–0.042	0.040	–0.077
	(2.20)	(2.13)	(3.70)	(0.01)	(0.01)	(0.02)	(0.09)	(0.11)	(0.11)
HT+	–4.667**	–1.889	2.531	0.000	0.023	0.036**	0.085	0.140	0.048
	(2.04)	(2.28)	(3.66)	(0.01)	(0.01)	(0.01)	(0.09)	(0.10)	(0.10)
R-squared statistic	0.884	0.814	0.711	0.095	0.084	0.066	0.113	0.166	0.151
H_0 (*HT only*=*All T*), *p*-value	0.214	0.038	0.634	0.386	0.204	0.600	0.658	0.951	0.647
H_0 (*HT only*=*HT+*), *p*-value	0.192	0.879	0.654	0.147	0.062	0.790	0.342	0.380	0.467
H_0 (*All T*=*HT+*), *p*-value	0.009	0.094	0.461	0.609	0.230	0.462	0.168	0.382	0.249
Observations	598	593	583	597	591	575	9,030	8,085	8,211

Notes: * denotes significant at the 10% level; ** at the 5% level; and *** at the 1% level. Robust standard errors are reported in parentheses. Standard errors in the enrollment and exam participation rate regressions are clustered at tehsil level. Standard errors in the student exam score regressions are clustered at the school level. All regressions control for districts and baseline outcomes and characteristics.

Table 4. Examining potential strategic manipulation of exam grade size

	Enrollment, grade 5		
	Program year		
	1st	2nd	3rd
	(1)	(2)	(3)
Pooled T	0.128	–0.353	–0.766
	(0.61)	(0.56)	(0.58)
Untreated group mean	15.36	14.71	14.72
	(9.74)	(9.50)	(9.67)
R-squared statistic	0.72	0.678	0.673
Treatment variants			
HT only	1.080**	–1.402***	–1.108***
	(0.53)	(0.53)	(0.40)
all T	0.371	0.665	–0.448
	(0.87)	(0.76)	(0.83)
HT+	–1.054	–0.302	–0.743
	(0.66)	(0.71)	(0.87)
R-squared statistic	0.725	0.683	0.673
H_0 (*HT only=all T*), *p*-value	0.314	0.013	0.383
H_0 (*HT only=HT+*), *p*-value	0.002	0.055	0.690
H_0 (*all T=HT+*), *p*-value	0.038	0.242	0.673
Observations	598	593	583

Notes: * denotes significant at the 10% level; ** at the 5% level; and *** at the 1% level. Robust standard errors, clustered at the tehsil level, are reported in parentheses. All regressions control for districts and baseline outcome and characteristics.

Table 5. Examining potential negative selection into examtaking
Tobit mean impact estimates for student exam scores

Treatment	Censoring percentile			
	5th	10th	15th	20th
Pooled T	–0.017	–0.014	–0.014	–0.014
	(0.09)	(0.09)	(0.09)	(0.09)
	Treatment variants			
HT only	–0.025	–0.022	–0.021	–0.009
	(0.11)	(0.11)	(0.11)	(0.11)
All T	–0.083	–0.081	–0.086	–0.092
	(0.11)	(0.11)	(0.12)	(0.12)
HT+	0.048	0.052	0.055	0.052
	(0.10)	(0.11)	(0.11)	(0.11)

Notes: Robust standard errors, clustered at the school level, are reported in parentheses. Student exam scores are standardized using the mean and standard deviation for students in untreated schools. All regressions control for districts and baseline outcomes and characteristics.

Table 6. Baseline means by location (urban versus rural)

	Urban	Rural	Difference	Treated–untreated	
			(1)–(2)	Urban	Rural
	(1)	(2)	(3)	(5)	(6)
		A. School related			
School age	42.96	42.65	0.312	5.419	–0.494
	(21.07)	(17.62)	(2.84)	(5.27)	(1.55)
Girls-only	0.06	0.11	–0.05	–0.06	–0.021
	(0.24)	(0.32)	(0.05)	(0.10)	(0.03)
Coeducational	0.50	0.51	–0.013	0.205	–0.112***
	(0.51)	(0.50)	(0.06)	(0.20)	(0.04)
Classrooms	3.70	3.14	0.562	0.181	0.002
	(2.07)	(1.58)	(0.44)	(0.53)	(0.10)
Building in satisfactory condition	0.85	0.86	–0.013	–0.184***	0.011
	(0.36)	(0.34)	(0.04)	(0.04)	(0.03)
Amenities index	3.38	3.07	0.306**	–0.085	–0.022
	(0.79)	(1.07)	(0.15)	(0.26)	(0.11)
H₀ (all differences jointly zero). *p*-value				0.103	0.116
		B. Teacher related			
Teachers	5.08	2.78	2.307***	1.06	–0.072
	(1.96)	(1.50)	(0.41)	(0.80)	(0.14)
Years of government service	20.41	17.00	3.418***	–1.263	1.038
	(4.77)	(7.14)	(0.96)	(1.92)	(0.65)
Years of service in current school	10.35	9.75	0.6	–2.201	0.288
	(4.46)	(5.57)	(0.63)	(3.11)	(0.52)
Less than high school diploma	0.39	0.33	0.053	–0.160***	0.042
	(0.28)	(0.35)	(0.04)	(0.05)	(0.04)
High school diploma	0.20	0.17	0.025	0.051	–0.005
	(0.26)	(0.26)	(0.03)	(0.05)	(0.03)
Bachelor’s (or equivalent) degree	0.30	0.35	–0.057	0.053	–0.021
	(0.27)	(0.35)	(0.05)	(0.07)	(0.03)
At least a master’s degree	0.11	0.14	–0.034	0.062	–0.012
	(0.15)	(0.25)	(0.03)	(0.04)	(0.02)
H₀ (all differences jointly zero). *p*-value				0.123	0.598
		C. Enrollment and exam scores related			
Enrollment, grades 1 to 5	136.96	96.77	40.188***	18.684	–1.709
	(69.54)	(48.61)	(10.39)	(22.97)	(6.15)
Enrollment, grade 5	24.44	17.71	6.729***	2.453	0.285
	(13.99)	(9.18)	(2.31)	(4.66)	(0.87)
Exam participation rate, grade 5	0.97	0.97	0.002	–0.004	–0.005
	(0.05)	(0.07)	(0.01)	(0.02)	(0.01)
Student classroom ratio	62.09	51.65	10.447	10.587	–1.611
	(52.13)	(35.32)	(8.13)	(14.71)	(4.47)
Student teacher ratio	39.48	57.52	–18.038***	1.406	1.072
	(21.64)	(37.42)	(5.12)	(4.56)	(4.76)
Mean exam scores (sds.)	0.37	–0.04	0.411***	0.038	–0.02
	(0.78)	(1.03)	(0.11)	(0.29)	(0.07)
H₀ (All differences jointly zero). *p*-value				0.961	0.856
Observations	48	552	600	48	552

Notes: .* denotes significant at the 10% level; ** at the 5% level; and *** at the 1% level. Standard deviations are reported in parentheses in Columns 1 and 2. Robust standard errors are reported in parentheses in Columns 3–5

Table 7. Program impacts by location (urban versus rural)

	Enrollment (grades 1 to 5)			Exam participation rate (grade 5)			Student exam scores (grade 5)		
	Program year								
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Pooled T	–1.77	–1.274	1.953	–0.006	0.007	0.036***	–0.014	0.077	-0.013
	(1.48)	(1.16)	(2.35)	(0.01)	(0.02)	(0.01)	(0.08)	(0.08)	(0.09)
T×Urban (*U*)	2.681	15.24	30.429***	0	–0.028	–0.050***	0.339**	0.008	-0.023
	(4.85)	(9.44)	(9.86)	(0.05)	(0.03)	(0.02)	(0.16)	(0.25)	(0.31)
R-squared statistic	0.882	0.813	0.715	0.093	0.073	0.068	0.112	0.165	0.149
				Treatment variants					
HT only	–1.752	–4.093*	1.285	–0.012	–0.015	0.037***	–0.04	0.036	-0.052
	(1.75)	(2.09)	(2.63)	(0.01)	(0.02)	(0.01)	(0.10)	(0.11)	(0.11)
HT only×U	2.696	21.285***	33.944***	–0.027	–0.009	–0.044**	0.374*	0.097	0.172
	(6.17)	(7.64)	(9.69)	(0.07)	(0.03)	(0.02)	(0.21)	(0.30)	(0.34)
All T	1.524	3.208	4.051	–0.006	0.01	0.030*	–0.065	0.015	-0.05
	(2.26)	(2.11)	(3.67)	(0.01)	(0.01)	(0.02)	(0.09)	(0.11)	(0.12)
All T×U	–0.64	6.597	26.857***	0.009	–0.035	–0.027	0.293	0.338	-0.282
	(7.14)	(5.59)	(8.09)	(0.05)	(0.03)	(0.02)	(0.20)	(0.38)	(0.42)
HT+	–5.067**	–2.874	0.605	–0.002	0.025	0.042***	0.058	0.173*	0.051
	(2.07)	(2.11)	(3.12)	(0.01)	(0.02)	(0.02)	(0.10)	(0.10)	(0.11)
HT+×U	5.564	15.796	29.251	0.029	–0.036	–0.077**	0.336	–0.349	-0.003
	(9.17)	(19.42)	(21.18)	(0.04)	(0.05)	(0.03)	(0.28)	(0.30)	(0.36)
R-squared statistic	0.884	0.816	0.716	0.098	0.085	0.072	0.114	0.171	0.154
Ho: ((HT only × U) = (All T×U)), p-value	0.722	0.038	0.491	0.596	0.455	0.383	0.732	0.544	0.215
Ho: ((HT only×U) = (HT+× U)), p-value	0.809	0.774	0.842	0.283	0.493	0.25	0.905	0.172	0.597
Ho: ((All T×U) = (HT+×U)), p-value	0.461	0.592	0.902	0.56	0.983	0.131	0.888	0.095	0.499
Observations	598	593	583	597	591	575	9,030	8,085	8,211

Notes: * denotes significant at the 10% level; ** at the 5% level; and *** at the 1% level. Robust standard errors are reported in parentheses. Standard errors in the enrollment and exam participation rate regressions are clustered at tehsil level; standard errors in the student exam score regressions are clustered at the school level. All regressions control for districts and baseline outcomes and characteristics.

Table 8. Program impacts on grade-specific enrollment by location (urban versus rural)

	Grade 1			Grade 2			Grade 3			Grade 4			Grade 5		
							Program year								
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
Pooled T	–1.095	0.543	2.257**	–0.278	–0.474	0.326	–0.637	–0.111	–0.541	–0.023	–0.805*	0.349	0.189	–0.246	–0.851
	(0.69)	(0.60)	(1.01)	(0.82)	(0.51)	(0.51)	(0.54)	(0.60)	(0.43)	(0.63)	(0.44)	(0.60)	(0.63)	(0.56)	(0.70)
T×Urban (*U*)	4.001*	5.685**	12.032***	3.712	5.915*	7.353***	–2.389	4.472*	4.691**	–1.913	0.649	4.690**	–0.561	–1.505	1.846
	(2.09)	(2.62)	(4.28)	(2.49)	(3.06)	(2.46)	(2.39)	(2.55)	(1.97)	(1.85)	(2.16)	(2.32)	(1.41)	(1.82)	(1.45)
R-squared statistic	0.593	0.598	0.452	0.758	0.641	0.552	0.753	0.702	0.625	0.728	0.72	0.673	0.719	0.681	0.671
Observations	600	599	597	600	599	597	600	599	597	600	599	597	600	599	597

Notes: * denotes significance at the 10% level; ** at the 5% level; and *** at the 1% level. Robust standard errors are reported in parentheses. Standard errors in the enrollment and exam participation rate regressions are clustered at tehsil level; standard errors in the student exam score regressions are clustered at the school level. All regressions control for districts and baseline outcomes and characteristics.