Forthcoming in the *Journal of Political Economy*

Teacher Performance Pay: Experimental Evidence from India

Karthik Muralidharan[†]
Venkatesh Sundararaman[‡]

5 January 2011[*]

Abstract: Performance pay for teachers is frequently suggested as a way of improving education outcomes in schools, but the theoretical predictions regarding its effectiveness are ambiguous and the empirical evidence to date is limited and mixed. We present results from a randomized evaluation of a teacher incentive program implemented across a large representative sample of government-run rural primary schools in the Indian state of Andhra Pradesh. The program provided bonus payments to teachers based on the average improvement of their students' test scores in independently administered learning assessments (with a mean bonus of 3% of annual pay). At the end of two years of the program, students in incentive schools performed significantly better than those in control schools by 0.27 and 0.17 standard deviations in math and language tests respectively. Students in incentive schools also performed better on subjects for which there were no incentives, suggesting positive spillovers. Group and individual incentive schools performed equally well in the first year of the program, but the individual incentive schools outperformed at the end of two years. Incentive schools performed significantly better than other randomly-chosen schools that received additional schooling inputs of a similar value.

JEL Classification: C93, I21, M52, O15

Keywords: teacher performance pay, teacher incentives, education, education policy, field experiments, compensation, India

[†] UC San Diego, NBER, BREAD, and J-PAL; E-mail: kamurali@ucsd.edu
[‡] South Asia Human Development Unit, World Bank. E-mail: vsundararaman@worldbank.org

[*] We are grateful to Caroline Hoxby, Michael Kremer, and Michelle Riboud for their support, advice, and encouragement at all stages of this project. We thank the editor Derek Neal, 2 anonymous referees, George Baker, Damon Clark, Julie Cullen, Gordon Dahl, Jishnu Das, Shanta Devarajan, Martin Feldstein, Richard Freeman, Robert Gibbons, Edward Glaeser, Roger Gordon, Sangeeta Goyal, Gordon Hanson, Richard Holden, Asim Khwaja, David Levine, Jens Ludwig, Sendhil Mullainathan, Ben Olken, Lant Pritchett, Halsey Rogers, Richard Romano, and various seminar participants for useful comments and discussions.

This paper is based on a project known as the Andhra Pradesh Randomized Evaluation Study (AP RESt), which is a partnership between the Government of Andhra Pradesh, the Azim Premji Foundation, and the World Bank. Financial assistance for the project has been provided by the Government of Andhra Pradesh, the UK Department for International Development (DFID), the Azim Premji Foundation, and the World Bank. We thank Dileep Ranjekar, Amit Dar, Samuel C. Carlson, and officials of the Department of School Education in Andhra Pradesh (particularly Dr. I.V. Subba Rao, Dr. P. Krishnaiah, and K. Ramakrishna Rao), for their continuous support and long-term vision for this research. We are especially grateful to DD Karopady, M Srinivasa Rao, and staff of the Azim Premji Foundation for their leadership and meticulous work in implementing this project. Sridhar Rajagopalan, Vyjyanthi Shankar, and staff of Education Initiatives led the test design. We thank Vinayak Alladi, Gokul Madhavan, and Ketki Sheth for outstanding research assistance. The findings, interpretations, and conclusions expressed in this paper are those of the authors and do not necessarily represent the views of the Government of Andhra Pradesh, the Azim Premji Foundation, or the World Bank.

1. Introduction

A fundamental question in education policy around the world is that of the relative effectiveness of input-based and incentive-based policies in improving the quality of schools. While the traditional approach to improving schools has focused on providing them with more resources, there has been growing interest in directly measuring and rewarding schools and teachers based on student learning outcomes. The idea of paying teachers based on direct measures of performance has attracted particular attention since teacher salaries are the largest component of education budgets and recent research shows that teacher characteristics rewarded under the status quo in most school systems – such as experience and master's degrees in education – are poor predictors of better student outcomes (see Rockoff 2004; Rivkin, Hanushek, and Kain 2005; and Gordon, Kane, and Staiger 2006).

However, while the idea of using incentive pay schemes for teachers as a way of improving school performance is increasingly making its way into policy,[1] the empirical evidence on the effectiveness of such policies is quite limited – with identification of the causal impact of teacher incentives being the main challenge. In addition, several studies have highlighted the possibility of perverse outcomes from teacher incentive and accountability programs (Jacob and Levitt 2003; Jacob 2005; Cullen and Reback 2006; and Neal and Schanzenbach 2010), suggesting the need for caution and better evidence before expanding teacher incentive programs based on student test scores.

In this paper, we contribute towards filling this gap with evidence from a large-scale randomized evaluation of a teacher performance pay program implemented in the Indian state of Andhra Pradesh (AP). We studied two types of teacher performance pay (group bonuses based on school performance, and individual bonuses based on teacher performance), with the average bonus calibrated to be around 3% of a typical teacher's annual salary. The incentive program was designed to minimize the likelihood of undesired consequences (see design details later) and the study was conducted by randomly allocating the incentive programs across a representative sample of 300 government-run schools in rural AP with 100 schools each in the group and individual incentive treatment groups and 100 schools serving as the comparison group.

[1] Teacher performance pay is being considered and implemented in several US states including Colorado, Florida, Tennessee, and Texas, and additional federal resources have been dedicated to such programs under the recent "Race to the Top" fund created by the US Department of Education in 2009. Other countries that have attempted to tie teacher pay to performance include Australia, Brazil, Chile, Israel, and the United Kingdom.

This large-scale experiment allows us to answer a comprehensive set of questions with regard to teacher performance pay including: (i) Can teacher performance pay based on test scores improve student achievement? (ii) What, if any, are the negative consequences of teacher incentives based on student test scores? (iii) How do school-level group incentives compare with teacher-level individual incentives? (iv) How does teacher behavior change in response to performance pay? and (v) How cost effective are teacher incentives relative to other uses for the same money?

We find that the teacher performance pay program was effective in improving student learning. At the end of two years of the program, students in incentive schools performed significantly better than those in comparison schools by 0.27 and 0.17 standard deviations (SD) in math and language tests respectively. The mean treatment effect of 0.22 SD is equal to 9 percentile points at the median of a normal distribution. We find a minimum average treatment effect of 0.1 SD at every percentile of baseline test scores, suggesting broad-based gains in test scores as a result of the incentive program.

We find no evidence of any adverse consequences as a result of the incentive programs. Students in incentive schools do significantly better not only in math and language (for which there were incentives), but also in science and social studies (for which there were no incentives), suggesting positive spillover effects. There was no difference in student attrition between incentive and control schools, and no evidence of any adverse gaming of the incentive program by teachers.

School-level group incentives and teacher-level individual incentives perform equally well in the first year, but the individual incentive schools outperformed the group incentive schools after two years of the program. At the end of two years, the average treatment effect was 0.28 SD in the individual incentive schools compared to 0.15 SD in the group incentive schools, with this difference being significant at the 10% level.

We measure changes in teacher behavior in response to the program with both teacher interviews as well as direct physical observation of teacher activity. Our results suggest that the main mechanism for the impact of the incentive program was not increased teacher attendance, but greater (and more effective) teaching effort conditional on being present.

We find that performance-based bonus payments to teachers were a significantly more cost effective way of increasing student test scores compared to spending a similar amount of money

unconditionally on additional schooling inputs. In a parallel initiative, two other sets of 100 randomly-chosen schools were provided with an extra contract teacher, and with a cash grant for school materials respectively. At the end of two years, students in schools receiving the input programs scored 0.08 SD higher than those in comparison schools. However, the incentive programs had a significantly larger impact on learning outcomes (0.22 versus 0.09 SD) over the same period, even though the total cost of the bonuses was around 25% lower than the amount spent on the inputs.

Our results contribute to a growing literature on the effectiveness of performance-based pay for teachers.[2] The best identified studies outside the US on the effect of paying teachers on the basis of student test outcomes are Lavy (2002) and (2009), and Glewwe, Ilias, and Kremer (2010), but their evidence is mixed. Lavy uses a combination of regression discontinuity, difference in differences, and matching methods to show that both group and individual incentives for high school teachers in Israel led to improvements in student outcomes (in the 2002 and 2009 papers respectively). Glewwe et al (2010) report results from a randomized evaluation that provided primary school teachers (grades 4 to 8) in Kenya with group incentives based on test scores and find that, while test scores went up in program schools in the short run, the students did not retain the gains after the incentive program ended. They interpret these results as being consistent with teachers expending effort towards short-term increases in test scores but not towards long-term learning.[3] Two recent experimental evaluations of performance pay in the US both reported no effect of performance-based pay for teachers on student learning outcomes (Goodman and Turner (2010) in New York, and Springer et al. (2010) in Tennessee).

There are several unique features in the design of the field experiment presented in this paper. We conduct the first randomized evaluation of teacher performance pay in a representative sample of schools.[4] We take incentive theory seriously and design the incentive

[2] Previous studies include Ladd (1999) in Dallas, Atkinson et al (2009) in the UK, and Figlio and Kenny (2007) who use cross-sectional data across multiple US states. See Umansky (2005) and Podgursky and Springer (2007) for reviews on teacher performance pay and incentives. The term "teacher incentives" is used very broadly in the literature. We use the term to refer to financial bonus payments on the basis of student test scores.

[3] It is worth nothing though that evidence from several contexts and interventions suggests that the effect of almost *all* education interventions appear to decay when the programs are discontinued (see Jacob et al. 2008, and Andrabi et al. 2009), and so this inference should be qualified.

[4] The random assignment of treatment provides high internal validity, while the random sampling of schools into the universe of the study provides greater external validity than typical experiments by avoiding the "randomization

program to reward gains at all points in the student achievement distribution, and to minimize the risk of perverse outcomes. The study design also allows us to test for a wide range of possible negative outcomes. We study group (school-level) and individual (teacher-level) incentives in the same field experiment. We measure changes in teacher behavior with both direct observations and with teacher interviews. Finally, we study both input and incentive based policies in the same field experiment to enable a direct comparison of their effectiveness.

While set in the context of schools and teachers, this paper also contributes to the broader literature on performance pay in organizations in general and public organizations in particular.[5] True experiments in compensation structure with contemporaneous control groups are rare (Bandiera et al. (2007) is a recent exception) and our results may be relevant to answering broader questions regarding performance pay in organizations.

The rest of this paper is organized as follows: section 2 provides a theoretical framework for thinking about teacher incentives. Section 3 describes the experimental design and the treatments, while section 4 discusses the test design. Sections 5 and 6 present results on the impact of the incentive programs on test score outcomes and teacher behavior. Section 7 discusses the cost effectiveness of the performance-pay programs. Section 8 concludes.

2. Theoretical Framework

2.1 Multi-task moral hazard

While basic incentive theory suggests that teacher incentives on the basis of improved test scores should have a positive impact on test scores, multi-tasking theory cautions that such incentives may increase the likelihood of undesired outcomes (Holmstrom and Milgrom 1991; Baker 1992; Baker 2002). The challenge of optimal compensation design in the presence of multi-tasking is illustrated by a simple model (based on Baker 2002 and Neal 2010).

Suppose teachers (agents) engage in two types of tasks in the classroom, T_1 and T_2, where T_1 represents teaching using curricular best practices, and T_2 represents activities designed to increase scores on exams (such as drilling, coaching on items likely to be on the test, and perhaps

bias", whereby entities that are in the experiment are atypical relative to the population that the result is sought to be extrapolated to (Heckman and Smith 1995).

[5] See Gibbons (1998) and Prendergast (1999) for general overviews of the theory and empirics of incentives in organizations. Dixit (2002) provides a discussion of these themes as they apply to public organizations.

even cheating). Let t_1 and t_2 represent the time spent on these two types of tasks and let the technology of human capital production (in gains) be given by:

$$H = f_1 t_1 + f_2 t_2 + \varepsilon$$

where f_1 and f_2 are the marginal products of time spent on T_1 and T_2 on human capital production, and ε is random noise in *H* representing all factors outside the teacher's control that also influence *H*. The social planner (principal) cannot observe any of *H*, t_1 or t_2 but can only observe an imperfect performance measure *P* (such as test scores) that is given by:

$$P = g_1 t_1 + g_2 t_2 + \phi$$

where g_1 and g_2 are the marginal products of time spent on T_1 and T_2 on test scores, and ϕ is random noise in *P* outside the teacher's control. The principal offers a wage contract as a function of *P*, such as: $w = s + b \cdot P$, where *w* is the total wage, *s* is the salary, and *b* is the bonus rate paid per unit of *P*. The teacher's utility function is given by:

$$U = E(w) - C(t_1, t_2)$$

where $E(w)$ is the expected wage (we abstract away from risk aversion to focus on multi-tasking), and $C(t_1, t_2; \bar{t})$ is the cost associated with any combination of t_1 and t_2. Here, we follow Holmstrom and Milgrom (1991) in allowing the cost of effort to depend on an effort norm, $\bar{t}$. Teachers may suffer psychic costs if their total effort levels fall below this norm (i.e. $t_1 + t_2 < \bar{t}$). The optimal bonus rate, b^*, depends on the functional form of this cost function, but if t_1 and t_2 are substitutes, it is easy to construct cases (typically when $f_1 > f_2$ and $g_2 > g_1$ as is believed to be the case by most education experts) where the optimal contract involves no incentive pay ($b^* = 0$). In these scenarios, it is optimal for the social planner to simply accept the output generated by the norm $\bar{t}$ because incentive provision *can reduce* human capital accumulation by causing teachers to reduce t_1 and increase t_2.

However, Neal (2010) notes that even when t_1 and t_2 are substitutes, the introduction of incentive pay may well be welfare improving in environments where $\bar{t}$ is small. When $\bar{t}$ is small, the gains from increasing *total effort* are more likely to exceed the costs from distorting the allocation of effort between t_1 and t_2. In addition, it is clear that incentive pay is more attractive

when f_1/f_2 is not much greater than one because, in these cases, substitution from t_1 to t_2 is less costly.

There is evidence to suggest that $\bar{t}$ may be quite low in India. A study using a nationally representative dataset of primary schools in India found that 25% of teachers were absent on any given day, and that less than half of them were engaged in any teaching activity (Kremer et al. 2005). There are also reasons to believe that f_1/f_2 may be close to one in India. The centrality of exam preparation in Indian and other Asian education systems may mean that the 'best practices' in the education system may not be very different from teaching practices meant to increase test scores. There is also evidence to suggest that the act of frequent test taking can increase comprehension and retention even of non-tested materials (Chan et al. 2006).

So, it is possible that setting $b > 0$ will not only increase test scores (*P*), but also increase underlying human capital of students (*H*), especially in contexts such as India for the reasons mentioned above. Whether or not this is true is an empirical question and is the focus of our research design and empirical analysis (sections 4 and 5).

2.2 Group versus Individual Incentives

The theoretical prediction of the relative effectiveness of individual and group teacher incentives is ambiguous. Group (school-level) incentives could induce free riding and thus normally be lower-powered than individual (teacher-level) incentives (Holmstrom 1982). However, social norms and peer monitoring (which may be feasible in the small groups of teachers in our setting) may enable community enforcement of the first-best level of effort, in which case, the costs of free-riding may be mitigated or eliminated (Kandori 1992; Kandel and Lazear 1992). Finally, if there are gains to cooperation or complementarities in production, then it is possible that group incentives might yield better results than individual incentives (Itoh 1991; Hamilton, Nickerson, and Owan 2003). The relative effectiveness of group and individual teacher performance pay is therefore an empirical question, and we study both types of incentives in the same field experiment over two full academic years.

3. Experimental Design

3.1 Context

While India has made substantial progress in improving access to primary schooling and primary school enrolment rates, the average levels of learning remain very low. The most recent *Annual Status of Education Report* found that nearly 60% of children aged 6 to 14 in an all-India survey of rural households could not read at the second grade level, though over 95% of them were enrolled in school (Pratham, 2010). Public spending on education has been rising as part of the "Education for All" campaign, but there are substantial inefficiencies in public delivery of education services. As mentioned earlier, a study using a representative sample of Indian schools found that 25% of teachers were absent on any given day, and that less than half of them were engaged in any teaching activity (Kremer et al. 2005).

Andhra Pradesh (AP) is the 5th most populous state in India, with a population of over 80 million (70% rural). AP is close to the all-India average on measures of human development such as gross enrollment in primary school, literacy, and infant mortality, as well as on measures of service delivery such as teacher absence (Figure 1a). The state consists of three historically distinct socio-cultural regions and a total of 23 districts (Figure 1b). Each district is divided into three to five divisions, and each division is composed of ten to fifteen mandals, which are the lowest administrative tier of the government of AP. A typical mandal has around 25 villages and 40 to 60 government primary schools.

The average rural primary school is quite small, with total enrollment of around 80 students and an average of 3 teachers across grades one through five. One teacher typically teaches all subjects for a given grade (and often teaches more than one grade simultaneously). All regular teachers are employed by the state, and their salary is mostly determined by experience and rank, with minor adjustments based on assignment location, but no component based on any measure of performance. The average salary of regular teachers at the time of the study was around Rs. 8,000/month and total compensation including benefits was over to Rs. 10,000/month (per capita income in AP is around Rs. 2,000/month; 1 US Dollar ≈ 45 Indian Rupees (Rs.)). Teacher unions are strong and disciplinary action for non-performance is rare.[6]

[6] Kremer et al (2005) find that 25% of teachers are absent across India, but only 1 head teacher in their sample of 3000 government schools had ever fired a teacher for repeated absence. See Kingdon and Muzammil (2001) for an illustrative case study of the power of teacher unions in India.

3.2 Sampling

We sampled 5 districts across each of the 3 socio-cultural regions of AP in proportion to population (Figure 1b).[7] In each of the 5 districts, we randomly selected one division and then randomly sampled 10 mandals in the selected division. In each of the 50 mandals, we randomly sampled 10 schools using probability proportional to enrollment. Thus, the universe of 500 schools in the study was representative of the schooling conditions of the typical child attending a government-run primary school in rural AP.

3.3 Design Overview

The performance-pay experiments were conducted as part of a larger research project implemented by the Azim Premji Foundation (APF) to evaluate the impact of policy options to improve the quality of primary education in AP. Four interventions were studied, with two being based on providing schools with additional inputs (an extra contract teacher, and a cash block grant), and two being based on providing schools and teachers with incentives for better performance (group and individual bonus programs for teachers based on student performance).

The overall design of the project is represented in the table below:

Table 3.1

	INCENTIVES (Conditional on Improvement in Student Learning			
INPUTS (Unconditional)		NONE	GROUP BONUS	INDIVIDUAL BONUS
	NONE	CONTROL (100 Schools)	100 Schools	100 Schools
	EXTRA CONTRACT TEACHER	100 Schools		
	EXTRA BLOCK GRANT	100 Schools		

As Table 3.1 shows, the input treatments (see section 7) were provided *unconditionally* to the selected schools at the beginning of the school year, while the incentive treatments consisted of an announcement that bonuses would be paid at the beginning of the next school year *conditional* on average improvements in test scores during the current school year. No school received more than one treatment, which allows the treatments to be analyzed independent of each other. The school year in AP starts in the middle of June, and the baseline tests were

[7] The districts were chosen so that districts within a region would be contiguous for logistical reasons.

conducted in the 500 sampled schools during late June and early July, 2005.[8] After the baseline tests were scored, 2 out of the 10 project schools in each mandal were randomly allocated to each of 5 cells (four treatments and one control). Since 50 mandals were chosen across 5 districts, there were a total of 100 schools (spread out across the state) in each cell (Table 3.1). The geographic stratification implies that every mandal was an exact microcosm of the overall study, which allows us to estimate the treatment impact with mandal-level fixed effects and thereby net out any common factors at the lowest administrative level of government.

Table 1 (Panel A) shows summary statistics of baseline school characteristics and student performance variables by treatment (control schools are also referred to as a 'treatment' for expositional ease). Column 4 provides the p-value of the joint test of equality, showing that the null of equality across treatment groups cannot be rejected for any of the variables.[9]

After the randomization, program staff from the Foundation personally went to each of the schools in the first week of August 2005 to provide them with student, class, and school performance reports, and with oral and written communication about the intervention that the school was receiving. They also made several rounds of unannounced tracking surveys to each of the schools during the school year to collect data on process variables including student attendance, teacher attendance and activity, and classroom observation of teaching processes.[10] All schools operated under identical conditions of information and monitoring and only differed in the treatment that they received. This ensures that Hawthorne effects are minimized and that a comparison between treatment and control schools can accurately isolate the treatment effect.[11]

End of year assessments were conducted in March and April, 2006 in all project schools. The results were provided to the schools in the beginning of the next school year (July – August,

[8] The selected schools were informed by the government that an external assessment of learning would take place in this period, but there was no communication to any school about any of the treatments at this time (since that could have led to gaming of the baseline test).

[9] Table 1 shows sample balance across control, group incentive, and individual incentive schools, which are the focus of the analysis in this paper. The randomization was done jointly across all 5 treatments shown in Table 3.1, and the sample was also balanced on observables across the other treatments.

[10] Six visits were made per school in the first year (05 – 06), while four were made in the second year (06 – 07)

[11] An independent question of interest is that of the impact on teacher behavior and learning outcomes of the diagnostic feedback reports and low-stakes monitoring that were provided to all schools (including the control schools). We study this by comparing the 'control' schools in this paper with another 'pure control' group that did not receive any of the baseline test, feedback reports, or regular low-stakes monitoring and find that there was no impact of low stakes measurement and monitoring on test scores (see Muralidharan and Sundararaman 2010a).

2006), and all schools were informed that the program would continue for another year.[12] Bonus checks based on first year performance were sent to qualifying teachers by the end of September 2006, following which the same processes were repeated for a second year.

3.4 Description of Incentive Treatments

Teachers in incentive schools were offered bonus payments on the basis of the average improvement in test scores (in math and language) of students taught by them subject to a minimum improvement of 5%. The bonus formula was:

Bonus = Rs. 500 * (% Gain in average test scores – 5%) if Gain > 5%

= 0 otherwise

All teachers in group incentive schools received the same bonus based on average school-level improvement in test scores, while the bonus for teachers in individual incentive schools was based on the average test score improvement of students taught by the specific teacher. [13] We use a (piecewise) linear formula for the bonus contract, both for ease of communication and implementation and also because it is the most resistant to gaming across periods (the end of year score in the first year determined the target score for the subsequent year).[14]

The 'slope' of Rs. 500 per percentage point gain in average scores was set so that the expected incentive payment per school would be approximately equal to the additional spending in the input treatments (based on calibrations from the project pilot).[15] The threshold of 5% average improvement was introduced to account for the fact that the baseline tests were in June/July and the end of year tests would be in March/April, and so the baseline score might be

[12] The communication to teachers with respect to the length of the program was that the program would continue as long as the government continued to support the project. The expectation conveyed to teachers during the first year was that the program was likely to continue but was not guaranteed to do so.

[13] 1st grade students were not tested in the baseline, and so their 'target' score for a bonus (above which the linear schedule above would apply) was set to be the mean baseline score of the 2nd grade students in the school. The target for the 2nd grade students was equal to their baseline score plus the 5% threshold described above. Schools selected for the incentive programs were given detailed letters and verbal communications explaining the incentive formula. Sample communication letters are available from the authors on request.

[14] Holmstrom and Milgrom (1987) show the theoretical optimality of linear contracts in a dynamic setting (under assumptions of exponential utility for the agent and normally distributed noise). Oyer (1998) provides empirical evidence of gaming in response to non-linear incentive schemes.

[15] The best way to set expected incentive payments to be exactly equal to Rs. 10,000/school would have been to run a tournament with pre-determined prize amounts. Our main reason for using a contract as opposed to a tournament was that contracts were more transparent to the schools in our experiment since the universe of eligible schools was spread out across the state. Individual contracts (without relative performance measurement) also dominate tournaments for risk-averse agents when specific shocks (at the school or class level) are more salient for the outcome measure than aggregate shocks (across all schools), which is probably the case here (see Kane and Staiger, 2002). See Lazear and Rosen (1981) and Green and Stokey (1983) for a discussion of tournaments and when they dominate contracts.

artificially low due to students forgetting material over the summer vacation. There was no minimum threshold in the second year of the program because the first year's end of year score was used as the second year's baseline and the testing was conducted at the same time of the school year on a 12-month cycle.[16]

The bonus formula was designed to minimize potentially undesirable 'threshold' effects, where teachers only focus on students near a performance target, by making the bonus payment a function of the average improvement of *all* students.[17] If the function transforming teacher effort into test-score gains is concave (convex) in the baseline score, teachers would have an incentive to focus on weaker (stronger) students, but no student is likely to be wholly neglected since each contributes to the class average. In order to discourage teachers from excluding students with weak gains from taking the end of year test, we assigned a zero improvement score to any child who took the baseline test but not the end of year test.[18] To make cheating as difficult as possible, the tests were conducted by external teams of 5 evaluators in each school (1 for each grade), the identity of the students taking the test was verified, and the grading was done at a supervised central location at the end of each day's testing.

4. Test Design

4.1 Test Construction and Normalization

We engaged India's leading education testing firm, "Educational Initiatives", to design the tests to our specifications. The baseline test (June-July 2005) tested math and language (Telugu)

[16] The convexity in reward schedule in the first year due to the threshold could have induced some gaming, but the distribution of mean class and school-level gains at the end of the first year of the program did not have a gap below the threshold of 5%. If there is no penalty for a reduction in scores, there is convexity in the payment schedule even if there is no threshold (at a gain of zero). To reduce the incentives for gaming in subsequent years, we use the higher of the baseline and year end scores as the target for the next year and so a school/class whose performance deteriorates does *not* have its target reduced for the next year.

[17] Many of the negative consequences of incentives discussed in Jacob (2005) are a response to the threshold effects created by the targets in the program he studied. Neal and Schanzenbach (2010) discuss the impact of threshold effects in the No Child Left Behind act on teacher behavior and show that teachers do in fact focus more on students on the 'bubble' and relatively neglect students far above or below the thresholds. We anticipated this concern and designed the incentive schedule accordingly.

[18] In the second year (when there was no threshold), students who took the test at the end of year 1 but not at the end of year 2 were assigned a score of -5. Thus, the cost of a dropping out student to the teacher was always equal to a negative 5% score for the student concerned. A higher penalty would have been difficult since most cases of attrition are out of the teacher's control. The penalty of 5% was judged to be adequate to avoid explicit gaming of the test taking population. We also cap negative gains at the student-level at -5% for the calculation of teacher bonuses. Thus, putting a floor on the extent to which a poor performing student brought down the class/school average at -5% ensured that a teacher/school could never do worse than having a student drop out to eliminate any incentive to get weak students to not appear for the test.

and covered competencies up to that of the previous school year. At the end of the school year (March-April, 2006), schools had two rounds of tests in each subject with a gap of two weeks between the rounds. The first test (referred to as the "lower end line" or LEL) covered competencies up to that of the previous school year, while the second test (referred to as the "higher end line" or HEL) covered materials from the current school year's syllabus. The same procedure was repeated at the end of the second year. Doing two rounds of testing at the end of each year allows for the inclusion of more materials across years of testing, reduces the impact of measurement errors specific to the day of the test, and also reduces sample attrition due to student absence on the day of the test.

For the rest of this paper, Year 0 (Y0) refers to the baseline tests in June-July 2005; Year 1 (Y1) refers to both rounds of tests conducted at the end of the first year of the program in March-April, 2006; and Year 2 (Y2) refers to both rounds of tests conducted at the end of the second year of the program in March-April, 2007. Scores in Y0 are normalized relative to the distribution of scores across all schools for the same test (pre-treatment), while scores in Y1 and Y2 are normalized with respect to the score distribution in the control schools for the same test.[19]

4.2 Use of repeat and multiple-choice questions

At the student-level, there were *no* identically repeated questions between Y0 and Y1. Between Y2 and Y1, 6% of questions were repeated in math (12 out of 205) and 1.5% in language (3 out of 201). At the school-level, 13% and 18% of questions were repeated in Y1 and Y2 in math and 14% and 10% in Y1 and Y2 in language.[20] The fraction of multiple-choice questions on any given test ranged from 22 to 28% in math, and 32 to 43% in language.

4.3 Basic versus higher-order skills

To distinguish between rote and conceptual learning, we asked the test-design firm to design the tests to include both 'mechanical' and 'conceptual' questions within each skill category on the test. Specifically, a mechanical question was considered to be one that conformed to the format

[19] Student test scores on each round (LEL and HEL), which are conducted two weeks apart, are first normalized relative to the score distribution in the control schools on that test, and then averaged across the 2 rounds to create the normalized test score for each student at each point in time. So a student can be absent on one testing day and still be included in the analysis without bias because the included score would have been normalized relative to the distribution of all control school students on the same test that the student took.

[20] A student-level repeated question is one that the same student would have seen in a previous round of testing. A school-level repeated question is one that any student in any grade could have seen in a previous test (this is therefore a better representation of the set of questions that the teacher may have been able to coach the students on using previous exams for test practice).

of the standard exercises in the text book, whereas a conceptual one was defined as a question that tested the same underlying knowledge or skill in an unfamiliar way.[21]

4.4 Incentive versus non-incentive subjects

Another dimension on which incentives can induce distortions is on the margin between incentive and non-incentive subjects. We study the extent to which this is a problem by conducting additional tests at the end of each year in science and social studies on which there was no incentive.[22] Since these subjects are introduced only in grade 3 in the school curriculum, these additional tests were administered in grades 3 to 5.

5. Results

5.1 Teacher Turnover and Student Attrition

Regular civil-service teachers in AP are transferred once every three years on average. While this could potentially bias our results if more teachers chose to stay in or tried to transfer into the incentive schools, it is unlikely that this was the case since the treatments were announced in August '05, while the transfer process typically starts earlier in the year. There was no statistically significant difference between any of the treatment groups in the extent of teacher turnover or attrition, and the transfer rate was close to 33%, which is consistent with the rotation of teachers once every 3 years (Table 1 – Panel B, rows 11-12). As part of the agreement between the Government of AP and the Azim Premji Foundation, the Government agreed to minimize transfers into and out of the sample schools for the duration of the study. The average teacher turnover in the second year was only 5%, and once again, there was no significant difference in the two-year teacher attrition and turnover rates across the various treatments (Table 1 – Panel B, rows 13 - 14).

The average student attrition rate in the sample (defined as the fraction of students in the baseline tests who did not take a test at the end of each year) was 7.1% and 20.6% in year 1 and

[21] See the working paper version of this paper (Muralidharan and Sundararaman 2009) for more details and examples. The percentage split between mechanical and conceptual questions on the tests was roughly 70-30. Koretz (2002) points out that test score gains are only meaningful if they generalize from the specific test to other indicators of mastery of the domain in question. While there is no easy solution to this problem given the impracticality of assessing every domain beyond the test, our inclusion of both mechanical and conceptual questions in each test attempts to address this concern.

[22] In the first year of the project, schools were not told about these additional subject tests till a week prior to the tests and were told that these tests were only for research purposes. In the second year, the schools knew that these additional tests would be conducted, but also knew from the first year that these tests would not be included in the bonus calculations.

year 2 respectively, but there is no significant difference in attrition across the treatments (rows 17 and 20). Beyond confirming sample balance, this is an important result in its own right because one of the concerns of teacher incentives based on test scores is that weaker children might be induced to drop out of testing in incentive schools (Jacob 2005). Attrition is higher among students with lower baseline scores, but this is true across all treatments, and we find no significant difference in mean baseline test score across treatment categories among the students who drop out from the test-taking sample (Table 1 – Panel B, rows 18, 19, 21, 22).[23]

5.2 Specification

We first discuss the impact of the incentive program as a whole by pooling the group and individual incentive schools and considering this to be the 'incentive' treatment. All estimation and inference is done with the sample of 300 control and incentive schools unless stated otherwise. Our default specification uses the form:

$$T_{ijkm}(Y_n) = \alpha + \gamma \cdot T_{ijkm}(Y_0) + \delta \cdot Incentives + \beta \cdot Z_m + \varepsilon_k + \varepsilon_{jk} + \varepsilon_{ijk} \qquad (5.1)$$

The main dependent variable of interest is T_{ijkm}, which is the normalized test score on the specific subject, where *i, j, k, m* denote the student, grade, school, and mandal respectively. Y_0 indicates the baseline tests, while Y_n indicates a test at the end of *n* years of the program. Including the normalized baseline test score improves efficiency due to the autocorrelation between test-scores across multiple periods.[24] All regressions include a set of mandal-level dummies (Z_m) and the standard errors are clustered at the school level. We also run the regressions with and without controls for household and school variables. The 'Incentives' variable is a dummy at the school level indicating treatment status, and the parameter of interest is δ, which is the effect on test scores of being in an incentive school. The random assignment of the incentive program ensures that this is an unbiased and consistent estimate of the one-year and two-year treatment effects.

5.3 Impact of Incentives on Test Scores

Averaging across both math and language, students in incentive schools scored 0.15 standard deviations (SD) higher than those in comparison schools at the end of the first year of the

[23] We estimate a model of student attrition using baseline scores and observable characteristics and cannot reject that the same model predicts attrition in both treatment and control schools. We also estimate treatment effects by re-weighting the sample by the inverse of the probability of continuing in the sample and the results are unchanged.
[24] Since grade 1 students did not have a baseline test, we set the normalized baseline score to zero for these students (similarly for students in grade 2 at the end of two years of the treatment).

program, and 0.22 SD higher at the end of the second year (Table 2 – Panel A, columns 1 and 3). The impact of the incentives at the end of two years is 0.27 SD in math and 0.17 SD in language (Panels B and C of Table 2). The addition of school and household controls does not significantly change the estimated value of δ in any of the regressions, confirming the validity of the randomization (columns 2 and 4).

We verify that teacher transfers do not affect the results by estimating equation (5.1) across different durations of teacher presence in the school, and there is no significant difference across these estimates. The testing process was externally proctored at all stages and we had no reason to believe that cheating was a problem in the first year, but there were two cases of cheating in the second year. The concerned schools/teachers were declared ineligible for bonuses, and both these cases were dropped from the analysis presented here.

5.4 Robustness of Treatment Effects

An important concern with interpreting these results is whether they represent real gains in learning or merely reflect drilling on past exams and better test-taking skills. We use question-level data to examine this issue further. We first break down the treatment effect by repeat and non-repeat questions. A question is classified as a repeat if it had appeared in any previous test in the project (for any grade and at any time).[25] Table 3 shows the percentage score obtained by students in control and incentive schools by repeat and non-repeat questions. We see that students in incentive schools score significantly higher on both repeat and non-repeat questions (rows 3 and 4). The incremental score on repeat questions is higher in the incentive schools, but this is not significantly higher than the extent to which they score higher on non-repeat questions suggesting that the main treatment effects are not being driven by improved student performance on repeated questions. We calculate the treatment effects estimated in Table 2 using only the non-repeat questions and find that the estimate is essentially unchanged.

We also break down the questions into multiple-choice and non multiple-choice questions, where performance on the former is more likely to be amenable to being improved by better test-taking skills. Table 4 presents a similar break down as Table 3 and we see that incentive schools do significantly better on both multiple-choice and free response questions, with no significant difference in performance across the two types of questions (in 5 of the 6 comparisons).

[25] This includes questions that appear in an LEL test for grade 'n' and then appear 2 weeks later in the HEL test for grade 'n-1'. The idea is to classify any question that a teacher could have seen before and drilled the students on as a 'repeat' question.

Finally, we also separately analyze student performance on both 'mechanical' and 'conceptual' parts of the test (as described in section 4.3) and find that incentive schools do significantly better on both the mechanical and conceptual components of the test, with no significant difference in improvement between the two types of questions (tables available on request).

5.5 Distribution of Treatment Effects

Figure 2 plots the quantile treatment effects of the performance pay program on student test scores (defined for each quantile τ as: $\delta(\tau) = G_n^{-1}(\tau) - F_m^{-1}(\tau)$ where G_n and F_m represent the empirical distributions of the treatment and control distributions with *n* and *m* observations respectively), with bootstrapped 95% confidence intervals, and shows that the quantile treatment effects are positive at every percentile and increasing. Note that this figure does *not* plot the treatment effect at different quantiles (since student rank order is not preserved between the baseline and end line tests even within the same treatment group). It simply plots the gap at each percentile of the treatment and control distributions after two years of the program and shows that test scores in incentive schools are higher at every percentile of the end line distribution, and that the program also increased the variance of test scores.

We next test for heterogeneity of the incentive treatment effect across baseline student, school, and teacher characteristics by testing if δ_3 is significantly different from zero in:

$$T_{ijkm}(Y_n) = \alpha + \gamma \cdot T_{ijkm}(Y_0) + \delta_1 \cdot Incentives + \delta_2 \cdot Characteristic$$
$$+ \delta_3 \cdot (Incentives \times Characteristic) + \beta \cdot Z_m + \varepsilon_k + \varepsilon_{jk} + \varepsilon_{ijk} \qquad (5.2)$$

Table 5 (Panel A) shows the results of these regressions on several school and household characteristics (each column in Table 5 represents one regression testing for heterogeneous treatment effects along the characteristic mentioned). We find very limited evidence of differential treatment effects by school characteristics such as total number of students, school infrastructure, or school proximity to facilities. We also find no evidence of a significant difference in the effect of the incentives by most of the student demographic variables, including an index of household literacy, the caste of the household, the student's gender, and the student's baseline score. The only evidence of heterogeneous treatment effects is across levels of family affluence, with students from more affluent families showing a better response to the teacher incentive program.

The lack of heterogeneous treatment effects by baseline score is an important indicator of broad-based gains since the baseline score is probably the best summary statistic of prior inputs into education. To see this more clearly, Figure 3 plots non-parametric treatment effects by percentile of baseline score,[26] and we see that there is a minimum treatment effect of 0.1 SD for students regardless of where they were in the initial test score distribution.

The lack of heterogeneous treatment effects by initial scores, suggests that the increase in variance of test scores in incentive schools (Figure 2) may be reflecting variance in teacher responsiveness to the incentive program, as opposed to variance in student responsiveness to the treatment by initial learning levels. We test this by estimating teacher value addition (measured as teacher fixed effects in a regression of current test scores on lagged scores) and plotting the difference in teacher fixed effects at each percentile of the control and treatment distributions. We find that both the mean and variance of teacher value-addition are significantly higher in the incentive schools (Figure 4).

Having established that there is variation in teacher responsiveness to the incentive program, we test for differential responsiveness by observable teacher characteristics (Table 5 – Panel B). We find that the interaction of teachers' education and training with incentives is positive and significant, while education and training by themselves are not significant predictors of value addition (columns 1-2). This suggests that teacher qualifications by themselves are not associated with better learning outcomes under the status quo, but that they could matter more if teachers had incentives to exert more effort (see Hanushek (2006)).

We also find that teachers with higher base pay as well as teachers with more experience respond less well to the incentives (columns 3-4). This suggests that the magnitude of the incentive mattered because the potential bonus (which was similar for all teachers) would have been a larger share of base pay for lower paid teachers. However, teachers with higher base pay are also more experienced, and so we cannot distinguish the impact of the incentive amount from that of other teacher characteristics that influence base pay.[27]

[26] The figure plots a kernel-weighted local polynomial regression of end line scores (after 2 years) on the percentile of baseline score separately for the incentive and control schools, and also plots the difference at each percentile of baseline scores. The confidence intervals of the treatment effects are constructed by drawing 1000 bootstrap samples of data that preserve the within school correlation structure in the original data, and plotting the 95% range for the treatment effect at each percentile of baseline scores.

[27] Of course, this is a caution that applies to any interpretation of interactions in an experiment, since the covariate is not randomly assigned and could be correlated with other omitted variables.

5.6 Impact on Non-Incentive Subjects

The impact of incentives on the performance in non-incentive subjects such as science and social studies is tested using a slightly modified version of specification (5.1) where lagged scores on both math and language are included to control for initial learning levels. We find that students in incentive schools also performed significantly better on non-incentive subjects at the end of each year of the program, scoring 0.11 and 0.18 SD higher than students in control schools in science and social studies at the end of two years of the program (Table 6 – Panel A). These results suggest that, in the context of primary education in a developing country with very low levels of learning, teacher efforts aimed at increasing test scores in math and language may also contribute to superior performance on non-incentive subjects suggesting complementarities among the measures and positive spillover effects between them.

We probe the possibility of spillovers further as follows: for each student we generate a *predicted* math and language score at each point in time as well as the *residual* test score formed by taking the difference between the actual score and the predicted score (this residual is therefore an estimate of the 'innovation' in learning that took place over the school year – and in light of Table 2 would be larger for students in incentive schools). We then run regressions of science and social studies scores on the predicted math and language scores, the residuals as defined above, a dummy for treatment status, and interactions of the residuals and treatment status and present the results in Table 6 (Panel B).

There are three noteworthy results here: (a) the coefficients on the residuals are highly significant with the coefficient on the language residual typically being larger than on the math residual, (b) the coefficient on the 'incentive' treatment dummy is close to zero, and (c) the interaction terms are mostly insignificant. In turn, these suggest that (a) improvements in language were more relevant for improved performance in other subjects – especially social studies, (b) the mechanism for the improved performance in science and social studies in the incentive schools was the improved performance in math and language – since the treatment dummy is close to zero after including the residuals, and (c) an innovation in math or language did not typically have a differential impact in incentive schools. Taken together, these results suggest that incentive schools did not do anything different with respect to non-incentive subjects, but that positive spillovers from improvements in math and especially language led to improved scores in non-incentive subjects as well.

5.7 Group versus Individual Incentives

Both the group and the individual incentive programs had significantly positive treatment effects at the end of each year of the program (Table 7, columns 1 and 4). In the first year of the program, students in individual incentive schools performed slightly better than those in group incentive schools, but the difference was not significant. By the end of the second year, students in individual incentive schools scored 0.28 SD higher than those in comparison schools, while those in group incentive schools scored 0.15 SD higher, with this difference being significant at the 10% level (column 4).

We find no significant impact of the number of teachers in the school on the relative performance of group and individual incentives (both linear and quadratic interactions of school size with the group incentive treatment are insignificant). However, the variation in school size is small with 92% of group incentive schools having between two and five teachers. The limited range of school size makes it difficult to precisely estimate the impact of group size on the relative effectiveness of group incentives. We repeat all the analysis presented above (in sections 5.3 – 5.6) treating group and individual incentive schools separately and find that the individual incentive schools always outperform the group incentive schools though the difference in point estimates is not always significant (tables available on request).

6. Teacher Behavior and Classroom Processes

We measure changes in teacher behavior in response to the incentive program with both direct observation as well as teacher interviews. As described in section 3.3, enumerators conducted several rounds of unannounced tracking surveys during the two school years across all schools in the project. To code classroom processes, an enumerator typically spent between 20 and 30 minutes at the back of a classroom (during each visit) without disturbing the class and coded whether specific actions took place during the period of observation. In addition to these observations, they also interviewed teachers about their teaching practices and methods, asking identical sets of questions in both incentive and control schools. These interviews were conducted in August 2006, around 4 months after the end of year tests, but before any results were announced, and a similar set of interviews was conducted in August 2007 after the second full year of the program.

There was no difference in either student or teacher attendance between control and incentive schools. We also find no significant difference between incentive and control schools on any of the various indicators of classroom processes as measured by direct observation.[28] This is similar to the results in Glewwe et al (2010) who find no difference in either teacher attendance or measures of teacher activity between treatment and control schools from similar surveys and raises the question of how the outcomes are significantly different when there don't appear to be any differences in observed processes between the schools.

The teacher interviews provide another way of testing for differences in behavior. Teachers in both incentive and control schools were asked *unprompted* questions about what they did differently during the school year at the end of each school year, but before they knew the results of their students. The interviews indicate that teachers in incentive schools are significantly more likely to have assigned more homework and class work, conducted extra classes beyond regular school hours, given practice tests, and paid special attention to weaker children (Table 8). While self-reported measures of teacher activity might be considered less credible than observations, we find a positive (and mostly significant) correlation between the reported activities of teachers and the performance of their students (Table 8 – column 4) suggesting that these self-reports were credible (especially since less than 50% of teachers in the incentive schools report doing *any* of the activities described in Table 8).

The interview responses suggest reasons for why salient dimensions of changes in teacher behavior might not have been captured in the classroom observations. An enumerator sitting in classrooms during the school day is unlikely to observe the extra classes conducted after school. Similarly, if the increase in practice tests occurred closer to the end of the school year (in March), this would not have been picked up by the tracking surveys conducted between September and February. Finally, while our survey instruments recorded if various activities took place, they did not have a way to capture the intensity of teacher efforts, which may be an important channel of impact.

One way to see this is to notice that there is no difference between treatment and control schools in the fraction of teachers coded as "actively teaching" when observed by the enumerator (Table 8 – row 2), but the interaction of "active teaching" and being in an incentive school is

[28] These include measures of teacher activity such as using the blackboard, reading from the textbook, asking questions to students, encouraging classroom participation, assigning homework, helping students individually, and measures of student activity such as using textbooks, and asking questions.

significantly positively correlated with measures of teacher value addition (Table 5B – column 7). This suggests that teachers changed the effectiveness of their teaching in response to the incentives in ways that would not be easily captured even by observing the teacher. In summary, it appears that the incentive program based on end of year test scores did not change the teachers' cost-benefit calculations on the attendance margin during the school year, but that it probably made them exert more effort when present.[29]

7. Comparison with Input Treatments & Cost-Benefit Analysis

As mentioned earlier, a parallel component of this study provided two other sets of 100 randomly-chosen schools with an extra contract teacher, and with a cash block grant for school materials respectively.[30] These interventions were calibrated so that the expected spending on the input and the incentive programs was roughly equal. To compare the effects across treatment types, we pool the 2 incentive treatments, the 2 input treatments, and the control schools and run the regression:

$$T_{ijkm}(Y_n) = \alpha + \gamma \cdot T_{ijkm}(Y_0) + \delta_1 \cdot Incentives + \delta_2 \cdot Inputs + \beta \cdot Z_m + \varepsilon_k + \varepsilon_{jk} + \varepsilon_{ijk} \qquad (7.1)$$

using the full sample of 500 schools. Both categories of treatments had a positive and significant impact on learning outcomes, but at the end of two years, the incentive schools scored 0.13 SD higher than the input schools and the difference is highly significant (Table 9 – Column 4). The incentive schools perform better than input schools in both math and language and both these differences are significant at the end of two years.

The total amount spent on each intervention was calibrated to be roughly equal, but the group incentive program ended up spending a lower amount per school. The average annual spending

[29] Duflo et al (2010) provide experimental estimates of the impact of teacher attendance on student learning in the Indian state of Rajasthan and estimate the effect on student learning to be roughly 0.1 SD for every 10 percentage point reduction in teacher absence. If we use this as a benchmark and assume that (a) the unit of 1 SD is comparable in their sample and ours and (b) the effects are linear over the relevant ranges of absence, then our treatment effect of 0.11 SD per year would require an increase in teacher attendance (at status quo levels of effort) of 11 percentage points. So we could interpret our results in terms of teacher attendance and argue that the increase in intensity of effort was equivalent to reducing teacher absence by over 40% from 25 to 14 percentage points.

[30] See our companion paper (Muralidharan and Sundararaman 2010b) for more details on the contract teacher program and its impact on student learning. We discuss the block grant intervention in Das et al. (2010). These input programs represented 2 out of the 3 most common input-based interventions (infrastructure, teachers, and materials). We did not conduct a randomized evaluation of infrastructure both due to practical difficulties, and because the returns would have to be evaluated over the depreciation life cycle of the infrastructure. Thus, the set of interventions studied here all represent "flow" expenditures that would be incurred annually and are therefore comparable to the "flow" spending on a teacher incentive program.

on each of the input treatments was Rs. 10,000/school, while the group and individual incentives programs cost roughly Rs. 6,000/school and Rs.10,000/school respectively.[31] Both the incentive programs were more cost effective than the input programs. The individual incentive program spent the same amount per school as the input programs but produced gains in test scores that were three times larger than those in the input schools (0.28 SD vs. 0.09 SD). The group incentive program had a smaller treatment effect than the individual incentive program (0.15 SD vs 0.27 SD), but was equally cost effective because smaller bonuses were paid.

A different way of thinking about the cost of the incentive program is to not consider the incentive payments as a cost at all, because it is simply a way of reallocating salary spending. For instance, if salaries were increased by 3% every year for inflation, then it might be possible to introduce a performance-based component with an expected payout of 3% of base pay in lieu of a standard increase across the board (using the formulation in Section 2, an increase in *b* could be offset by a reduction in *s*, without violating the participation constraint). Under this scenario, the 'incentive cost' would only be the risk premium needed to keep expected utility constant compared to the guaranteed increase of 3%. This is a very small number with an upper bound of 0.1% of base pay if teachers' coefficient of absolute risk aversion (CARA) is 2 and 0.22% of base pay even if the CARA is as high 5.[32] Finally, if performance-pay programs are designed on the basis of multiple years of performance, differences in compensation across teachers would be less due to random variation, and more due to heterogeneity in ability. This will not only reduce the risk of performance pay but could also attract higher-ability teachers into the profession, and reduce the rents paid to less effective teachers (see Muralidharan and Sundararaman 2011).

A full discussion of cost effectiveness should include an estimate of the cost of administering the program. The main cost outside the incentive payments is that of independently

[31] The bonus payment in the group incentive schools was lower than that in the individual incentive schools both because the treatment effect was smaller and also because classes with scores below their target brought down the average school gain in the group incentive schools, while teachers with negative gains (relative to targets) did not hurt teachers with positive gains in the individual incentive schools. So, even conditional on the same distribution of scores, the individual incentive payout would be higher as long as there are some classes with negative gains relative to the target because of truncation of teacher-level bonuses at zero in the individual incentive calculations.

[32] The risk premium here is the value of ε such that $0.5[u(0.97w+\varepsilon)+u(1.03w+\varepsilon)]=u(w)$, and is easily estimated for various values of CARA using a Taylor expansion around *w*. This is a conservative upper bound since the incentive program is modeled as an even lottery between the extreme outcomes of a bonus of 0% and 6%. In practice, the support of the incentive distribution would be non-zero everywhere on [0, 6] and the risk premium would be considerably lower.

administering and grading the tests. The approximate cost of each annual round of testing was Rs. 5,000 per school, which includes the cost of two rounds of independent testing and data entry but not the additional costs borne for research purposes. The incentive program would be more cost effective than the input programs even after adding these costs and even more so if we take the long-run view that the fiscal cost of performance pay can be lower than the amount of the bonus, if implemented in lieu of a scheduled across the board increase in pay.

Finally, we attempt a more speculative back of the envelope estimate of the absolute rate of return of the program by looking at the labor-market returns to improved test scores. Recent cross-sectional estimates of the returns to cognitive achievement in India suggest returns of 16% for scoring one SD higher on a standardized math test and 20% for scoring one SD higher on a standardized language test (Aslam et al. 2011). Assuming that the test score gains in this program correspond to a similar long-term difference in human capital accumulation,[33] the two year treatment effect would correspond to a 7.7% increase in wages (0.27 SD x 0.16 + 0.17 SD x 0.20). Depending on assumptions on rate of wage growth and discount rates, we obtain estimates of an internal rate of return ranging from 1600% to 18500% (or a return ranging from 16 to 185 times the initial cost).[34] These estimates are large enough that even if the estimates on the labor market returns to test scores were to be substantially lower, or the program costs much higher, the program would still have a very high rate of return. An important reason for this is that the cost of the incentive program was very low and combining estimates from our companion papers suggests that the performance pay program would be ten times more cost effective than reducing class size by hiring another civil-service teacher.[35] Thus, the optimal

[33] Chetty et al. (2010) show that there were significant long-term benefits to the class-size reductions under the Tennessee STAR program even though the test score gains faded away a few years into the program. Deming (2009) shows similar long-term gains to Head Start, though the test score gains fade away here as well. Of course, these studies are only suggestive about the long-term effects of programs that produce test-score gains, because there is no precise measure of the extent to which test-score gains in school translate into higher long-term wages.

[34] The minimum wage for agricultural labor in AP is Rs. 112/day. Assuming 250 working days/year yields an annual income of Rs. 28,000 and a 7.7% increase in wage would translate into additional income of Rs. 2,156/year. We treat this as a 40-year stream of *fixed* additional earnings (which is *very conservative* since we don't assume wage growth) and discount at 10% a year to obtain a present value of Rs. 21,235 per student at the time of entering the labor market. Since the average student in our project is 8 years old, we assume that they will enter the labor market at age 20 and further discount the present value by 10% annually for another 12 years to obtain a present value of Rs. 6,750/student. The average school had 65 students who took the tests, which provides an estimate of the total present value of Rs. 438,750. The cost of the program per school for two years was Rs. 27,500 (including both bonus and administrative costs), which provides an IRR estimate of 1600%. If we were to assume that wages would grow at the discount rate, the calculation yields an IRR estimate of 18500%.

[35] The performance pay intervention was twice as cost effective as providing schools with an extra contract teacher. We also find that the contract teacher was no less effective than a regular civil service teacher in spite of being paid

wage contract for teachers probably has a non-zero weight on student test score gains in this context.

8. Conclusion

Performance pay for teachers is an idea with strong proponents, as well as opponents, and the evidence to date on its effectiveness has been mixed. In this paper, we present evidence from a randomized evaluation of a teacher incentive program in a representative sample of government-run rural primary schools in the Indian state of Andhra Pradesh, and show that teacher performance pay led to significant improvements in student test scores, with no evidence of any adverse consequences of the program. Additional school inputs were also effective in raising test scores, but the teacher incentive programs were three times as cost effective.

The longer-term benefits to performance pay include not only greater teacher effort, but also potentially attracting better teachers into the profession (Lazear 2000, 2003; Hoxby and Leigh 2005). We find a positive and significant correlation between teachers' *ex ante* reported support for performance pay and their actual *ex post* performance (as measured by value addition). This suggests that effective teachers know who they are, and that teacher compensation systems that reward effectiveness may attract higher ability teachers (see Muralidharan and Sundararaman 2011 for further details on teacher opinions regarding the program and their correlates).

While certain features of our experiment may be difficult to replicate in other settings, and certain aspects of the Indian context (like low average levels of learning and low norms for teacher effort), may be most relevant to developing countries, our results suggest that performance pay for teachers could be an effective policy tool in India, and perhaps in other similar contexts as well. Input and incentive-based policies for improving school quality are not mutually exclusive, but our results suggest that conditional on the status quo patterns of spending in India, the marginal returns to spending additional resources on performance-linked bonuses for teachers may be higher than additional spending on unconditionally-provided school inputs. Finally, the finding that more educated and better trained teachers responded better to the incentives (while teacher education and training were not correlated with learning outcomes in

---

five times lower salaries (Muralidharan and Sundararaman 2010b). Combining the results would suggest that introducing a performance pay program would be ten times more effective at increasing test scores than reducing class size with an extra civil-service teacher.

comparison schools), highlights the potential for incentives to be a productivity-enhancing measure that can improve the effectiveness of other school inputs (including teacher human capital).

However, there are several unresolved issues and challenges that need to be addressed before scaling up teacher performance pay programs. One area of uncertainty is the optimal ratio of base and bonus pay. Setting the bonus too low might not provide adequate incentives to induce higher effort, while setting it too high increases both the risk premium and the probability of undesirable distortions. We have also not devised or tested the optimal long-term formula for teacher incentive payments. While the formula used in this project avoided the most common pitfalls of performance pay from an incentive design perspective, its accuracy was limited by the need for the bonus formula to be transparent to all teachers (most of whom were encountering a performance-based bonus for the first time in their careers). A better formula for teacher bonuses would net out home inputs to estimate a more precise measure of teachers' value addition. It would also try and account for the fact that the transformation function from teacher effort into student outcomes is likely to be different at various points in the achievement distribution. A related concern is measurement error and the potential lack of reliability of test scores and estimates of teacher value addition at the class and school levels.

The incentive formula can be improved with teacher data over multiple years and by drawing on the growing literature on estimating teacher value-added models (see the essays in Haertel and Herman 2005 and the special issue of *Education Finance and Policy* in Fall 2009) as well as papers complementary to ours that focus on the theoretical properties of optimal incentive formulae for teachers (see Barlevy and Neal 2010 and Neal 2010 for recent contributions). However, there may be a practical trade-off between the accuracy and precision of the bonus formula on one hand and the transparency of the system to teachers on the other. Teachers accepted the intuitive 'average gain' formula and trusted the procedure used and communicated by the Azim Premji Foundation. If such a program were to become policy, it is likely that teachers will start getting more sophisticated about the formula, at which point the decision regarding where to locate on the accuracy-transparency frontier can be made in consultation with

teachers. At the same time, it is possible that there may be no satisfactory resolution of the tension between accuracy and transparency.[36]

While the issue of the optimal formula for teacher performance pay has not been resolved, and implementation concerns are very real, this paper presents rigorous experimental evidence that even modest amounts of performance-based pay for teachers can lead to substantial improvements in student learning outcomes, with limited negative consequences (when implemented in a transparent and credible way). As school systems around the world consider adopting various forms of performance pay for teachers, attempts should be made to build in rigorous impact evaluations of these programs. A related point is that the details of the design of teacher incentive systems matter and should be informed by economic theory to improve the likelihood of their success (see Neal 2010). Programs and studies could also attempt to vary the magnitude of the incentives to estimate outcome elasticity with respect to the extent of variable pay, and thereby gain further insights not only on performance pay for teachers, but on performance pay in organizations in general.

[36] Murnane and Cohen (1986) point out that one of the main reasons why merit-pay plans fail is that it is difficult for principals to clearly explain the basis of evaluations to teachers. However, Kremer and Chen (2001) show that performance incentives, even for something as objective as teacher attendance did not work when implemented through head teachers in schools in Kenya. The head teacher marked all teachers present often enough for all of them to qualify for the prize. These results suggest that the bigger concern is not complexity, but rather human mediation, and so a sophisticated algorithm might be acceptable as long as it is clearly objective and based on transparently established ex-ante criteria.

References:

ANDRABI, T., J. DAS, A. KHWAJA, and T. ZAJONC (2009): ""Do Value-Added Estimates Add Value? Accounting for Learning Dynamics," Harvard University.

ASLAM, M., A. DE, G. KINGDON, and R. KUMAR (2011): "Economic Returns to Schooling and Skills – an Analysis of India and Pakistan," in *Education Outcomes and Poverty in the South*, ed. by C. Colclough. London: Routledge.

ATKINSON, A., S. BURGESS, B. CROXSON, P. GREGG, C. PROPPER, H. SLATER, and D. WILSON (2009): "Evaluating the Impact of Performance-Related Pay for Teachers in England," *Labour Economics*, 16, 251-261.

BAKER, G. (1992): "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, 100, 598-614.

— (2002): "Distortion and Risk in Optimal Incentive Contracts," *Journal of Human Resources*, 37, 728-51.

BANDIERA, O., I. BARANKAY, and A. RASUL (2007): "Incentives for Managers and Inequality among Workers: Evidence from a Firm Level Experiment," *The Quarterly Journal of Economics*, 122, 729-773.

BARLEVY, G., and D. NEAL (2010): "Pay for Percentile," University of Chicago.

CHAN, J. C. K., K. B. MCDERMOTT, and H. L. ROEDIGER III (2006): "Retrieval-Induced Facilitation: Initially Nontested Material Can Benefit from Prior Testing of Related Material," *Journal of Experimental Psychology: General*, 135, 553-571.

CHETTY, R., J. N. FRIEDMAN, N. HILGER, E. SAEZ, D. W. SCHANZENBACH, and D. YAGAN (2010): "How Does Your Kindergarten Classroom Affect Your Earnings: Evidence from Project Star," National Bureau of Economic Research Working Paper 16381.

CULLEN, J. B., and R. REBACK (2006): "Tinkering Towards Accolades: School Gaming under a Performance Accountability System," in *Advances in Applied Microeconomics, Volume 14*: Elsiever, 1-34.

DAS, J., S. DERCON, P. KRISHNAN, J. HABYARIMANA, K. MURALIDHARAN, and V. SUNDARARAMAN (2010): "When Can School Inputs Improve Test Scores?," UC San Diego.

DEMING, D. (2009): "Early Childhood Intervention and Life-Cycle Skill Development: Evidence from Head Start," *American Economic Journal: Applied Economics*, 1, 111-34.

DIXIT, A. (2002): "Incentives and Organizations in the Public Sector: An Interpretative Review," *Journal of Human Resources*, 37, 696-727.

DUFLO, E., R. HANNA, and S. RYAN (2007): "Monitoring Works: Getting Teachers to Come to School," Cambridge, MA: MIT.

FIGLIO, D. N., and L. KENNY (2007): "Individual Teacher Incentives and Student Performance," *Journal Public Economics*, 91, 901-914.

GIBBONS, R. (1998): "Incentives in Organizations," *Journal of Economic Perspectives*, 12, 115-32.

GLEWWE, P., N. ILIAS, and M. KREMER (2010): "Teacher Incentives," *American Economic Journal: Applied Economics*, 2, 205-227.

GLEWWE, P., M. KREMER, and S. MOULIN (2009): "Many Children Left Behind? Textbooks and Test Scores in Kenya," *American Economic Journal: Applied Economics*, 1, 112-135.

GOODMAN, S., and L. TURNER (2010): "Teacher Incentive Pay and Educational Outcomes: Evidence from the Nyc Bonus Program," Columbia University.

GORDON, R., T. KANE, and D. STAIGER (2006): "Identifying Effective Teachers Using Performance on the Job," Washington DC: The Brookings Institution.
GREEN, J. R., and N. L. STOKEY (1983): "A Comparison of Tournaments and Contracts," *Journal of Political Economy*, 91, 349-64.
HAERTEL, E. H., and J. L. HERMAN (2005): *Uses and Misuses of Data for Educational Accountability and Improvement*. Blackwell Synergy.
HAMILTON, B. H., J. A. NICKERSON, and H. OWAN (2003): "Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation," *Journal of Political Economy* 111, 465-97.
HANUSHEK, E. (2006): "School Resources," in *Handbook of the Economics of Education (Vol 2)*, ed. by E. Hanushek, and F. Welch: North-Holland.
HANUSHEK, E., and S. RIVKIN (2006): "Teacher Quality," in *Handbook of the Economics of Education*, ed. by E. Hanushek, and F. Welch: North-Holland.
HECKMAN, J., and J. SMITH (1995): "Assessing the Case of Social Experiments," *Journal of Economic Perspectives*, 9, 85-110.
HOLMSTROM, B. (1982): "Moral Hazard in Teams," *Bell Journal of Economics*, 13, 324-40.
HOLMSTROM, B., and P. MILGROM (1987): "Aggregation and Linearity in the Provision of Intertemporal Incentives," *Econometrica*, 55, 303-28.
— (1991): "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organization*, 7, 24-52.
HOXBY, C. M., and A. LEIGH (2005): "Pulled Away or Pushed Out? Explaining the Decline of Teacher Aptitude in the United States," *American Economic Review*, 94, 236-40.
ITOH, H. (1991): "Incentives to Help in Multi-Agent Situations," *Econometrica*, 59, 611-36.
JACOB, B.-A., L. LEFGREN, and D. SIMS (2008): "The Persistence of Teacher-Induced Learning Gains," National Bureau of Economic Research Working Paper 14065.
JACOB, B. A. (2005): "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools," *Journal of Public Economics*, 89, 761-96.
JACOB, B. A., and S. D. LEVITT (2003): "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics* 118, 843-77.
KANDEL, E., and E. LAZEAR (1992): "Peer Pressure and Partnerships," *Journal of Political Economy*, 100, 801-17.
KANDORI, M. (1992): "Social Norms and Community Enforcement," *Review of Economic Studies*, 59, 63-80.
KANE, T. J., and D. O. STAIGER (2002): "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives*, 16, 91-114.
KINGDON, G. G., and M. MUZAMMIL (2001): "A Political Economy of Education in India: The Case of U.P.," *Economic and Political Weekly*, 36.
KORETZ, D. M. (2002): "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity," *Journal of Human Resources*, 37, 752-77.
KREMER, M., and D. CHEN (2001): "An Interim Program on a Teacher Attendance Incentive Program in Kenya," Harvard University.
KREMER, M., K. MURALIDHARAN, N. CHAUDHURY, F. H. ROGERS, and J. HAMMER (2005): "Teacher Absence in India: A Snapshot," *Journal of the European Economic Association*, 3, 658-67.
LADD, H. F. (1999): "The Dallas School Accountability and Incentive Program: An Evaluation of Its Impacts on Student Outcomes," *Economics of Education Review*, 18, 1-16.

LAVY, V. (2002): "Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement," *Journal of Political Economy*, 110, 1286-1317.
— (2009): "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics," *American Economic Review*, 99, 1979 - 2011.
LAZEAR, E. (2000): "Performance Pay and Productivity," *American Economic Review*, 90, 1346-61.
— (2003): "Teacher Incentives," *Swedish Economic Policy Review*, 10, 179-214.
LAZEAR, E., and S. ROSEN (1981): "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, 89, 841-64.
MURALIDHARAN, K., and V. SUNDARARAMAN (2009): "Teacher Performance Pay: Experimental Evidence from India," National Bureau of Economic Research Working Paper 15323.
— (2010): "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India," *Economic Journal*, 120, F187-F203.
— (2010): "Contract Teachers: Experimental Evidence from India," UC San Diego.
— (2011): "Teacher Opinions on Performance Pay: Evidence from India," *Economics of Education Review*, 30.
MURNANE, R. J., and D. K. COHEN (1986): "Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive," *Harvard Educational Review*, 56, 1-17.
NEAL, D. (2010): "The Design of Performance Pay in Education," University of Chicago.
NEAL, D., and D. SCHANZENBACH (2010): "Left Behind by Design: Proficiency Counts and Test-Based Accountability," *The Review of Economics and Statistics*, 92, 263-283.
OYER, P. (1998): "Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality," *Quarterly Journal of Economics*, 113, 149-85.
PODGURSKY, M., and M. SPRINGER (2007): "Teacher Performance Pay: A Review," *Journal of Policy Analysis and Management*, 26, 909-950.
PRATHAM (2010): *Annual Status of Education Report*.
PRENDERGAST, C. (1999): "The Provision of Incentives in Firms," *Journal of Economic Literature*, 37, 7-63.
RIVKIN, S. G., E. A. HANUSHEK, and J. F. KAIN (2005): "Teachers, Schools, and Academic Achievement," *Econometrica*, 73, 417-58.
ROCKOFF, J. E. (2004): "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review*, 94, 247-252.
SPRINGER, M. G., D. BALLOU, L. HAMILTON, V.-N. LE, J. R. LOCKWOOD, D. MCCAFFREY, M. PEPPER, and B. STECHER (2010): "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching," Nashville, TN: National Center for Performance Incentives at Vanderbilt University.
UMANSKY, I. (2005): "A Literature Review of Teacher Quality and Incentives: Theory and Evidence," in *Incentives to Improve Teaching: Lessons from Latin America*, ed. by E. Vegas. Washington, D.C: World Bank, 21-61.

Table 1: Sample Balance Across Treatments

		Panel A (Means of Baseline Variables)			
		[1]	[2]	[3]	**[4]**
		Control	Group Incentive	Individual Incentive	**P-value (Equality of all groups)**
	School-level Variables				
1	Total Enrollment (Baseline: Grades 1-5)	113.2	111.3	112.6	**0.82**
2	Total Test-takers (Baseline: Grades 2-5)	64.9	62.0	66.5	**0.89**
3	Number of Teachers	3.07	3.12	3.14	**0.58**
4	Pupil-Teacher Ratio	39.5	40.6	37.5	**0.66**
5	Infrastructure Index (0-6)	3.19	3.14	3.26	**0.84**
6	Proximity to Facilities Index (8-24)	14.65	14.66	14.72	**0.98**
	Baseline Test Performance				
7	Math (Raw %)	18.5	18.0	17.5	**0.69**
8	Math (Normalized - in Std. deviations)	0.032	0.001	-0.032	**0.70**
9	Telugu (Raw %)	35.1	34.9	33.5	**0.52**
10	Telugu (Normalized - in Std. deviations)	0.026	0.021	-0.046	**0.53**
		Panel B (Means of Endline Variables)			
	Teacher Turnover and Attrition				
	Year 1 (relative to Year 0)				
11	Teacher Attrition (%)	0.30	0.34	0.31	**0.63**
12	Teacher Turnover (%)	0.34	0.33	0.32	**0.90**
	Year 2 (relative to Year 0)				
13	Teacher Attrition (%)	0.35	0.37	0.34	**0.67**
14	Teacher Turnover (%)	0.34	0.36	0.33	**0.77**
	Student Turnover and Attrition				
	Year 1 (relative to Year 0)				
15	Student Attrition from baseline to end of year tests	0.081	0.065	0.066	**0.15**
16	Baseline Maths test score of attritors (Equality of all groups)	-0.17	-0.13	-0.22	**0.77**
17	Baseline Telugu test score of attritors (Equality of all groups)	-0.26	-0.17	-0.25	**0.64**
	Year 2 (relative to Year 0)				
18	Student Attrition from baseline to end of year tests	0.219	0.192	0.208	**0.23**
19	Baseline Maths test score of attritors (Equality of all groups)	-0.13	-0.05	-0.14	**0.56**
20	Baseline Telugu test score of attritors (Equality of all groups)	-0.18	-0.11	-0.21	**0.64**

Notes:

1. The infrastructure index is the sum of six binary variables showing the existence of a brick building, a playground, a compound wall, a functioning source of water, a functional toilet, and functioning electricity.

2. The proximity index is the sum of 8 variables (each coded from 1-3) indicating proximity to a paved road, a bus stop, a public health clinic, a private health clinic, public telephone, bank, post office, and the mandal educational resource center.

3. Teacher attrition refers to the fraction of teachers in the school who left the school during the year, while teacher turnover refers to the fraction of new teachers in the school at the end of the year (both are calculated relative to the list of teachers in the school at the start of the year)

4. The p-values for the baseline test scores and attrition are computed by treating each student/teacher as an observation and clustering the standard errors at the school level (Grade 1 did not have a baseline test). The other p-values are computed treating each school as an observation.

Table 2: Impact of Incentives on Student Test Scores

Panel A: Combined (Math and Language)

	Dependent Variable = Normalized End of Year Test Score			
	Year 1 on Year 0		**Year 2 on Year 0**	
	[1]	[2]	[3]	[4]
Normalized Lagged Test Score	0.503***	0.498***	0.452***	0.446***
	(0.013)	(0.013)	(0.015)	(0.015)
Incentive School	**0.149*** **	**0.165*** **	**0.219*** **	**0.224*** **
	(0.042)	**(0.042)**	**(0.047)**	**(0.048)**
School and Household Controls	No	Yes	No	Yes
Observations	42145	37617	29760	24665
R-squared	0.31	0.34	0.24	0.28

Panel B: Math

	Dependent Variable = Normalized End of Year Test Score			
	Year 1 on Year 0		**Year 2 on Year 0**	
	[1]	[2]	[3]	[4]
Normalized Lagged Test Score	0.492***	0.491***	0.414***	0.408***
	(0.016)	(0.016)	(0.022)	(0.022)
Incentive School	**0.180*** **	**0.196*** **	**0.273*** **	**0.280*** **
	(0.049)	**(0.049)**	**(0.055)**	**(0.056)**
School and Household Controls	No	Yes	No	Yes
Observations	20946	18700	14797	12255
R-squared	0.30	0.33	0.25	0.28

Panel C: Telugu (Language)

	Dependent Variable = Normalized End of Year Test Score			
	Year 1 on Year 0		**Year 2 on Year 0**	
	[1]	[2]	[3]	[4]
Normalized Lagged Test Score	0.52***	0.510***	0.49***	0.481***
	(0.014)	(0.014)	(0.014)	(0.014)
Incentive School	**0.118*** **	**0.134*** **	**0.166*** **	**0.168*** **
	(0.040)	**(0.039)**	**(0.045)**	**(0.044)**
School and Household Controls	No	Yes	No	Yes
Observations	21199	18917	14963	12410
R-Squared	0.33	0.36	0.26	0.30

Notes:

1. All regressions include mandal (sub-district) fixed effects and standard errors clustered at the school le
2. School controls include an infrastructure and proximity index (as defined in Table 1).
3. Household controls include student caste, parental education, and affluence (as defined in Table 5A).

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 3 : Impact of Incentives by Repeat and Non-Repeat Questions

	Dependent Variable : Percentage Score					
	Combined		**Math**		**Telugu**	
	Year 1	Year 2	Year 1	Year 2	Year 1	Year 2
Percentage Score on Non-repeat Questions	0.335***	0.328***	0.256***	0.257***	0.414***	0.397***
	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)	(0.007)
Percentage Score on Repeat Questions	0.352***	0.42***	0.252***	0.386***	0.452***	0.468***
	(0.006)	(0.005)	(0.007)	(0.006)	(0.007)	(0.007)
Incremental Score in Incentive Schools for Non-repeats	0.030***	0.039***	0.033***	0.046***	0.027***	0.033***
	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)
Incremental Score in Incentive Schools for Repeats	0.043***	0.043***	0.042***	0.044***	0.043***	0.041***
	(0.011)	(0.011)	(0.013)	(0.012)	(0.011)	(0.013)
Test For Equality of Treatment Effect for Repeat and Non-repeat Questions (F-stat p-value)	**0.141**	**0.584**	**0.374**	**0.766**	**0.076**	**0.354**
Observations	62872	54972	31225	29594	31647	25378
R-Squared	0.24	0.18	0.26	0.23	0.29	0.18

Notes

1. Repeat questions are questions that at the time of administering the particular test had appeared identically on ANY earlier test (across grades)

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 4 : Impact of Incentives by Multiple Choice and Non-Multiple Choice Questions

	Dependent Variable : Percentage Score					
	Combined		**Math**		**Telugu**	
	Year 1	Year 2	Year 1	Year 2	Year 1	Year 2
Percentage Score on non Multiple-choice Questions	0.311***	0.311***	0.258***	0.278***	0.364***	0.344***
	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)	(0.008)
Percentage Score on Multiple-choice Questions (MCQ's)	0.379***	0.391***	0.227***	0.284***	0.529***	0.497***
	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)	(0.005)
Incremental Score on non MCQ's in Incentive Schools	0.028***	0.037***	0.032***	0.047***	0.023**	0.027**
	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)
Incremental Score on MCQ's in Incentive Schools	0.034***	0.042***	0.034***	0.041***	0.034***	0.042***
	(0.009)	(0.009)	(0.009)	(0.009)	(0.011)	(0.009)
Test For Equality of Treatment Effect for MCQ's and non-MCQ's (F-stat p-value)+A79	**0.168**	**0.282**	**0.671**	**0.341**	**0.119**	**0.025**
Observations	84290	59520	41892	29594	42398	29926
R-Squared	0.197	0.187	0.213	0.178	0.302	0.289

Notes

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 5: Heterogenous Treatment Effects

Panel A: Household and School Characteristics

	[1] Log School Enrollment	[2] School Proximity (8 - 24)	[3] School Infrastructure (0 - 6)	[4] Household Affluence (0 - 7)	[5] Parental Literacy (0 - 4)	[6] Scheduled Caste/ Tribe	[7] Male	[8] Normalised Baseline Score
	Two-Year Effect							
Incentive	-0.198	-0.019	0.28**	0.09	0.224***	0.226***	0.233***	0.219***
	(0.354)	(0.199)	(0.130)	(0.073)	(0.054)	(0.049)	(0.049)	(0.047)
Covariate	-0.065	-0.005	0.025	0.017	0.068***	-0.066	0.029	0.448***
	(0.058)	(0.010)	(0.038)	(0.014)	(0.015)	(0.042)	(0.027)	(0.024)
Interaction	**0.083**	**0.018**	**-0.02**	**0.038****	**-0.003**	**-0.013**	**-0.02**	**0.006**
	(0.074)	**(0.014)**	**(0.040)**	**(0.019)**	**(0.019)**	**(0.056)**	**(0.034)**	**(0.031)**
Observations	29760	29760	29760	25231	25226	29760	25881	29760
R-squared	0.244	0.244	0.243	0.272	0.273	0.244	0.266	0.243
	One-Year Effect							
Incentive	-0.36	-0.076	0.032	0.004	0.166***	0.164***	0.157***	0.149***
	(0.381)	(0.161)	(0.110)	(0.060)	(0.047)	(0.045)	(0.044)	(0.042)
Covariate	-0.128**	-0.016*	-0.001	0.017	0.08***	0.007	0.016	0.502***
	(0.061)	(0.008)	(0.025)	(0.013)	(0.012)	(0.035)	(0.020)	(0.021)
Interaction	**0.103**	**0.017**	**0.041**	**0.042****	**-0.013**	**-0.06**	**0.002**	**0.000**
	(0.081)	**(0.011)**	**(0.031)**	**(0.017)**	**(0.016)**	**(0.048)**	**(0.025)**	**(0.026)**
Observations	42145	41131	41131	38545	38525	42145	39540	42145
R-squared	0.31	0.32	0.32	0.34	0.34	0.31	0.33	0.31

Notes:

1. The infrastructure and proximity indices are defined as in Table 1.

2. The household affluence index ranges sums seven binary variables including ownership of land, ownership of current residence, residing in a "pucca" house (house with four walls and a cement and concrete roof) and having each of electricity, water, toilet, and a television at home.

3. Parental education ranges from 0 to 4 in which a point is added for each of the following: father's literacy, mother's literacy, father having completed 10th grade, and mother having completed 10th grade.

4. Scheduled Castes and Tribes are considered the most socioeconomically backward groups in India.

Panel B: Teacher Characteristics

	[1] Education	[2] Training	[3] Years of experience	[4] Salary (log)	[5] Male	[6] Teacher Absence	[7] Active Teaching	[8] Active or Passive Teaching
	Pooled regression using both years of data							
Incentive	-0.113	-0.224	0.258***	1.775**	0.031	0.15***	0.084	0.118
	(0.163)	(0.176)	(0.059)	(0.828)	(0.091)	(0.050)	(0.054)	(0.074)
Covariate	0.003	-0.051	-0.001	-0.034	-0.084	-0.149	0.055	0.131
	(0.032)	(0.041)	(0.003)	(0.066)	(0.057)	(0.137)	(0.078)	(0.093)
Interaction	**0.086***	**0.138****	**-0.009****	**-0.179***	**0.09**	**0.013**	**0.164***	**0.064**
	(0.050)	**(0.061)**	**(0.004)**	**(0.091)**	**(0.069)**	**(0.171)**	**(0.098)**	**(0.111)**
Observations	53737	53890	54142	53122	54142	53609	53383	53383
R-squared	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29

Notes:

1. Teacher education is coded from 1-4 indicating 10th grade, 12th grade, College degree and Master's or higher degree

2. Teacher training is coded from 1-4 indicating no training, a Diploma, a bachelor's degree in Education, and a Master's degree in Education.

3. Teacher absence and active teaching are determined from direct observations 4-6 times a year.

4. All regressions include mandal (sub-district) fixed effects, lagged normalized test scores, and standard errors clustered at the school level.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 6 : Impact of Incentives on Non-Incentive Subjects

Panel A: Reduced Form Impact

	Dependent Variable : Normalized Endline Score			
	Year 1		Year 2	
	Science	**Social Studies**	**Science**	**Social Studies**
Normalized Baseline Math Score	0.215***	0.224***	0.156***	0.167***
	(0.019)	(0.018)	(0.023)	(0.024)
Normalized Baseline Language Score	0.209***	0.289***	0.212***	0.189***
	(0.019)	(0.019)	(0.023)	(0.024)
Incentive School	**0.112****	**0.141*****	**0.113****	**0.18*****
	(0.052)	**(0.048)**	**(0.044)**	**(0.050)**
Observations	11786	11786	9143	9143
R-squared	0.26	0.31	0.19	0.18

Panel B: Mechanism of Impact

	Dependent Variable : Normalized Endline Score			
	Year 1		Year 2	
	Science	**Social Studies**	**Science**	**Social Studies**
Normalised Math predicted score	0.382***	0.340***	0.274***	0.330***
	(0.032)	(0.027)	(0.041)	(0.044)
Normalised Telugu predicted score	0.298***	0.487***	0.429***	0.360***
	(0.028)	(0.026)	(0.036)	(0.036)
Normalised Math residual score	0.319***	0.276***	0.232***	0.247***
	(0.025)	(0.024)	(0.032)	(0.035)
Normalised Telugu residual score	0.343***	0.425***	0.399***	0.341***
	(0.024)	(0.025)	(0.032)	(0.036)
Incentive School	**-0.01**	**0.011**	**-0.054***	**0.009**
	(0.031)	**(0.027)**	**(0.030)**	**(0.033)**
Incentive School * Normalised math residual score	0.048	0.045	-0.007	0.014
	(0.035)	(0.031)	(0.038)	(0.042)
Incentive School * Normalised telugu residual score	-0.006	0.024	0.058	0.099**
	(0.029)	(0.031)	(0.039)	(0.043)
Test for equality math and telugu residuals	0.548	0.001	0.002	0.128
Observations	11228	11228	8949	8949
R-squared	0.48	0.54	0.41	0.39

Notes:

1. Social Studies and Science tests were only administered to grades 3 to 5

2. Predicted and residual scores in Panel B are generated from a regression of the normalised test score (by subject and year) on baseline test score and other school and household characteristics in the control schools

3. All regressions include mandal (sub-district) fixed effects and standard errors clustered at the school level.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 7: Group versus Individual Incentives

Dependent Variable = Normalized End of Year Test Score

	Year 1 on Year 0			Year 2 on Year 0		
	Combined	Maths	Telugu	Combined	Maths	Telugu
	[1]	[2]	[3]	[4]	[5]	[6]
Individual Incentive School (II)	**0.156*** **	0.184***	0.130***	**0.283*** **	0.329***	0.239***
	(0.050)	(0.059)	(0.045)	**(0.058)**	(0.067)	(0.054)
Group Incentive School (GI)	**0.141*** **	0.175***	0.107**	**0.154*** **	0.216***	0.092*
	(0.050)	(0.057)	(0.047)	**(0.057)**	(0.068)	(0.052)
F-Stat p-value (Testiing GI = II)	**0.765**	**0.889**	**0.610**	**0.057**	**0.160**	**0.016**
Observations	42145	20946	21199	29760	14797	14963
R-squared	0.31	0.299	0.332	0.25	0.25	0.26

Notes:

1. All regressions include mandal (sub-district) fixed effects and standard errors clustered at the school level.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 8: Teacher Behavior (Observation and Interviews)

Teacher Behavior	Incentive versus Control Schools (All figures in %)			
	Incentive Schools	**Control Schools**	**p-Value of Difference**	**Correlation with student test score**
	[1]	[2]	[3]	[4]
Teacher Absence (%)	0.25	0.23	**0.199**	-0.103
Actively Teaching at Point of Observation (%)	0.42	0.43	**0.391**	0.135***
Did you do any special preparation for the end of year tests? (% Yes)	0.64	0.32	**0.000*** **	0.095**
What kind of preparation did you do? (UNPROMPTED) (% Mentioning)				
Extra Homework	0.42	0.20	**0.000*** **	0.061
Extra Classwork	0.47	0.23	**0.000*** **	0.084**
Extra Classes/Teaching Beyond School Hours	0.16	0.05	**0.000*** **	0.198***
Gave Practice Tests	0.30	0.14	**0.000*** **	0.105**
Paid Special Attention to Weaker Children	0.20	0.07	**0.000*** **	0.010

Notes:

1. All teacher response variables from the teacher interviews are binary and column 4 reports the correlation between a teacher's stated response and the test scores of students taught by that teacher (controlling for lagged test scores as in the default specifications throughout the

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 9: Impact of Inputs versus Incentives on Learning Outcomes

Dependent Variable = Normalized End of Year Test Score

	Year 1 on Year 0			Year 2 on Year 0		
	Combined	**Math**	**Language**	**Combined**	**Math**	**Language**
	[1]	[2]	[3]	[4]	[5]	[6]
Normalised Lagged Score	0.512***	0.494***	0.536***	0.458***	0.416***	0.499***
	(0.010)	(0.012)	(0.011)	(0.012)	(0.016)	(0.012)
Incentives	**0.15*** **	0.179***	0.121***	**0.218*** **	0.272***	0.164***
	(0.041)	(0.048)	(0.039)	**(0.049)**	(0.057)	(0.046)
Inputs	**0.102*** **	0.117***	0.086**	**0.085* **	0.089*	0.08*
	(0.038)	(0.042)	(0.037)	**(0.046)**	(0.052)	(0.044)
F-Stat p-value (Inputs = Incentives)	**0.178**	0.135	0.298	**0.003**	0.000	0.044
Observations	69157	34376	34781	49503	24628	24875
R-squared	0.30	0.29	0.32	0.225	0.226	0.239

Notes:

1. These regressions pool data from all 500 schools in the study: 'Group' and 'Individual' incentive treatments are pooled together as "Incentives", and the 'Extra contract teacher' and 'Block grant' treatments are pooled together as "Inputs".

2. All regressions include mandal (sub-district) fixed effects and standard errors clustered at the school level.

* significant at 10%; ** significant at 5%; *** significant at 1%

Figure 1a: Andhra Pradesh (AP)

	India	AP
Gross Enrollment (Ages 6-11) (%)	95.9	95.3
Literacy (%)	64.8	60.5
Teacher Absence (%)	25.2	25.3
Infant Mortality (per 1000)	63	62

Figure 1b: District Sampling (Stratified by Socio-cultural Region of AP)

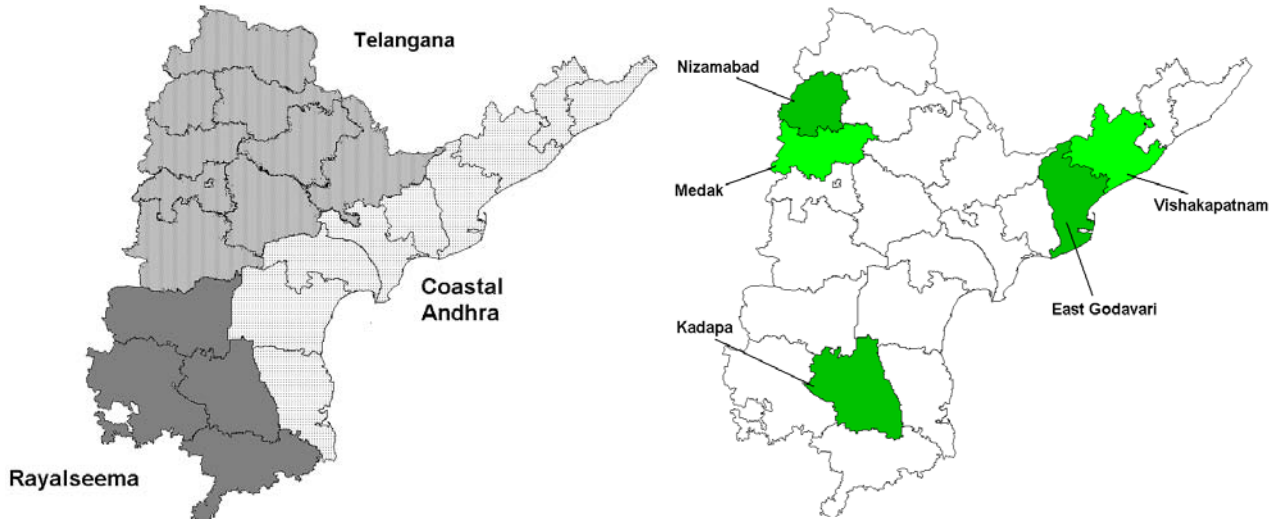

Figure 2: Test Score Distribution after Two Years of Program by Treatment Status

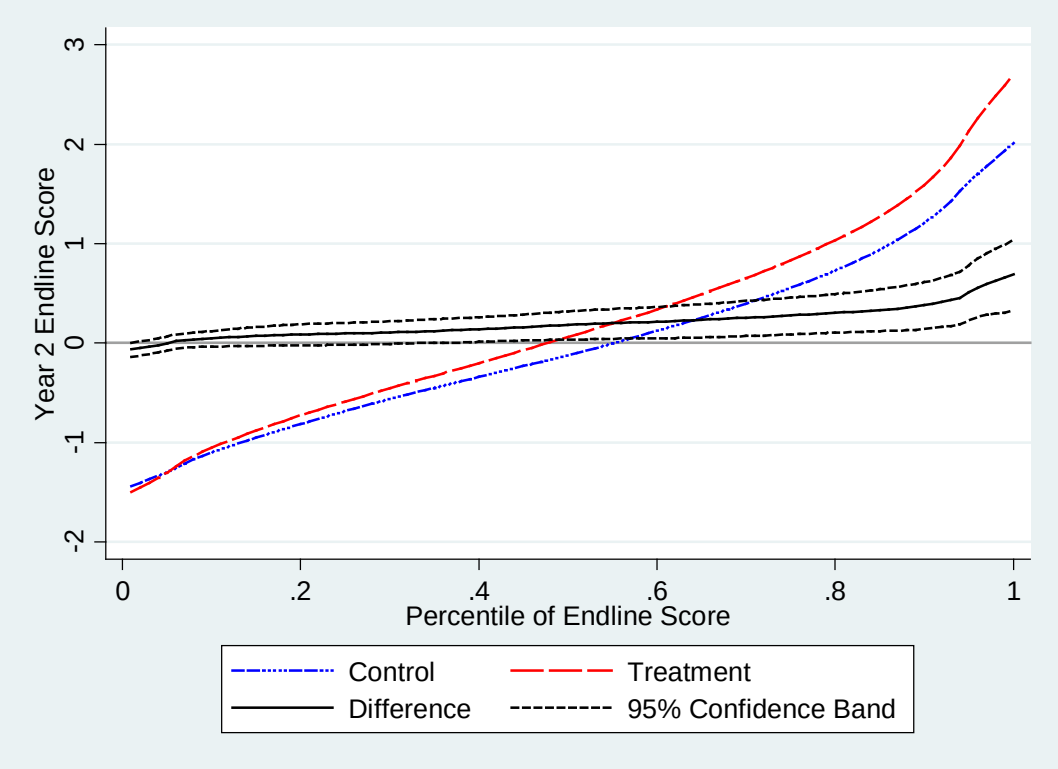

Figure 3: Heterogeneous Treatment Effects by Baseline Score Percentile

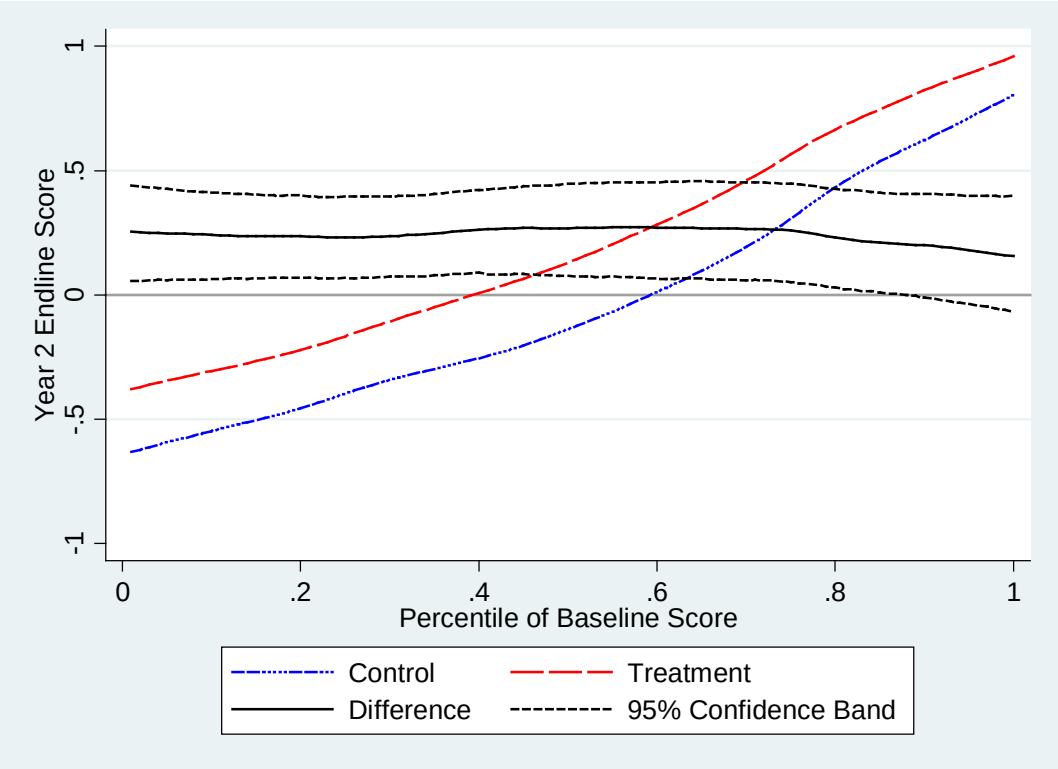

Figure 4: Teacher Fixed Effects by Treatment Status