Teacher Performance Pay: Experimental Evidence from India

Karthik Muralidharan[†]

Venkatesh Sundararaman[‡]

This Draft: 12 May 2008[*]

Abstract: Performance pay for teachers is frequently suggested as a way of improving educational outcomes in schools, but the empirical evidence to date on its effectiveness is limited and mixed. We present results from a randomized evaluation of a teacher incentive program implemented across a 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 0.28 and 0.16 standard deviations in math and language tests respectively. They scored significantly higher on "conceptual" as well as "mechanical" components of the tests suggesting that the gains in test scores represented an actual increase in learning outcomes. Incentive schools also performed better on subjects for which there were no incentives. Group and individual incentive schools perform equally well in the first year of the program, but the individual incentive schools significantly outperform in the second year. 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: performance pay, teacher incentives, group and individual incentives, education policy, field experiments

---

[†] Harvard University Graduate School of Education; E-mail: muralika@gse.harvard.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 Amrita Ahuja, George Baker, Efraim Benmelech, Jishnu Das, Martin Feldstein, Richard Freeman, Robert Gibbons, Edward Glaeser, Richard Holden, Asim Khwaja, Sendhil Mullainathan, Ben Olken, Lant Pritchett, Halsey Rogers, Philipp Schnabl, Kartini Shastry, Heidi Williams, Jeff Williamson, and various seminar participants for useful comments and discussions. We thank officials of the Department of School Education in Andhra Pradesh 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 outstanding work in managing the implementation of the project. Sridhar Rajagopalan, Vyjyanthi Shankar, and staff of Educational Initiatives led the test design. Vinayak Alladi, Richman Dzene and Gokul Madhavan provided excellent research assistance. The project would not have been possible without generous financial support from the UK Department for International Development (DFID) and the Government of Andhra Pradesh. Karthik Muralidharan thanks the Bradley and Spencer Foundations for fellowship support. The findings, interpretations, and conclusions expressed in this paper are those of the authors and do not necessarily represent the views of the World Bank, its Executive Directors, or the governments they represent.

1. Introduction

The focus of primary education policy in developing countries such as India has typically been on access, enrollment, and retention; however, much less attention has been paid to the low quality of learning in schools.[1] The most recent *Annual Status of Education Report* found that over 58% of children aged 6 to 14 in an all-India sample of over 300,000 rural households could not read at the second grade level, though over 95% of them were enrolled in school (Pratham, 2008).

Attempts to improve education in developing countries have typically focused on providing more inputs to schools – and have usually expanded spending along existing patterns. However, a recent 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, Muralidharan, Chaudhury, Hammer, and Rogers, 2005). Since over 90% of non-capital education spending in India goes to regular teacher salaries and benefits, it is not clear that a "business as usual" policy of expanding inputs along existing patterns is the most effective way of improving educational outcomes.

We present results from the Andhra Pradesh Randomized Evaluation Study (AP RESt)[2] that piloted and evaluated alternative approaches to improving primary education in the Indian state of Andhra Pradesh (AP) including the provision of performance bonuses to teachers based on the average improvement in test scores of their students. We studied two types of teacher performance pay (group bonuses based on school performance; and individual bonuses based on teacher performance), and the additional spending in both programs was calibrated to be slightly over 3% of a typical school's annual budget. The study was conducted by randomly allocating the incentive programs across a representative sample of 300 government-run schools in rural AP with 100

[1] For instance, the Millennium Development Goal for education is to "ensure that all boys and girls complete a full course of primary schooling," but it makes no mention of the level of learning achieved by students completing primary school.

[2] The AP RESt is a partnership between the government of AP, the Azim Premji Foundation (a leading non-profit organization working to improve primary education in India), and the World Bank. The Azim Premji Foundation (APF) was the main implementing agency for the study. We have served as technical consultants and have overseen the design, implementation, and evaluation of the various interventions.

schools in each of the two incentive treatment groups and 100 schools serving as the comparison group.

This paper attempts to answer the following questions in the context of primary education in a developing country: (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 perform relative to teacher-level individual incentives? (iv) How does teacher behavior change in response to incentives? (v) How cost effective are teacher incentives relative to other uses for the same money? and (vi) Will teachers support the idea?

At the end of two years of the program, students in incentive schools performed significantly better than those in control schools by 0.28 and 0.16 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. Incentive schools score higher in each of the 5 grades (1-5), across all quintiles of question difficulty, and in all the 5 districts where the project was conducted, with most of these differences being statistically significant. The distribution of student test score gains in the incentive schools first order stochastically dominates that of the control school distribution suggesting that no student was made worse off as a result of the program.

We find no evidence of any adverse consequences as a result of the incentive programs. Incentive schools do significantly better on both mechanical components of the test (designed to reflect rote learning) and conceptual components of the test (designed to capture deeper understanding of the material), suggesting that the gains in test scores represent an actual increase in learning outcomes. Students in incentive schools also do significantly better in science and social studies (on which there were no incentives), suggesting positive spillover effects from incentive to non-incentive subjects. There was no difference in student attrition between incentive and control schools.

In the first year, there was almost no difference between the effectiveness of school-level group incentives and teacher-level individual incentives. Since the average government-run school in rural AP is quite small with only 3 teachers, these results probably reflect a context of relatively easy peer monitoring. However, in the second year, the individual incentive schools significantly outperformed the group incentive

schools, and at the end of two years, the average treatment effect was 0.27 SD in the individual incentive schools compared to 0.16 SD in the group incentive schools, with this difference being nearly significant at the 10% level.

Teacher absence did not differ across treatments, and neither did teaching activity as measured by observations. However, teacher interviews indicate that teachers in incentive schools were more likely to have exerted extra effort such as assigning additional homework and class work, providing practice tests, and conducting extra classes after school. These measures of teacher behavior are all significantly correlated with student test score gains suggesting that the incentives did not work through increased teacher attendance but on greater teaching effort conditional on being present

In a parallel initiative, two other sets of 100 randomly chosen schools were provided with additional 'smart inputs' that were hypothesized to be more cost effective than the status quo patterns of spending. The interventions were calibrated so that the expected spending on the inputs and the incentives was roughly equal.[3] At the end of two years of the various programs, students in schools receiving the input programs scored 0.08 SD higher than those in comparison schools. However, the incentive programs spent around 25% less in bonuses paid and had a significantly larger impact on learning outcomes (0.22 versus 0.08 SD). Thus, performance-based bonus payments to teachers were a significantly more cost effective way of increasing student test scores compared to spending the money unconditionally on additional schooling inputs.

Bonuses are another way of paying a salary. So, it may be possible to introduce a performance pay component to the wage structure in lieu of scheduled 'across the board' increase in salaries. In this scenario, the cost of performance pay is not the bonus payment itself, but the risk premium that has to be paid to keep teachers' expected utility constant under a system of variable pay. Since the risk premium would be much lower than the expected bonus payment, the long-run cost of a teacher incentive program could be even lower than the cost of additional bonus payments made in the short run.

There was broad-based support from teachers for the program. Over 85% of them were in favor of the idea of bonus payments on the basis of performance, and over 75%

[3] These inputs consisted of either an extra para-teacher, or a cash block grant to schools. See our companion papers (Muralidharan and Sundararaman 2008b and 2008c) for details on the input interventions and their impact on learning outcomes.

favored such a scheme even if their expected wage were to be held constant. We also find that the extent of teachers' ex-ante support for performance pay (over a series of mean-preserving spreads of pay) is positively correlated with their ex-post performance. This suggests that teachers are aware of their own effectiveness (as measured by test scores) and that performance pay might not only increase effort among existing teachers, but systematically draw more effective teachers into the profession over time.[4]

Our results contribute to a small but growing literature on the effectiveness of performance-based pay for teachers.[5] Unfortunately, a majority of teacher incentive programs have been implemented in ways that make it difficult to construct a statistically valid comparison group against which the impact of the incentives can be assessed. The best identified studies on the effect of paying teachers on the basis of student test outcomes are Lavy (2002) and (2007), and Glewwe, Ilias, and Kremer (2003), but their evidence is mixed. Lavy uses regression discontinuity 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 2007 papers respectively). Glewwe et al (2003) 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 conclude that the results are consistent with teachers expending effort towards short-term increases in test scores but not towards long-term learning.

We make several original contributions in this paper. We present results from the first randomized evaluation of teacher performance pay in a representative sample of schools.[6] We take the test design seriously and include both 'mechanical' and 'conceptual' questions in the tests to distinguish rote learning from a broader increase in

[4] Lazear (2000) shows that around half the gains from performance-pay in the company he studied were due to more productive workers being attracted to join the company under a performance-pay system. Similarly, Hoxby and Leigh (2005) argue that compression of teacher wages in the US is an important reason for the decline in teacher quality, with higher-ability teachers exiting the teacher labor market.
[5] Previous studies include Ladd (1999) in Dallas, Atkinson et al (2004) in the UK, and Figlio and Kenny (2006) who use cross-sectional data across multiple US states. See Umansky (2005) for a current literature review on various kinds of teacher incentive programs. 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.
[6] 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 previous studies.

learning outcomes. We study group (school-level) and individual (teacher-level) incentives in the same field experiment. We record differences in teacher behavior with both direct observations based on tracking surveys and teacher interviews. We study both input and incentive based policies in the same field experiment and calibrate the spending on each of these options to be similar. Finally, we interview teachers after each year of the program, but before they know their own performance bonus for the year, to determine the extent and correlates of their support for performance pay.

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.[7] The most common source of identification in this literature is to look for changes in compensation systems and compare outcomes before and after the change. However, these studies typically cannot rule out the possibility that other management practices also changed at the time the compensation system was changed, or the possibility that the change in compensation system was endogenously determined by unmeasured factors that directly impact the measured outcomes. True experiments in compensation structure with identical contemporaneous control groups are rare,[8] and our results can help to answer broader questions regarding performance pay in organizations, and group versus individual incentives in small group situations.[9]

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, while section 8 discusses teacher responsiveness to the idea of performance pay. Section 9 concludes.

---

[7] 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. Chiappori and Salanié (2003) survey recent empirical work in contract theory and emphasize the identification problems in testing incentive theory.

[8] Bandiera, Barankay, and Rasul (2006) is a recent exception that studies the impact of exogenously varied compensation schemes (though with a sequential as opposed to contemporaneous comparison group).

[9] Of course, as Dixit (2002) warns, it is important for empirical work to be cautious in making generalizations about performance-based incentives, and to focus on relating success or failure of incentive pay to context-specific characteristics such as the extent and nature of multi-tasking.

2. Theoretical Framework

2.1 Incentives and intrinsic motivation

It is not obvious that paying teachers bonuses on the basis of student test scores will raise test scores. Evidence from psychological studies suggests that monetary incentives (especially of small amounts) can sometimes crowd out intrinsic motivation and lead to inferior outcomes.[10] Teaching may be especially susceptible to this concern since many teachers are thought to enter the profession due to strong intrinsic motivation. The AP context, however, suggested that an equally valid concern was the lack of differentiation among high and low-performing teachers. Kremer et al (2006) show that in Indian government schools, teachers reporting high levels of job satisfaction are *more likely* to be absent. In subsequent focus group discussions with teachers, it was suggested that this was because teachers who were able to get by with low effort were quite satisfied, while hard-working teachers were dissatisfied because there was no difference in professional outcomes between them and those who shirked. Thus, it is also possible that the lack of external reinforcement for performance can erode intrinsic motivation.[11]

2.2 Multi-task moral hazard

Even those who agree that incentives based on test scores could improve test performance worry that such incentives could lead to sub-optimal behavioral responses from teachers. Examples of such behavior include rote 'teaching to the test' and neglecting higher-order skills (Holmstrom and Milgrom, 1991), manipulating performance by short-term strategies like boosting the caloric content of meals on the day of the test (Figlio and Winicki, 2005), excluding weak students from testing (Jacob, 2005), or even outright cheating (Jacob and Levitt, 2003).

These are all examples of the problem of multi-task moral hazard, which is illustrated by the following formulation from Baker (2002).[12] Let $\mathbf{a}$ be an n-dimensional vector of

[10] A classic reference in psychology is Deci and Ryan (1985). References in economics include Frey and Oberholzer-Gee (1997), and Gneezy and Rustichini (2000). Chapter 5 of Baron and Kreps (1999) provides an excellent discussion relating intrinsic motivation to practical incentive design and communication.
[11] Mullainathan (2006) describes how high initial intrinsic motivation of teachers can diminish over time if they feel that the government does not appreciate or reciprocate their efforts.
[12] The original references are Holmstrom and Milgrom (1991), and Baker (1992). The treatment here follows Baker (2002) which motivates the multi-tasking discussion by focusing on the divergence between the performance measure and the principal's objective function.

potential agent (teacher) actions that map into a risk-neutral principal's (social planner's) value function (V) through a linear production function of the form:

$$V(\mathbf{a}, \varepsilon) = \mathbf{f} \cdot \mathbf{a} + \varepsilon$$

where **f** is a vector of marginal products of each action on V, and ε is noise in V.

Assume the principal can observe V (but not **a**) and offers a linear wage contract of the form $w = s + b_v \cdot V$. If the agent's expected utility is given by:

$$E(s + b_v \cdot V) - h \cdot \text{var}(s + b_v \cdot V) - \sum_{i=1}^{n} a_i^2 / 2$$

where h is her coefficient of absolute risk aversion and $a_i^2 / 2$ is the cost of each action, then the optimal slope on output (b_v^*) is given by:

$$b_v^* = \frac{F^2}{F^2 + 2h\sigma_\varepsilon^2} \qquad (2.2.1)$$

where $F = \sqrt{\sum_{i=1}^{n} f_i^2}$. Expression (2.2.1) reflects the standard trade-off between risk and aligning of incentives, with the optimal slope b_v^* decreasing as h and σ_ε^2 increase.

Now, consider the case where the principal cannot observe V but can only observe a performance measure (P) that is also a linear function of the action vector **a** given by:

$$P(\mathbf{a}, \phi) = \mathbf{g} \cdot \mathbf{a} + \phi$$

Since $\mathbf{g} \neq \mathbf{f}$, P is an imperfect proxy for V (such as test scores for broader learning). However, since V is unobservable, the principal is constrained to offer a wage contract as a function of P such as $w = s + b_p \cdot P$.

The key result in Baker (2002) is that the optimal slope b_p^* on P is given by:

$$b_p^* = \frac{F \cdot G \cdot \cos\theta}{G^2 + 2h\sigma_\phi^2} \qquad (2.2.2)$$

where $G = \sqrt{\sum_{i=1}^{n} g_i^2}$, and θ is the angle between **f** and **g**. The cosine of θ is a measure of how much b_p^* needs to be reduced relative to b_v^* due to the distortion arising from $\mathbf{g} \neq \mathbf{f}$. If $\cos\theta = 1$, both expressions are equivalent except for scaling and there is no distortion.

The empirical literature in education showing that agents often respond to incentives by increasing actions on dimensions that are not valued by the principal highlights the

need to be cautious in designing incentive programs. However, in most practical cases, $\mathbf{g} \neq \mathbf{f}$ (and $\cos\theta \neq 1$), and so it is perhaps inevitable that a wage contract with $b_p > 0$ will induce some actions that are unproductive. The implication for incentive design is that $b_p^* > 0$, as long as $V(\mathbf{a}(b_p > 0)) > V(\mathbf{a}(b_p = 0))$, even if there is some deviation relative to the first-best action in the absence of distortion and $V(\mathbf{a}(b_p^*)) < V(\mathbf{a}(b_v^*))$.[13] In other words, what matters is not whether teachers engage in more or less of some activity than they would in a first-best world (with incentives on the underlying social value function), but whether the sum of their activities in a system with incentives on test scores generates more learning (broadly construed) than in a situation with no such incentives.

There are several reasons why test scores might be an adequate performance measure in the context of primary education in a developing country. First, given the extremely low levels of learning, it is likely that even an increase in routine classroom teaching of basic material will lead to better learning outcomes. Second, even if some of the gains merely reflect an improvement in test-taking skills, the fact that the education system in India is largely structured around test-taking suggests that it might be unfair to deny disadvantaged children in government-schools the benefits of test-taking skills that their more privileged counterparts in private schools develop.[14] Finally, the design of tests can get more sophisticated over time, making it difficult to do well on the tests without a deeper understanding of the subject matter. So, it is possible that additional efforts taken by teachers to improve test scores for primary school children can also lead to improvements in broader educational outcomes. Whether this is true is an empirical question and is a focus of our research design (see section 4).

2.3 Group versus Individual Incentives

The theoretical prediction of the relative effectiveness of individual and group teacher incentives is ambiguous. Let *w* = wage, *P* = performance measure, and *c(a)* = cost of

[13] Thus a key challenge is choosing the appropriate performance measure *P*. Duflo, Hanna, and Ryan (2007) evaluate the effectiveness of a program run by an NGO in rural Rajasthan (a north Indian state) that provided high-powered incentives based on teacher attendance rather than test scores and find a significant increase in both teacher attendance and student test scores, but no effect on teaching conditional on teachers being present in school. This is a promising option because the multi-tasking problem is potentially less severe with respect to attendance than with classroom activity.

[14] While the private returns to test-taking skills may be greater than the social returns, the social returns could be positive if they enable disadvantaged students to compete on a more even basis with privileged students for scarce slots in higher levels of education.

exerting effort a with $c'(a) > 0$, $c''(a) > 0$, $P'(a) > 0$, and $P''(a) < 0$. Unlike typical cases of team production, an individual teacher's output (test scores of his students) is observable, making contracts on individual output feasible. The optimal effort for a teacher facing individual incentives is to choose a_i so that: $\frac{\partial w_i}{\partial P_i} \cdot \frac{\partial P_i}{\partial a_i} = c'(a_i)$ (2.3.1)

Now, consider a group incentive program where the bonus payment is a function of the average performance of all teachers. The optimality condition for each teacher is:

$$\frac{\partial w_i}{\partial \left[\left(P_i + \sum P_{-i} \right) / n \right]} \cdot \frac{\partial \left[\left(P_i + \sum P_{-i} \right) / n \right]}{\partial a_i} = c'(a_i) \quad (2.3.2)$$

If the same bonus is paid to a teacher for a unit of performance under both group and individual incentives then $\frac{\partial w_i}{\partial \left[\left(P_i + \sum P_{-i} \right) / n \right]} = \frac{\partial w_i}{\partial P_i}$, but $\frac{\partial \left[\left(P_i + \sum P_{-i} \right) / n \right]}{\partial a_i} = \frac{1}{n} \cdot \frac{\partial P_i}{\partial a_i}$.

Since $c''(a) > 0$, the equilibrium effort exerted by each teacher under group incentives is lower than that under individual incentives. Thus, in the basic theory, group (school-level) incentives induce free riding and are therefore inferior to individual (teacher-level) incentives, when the latter are feasible.[15]

However, if the teachers jointly choose their effort levels, they will account for the externalities within the group. In the simple case where they each have the same cost and production functions and these functions do not depend on the actions of the other teachers, they will each (jointly) choose the level of effort given by (2.3.1). Of course, each teacher has an incentive to shirk relative to this first best effort level, but if teachers in the school can monitor each other at low cost, then it is possible that the same level of effort can be implemented as under individual incentives. This is especially applicable to smaller schools where peer monitoring is likely to be easier.[16]

Finally, if there are gains to cooperation, then it is possible that group incentives might yield better results than individual incentives.[17] Consider a case where teachers

[15] See Holmstrom (1982) for a solution to the problem of moral hazard in teams.
[16] See Kandori (1992) and Kandel and Lazear (1992) for discussions of how social norms and peer pressure in groups can ensure community enforcement of the first best effort level.
[17] Holmstrom and Milgrom (1990) and Itoh (1991) model incentive design when cooperation is important. Hamilton, Nickerson, and Owan (2003) present empirical evidence from a garment factory showing that group incentives for workers improved productivity relative to individual incentives.

have comparative advantages in teaching different subjects or different types of students. If teachers specialize in their area of advantage and reallocate students/subjects to reflect this, they could raise $P'(a)$ $(\forall a)$ relative to a situation where each teacher had to teach all students/subjects. Since $P''(a) < 0$, the equilibrium effort would also be higher and the outcomes under group incentives might be superior to those under individual incentives.[18]

Lavy (2002) and (2007) report results from high-school teacher incentive program in Israel at the individual and group level respectively. However, the two programs were implemented at different (non-overlapping) times and the schools were chosen by different (non-random) eligibility criteria, and the individual incentive program was only studied for one year. We study both group and individual incentives in the same field experiment over two full academic years.

3. Experimental Design

3.1 Context

Andhra Pradesh (AP) is the 5th most populous state in India, with a population of over 80 million, 73% of whom live in rural areas. 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. There are a total of over 60,000 such schools in AP and over 80% of children in rural AP attend government-run schools (Pratham, 2005).

The average rural primary school is quite small, with total enrollment of around 80 to 100 students and an average of 3 teachers across grades one through five.[19] One teacher typically teaches all subjects for a given grade (and often teaches more than one grade

[18] The additive separability of utility between income and cost of effort implies that there is no 'income effect' of higher productivity on the cost of effort, and so effort goes up in equilibrium since $P'(a)$ is higher.
[19] This is a consequence of the priority placed on providing all children with access to a primary school within a distance of 1 kilometer from their homes.

simultaneously). All regular teachers are employed by the state, and their salary is mostly determined by experience and rank,[20] with minor adjustments based on postings, but no component based on any measure of performance. The average salary of regular teachers is over Rs. 7,500/month and total compensation including benefits is close to Rs. 10,000/month (per capita income in AP is around Rs. 2,000/month; 1 US Dollar ≈ 40 Indian Rupees (Rs.)). Regular teachers' salaries and benefits comprise over 90% of non-capital expenditure on primary education in AP. Teacher unions are strong and disciplinary action for non-performance is rare.[21]

3.2 Sampling

We sampled 5 districts across each of the 3 socio-cultural regions of AP in proportion to population (Figure 1b).[22] 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 AP RESt Design Overview

The overall design of AP RESt 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 PARA TEACHER	100 Schools		
	EXTRA BLOCK GRANT	100 Schools		

As Table 3.1 shows, the inputs were provided *unconditionally* to the selected schools at the beginning of the school year, while the incentive treatments consisted of an

[20] A regression of teacher salary on experience and rank (in our sample) has an R-squared of 0.8.
[21] See Kingdon and Muzammil (2001) for an illustrative case study of the power of teacher unions in India. 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.
[22] Subject to the selected districts within a region being contiguous for ease of logistics and supervision.

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. The school year in AP starts in the middle of June, and the baseline tests were conducted in the 500 sampled schools during late June and early July, 2005.[23] After the baseline tests were scored, we randomly allocated 2 out of the 10 project schools in each mandal 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. 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 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 and that the randomization worked properly.[24]

After the randomization, mandal coordinators (MCs) from the Azim Premji Foundation (APF) 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. The MCs 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.[25] 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

[23] See Appendix A for the project timeline and activities and Appendix B for details on test administration. 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).

[24] 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.

[25] Six visits were made to each school in the first year (2005 – 06), while four visits were made in the second year (2006 – 07)

minimized and that a comparison between treatment and control schools can accurately isolate the treatment effect.

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, 2006), and all schools were informed that the program would continue for another year.[26] Bonus checks based on first year performance were sent to qualifying teachers by the end of August 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[27]

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. 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 endline test score for the first year will be the baseline score for the second year).[28]

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

[26] 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.
[27] 1st grade children were not tested in the baseline, but were in the endline. The 'baseline' for grade 1 was computed as the mean baseline score of the 2nd grade children in the school, and the 'improvement' for grade 1 was calculated relative to this. The 5% threshold did not apply to the 1st grade. 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.
[28] 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.

spending in the input treatments (based on calibrations from the project pilot).[29] The threshold of 5% average improvement was introduced to account for the fact that the baseline tests were in June/July and the endline tests would be in March/April, and so the baseline score might be artificially low due to students forgetting material over the vacation. There was no improvement 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.[30]

We tried to minimize potentially undesirable 'target' 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.[31] If the function transforming teacher effort into test-scores is concave (convex) in the baseline score, teachers 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 endline test, we assigned a zero improvement score to any child who took the baseline test but not the endline test.[32] To make cheating as difficult as possible, the tests were conducted by external teams of 5 evaluators in each

[29] 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 (1982) and Green and Stokey (1983) for a discussion of tournaments and when they dominate contracts.

[30] 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 endline scores as the baseline for the next year and so a school/class whose performance deteriorates does *not* have its baseline reduced for the next year.

[31] 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 (2008) discuss the effects 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.

[32] 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 attrition to the teacher was always equal to a loss of 5% in average computation. 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.

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

We engaged India's leading educational testing firm, "Educational Initiatives" (EI), to design the tests to our specifications. The test design activities included mapping the syllabus from the text books into skills, creating a universe of questions to represent each skill, and calibrating question difficulty in a pilot exercise in 40 schools during the prior school year (2004-05) to ensure adequate discrimination on the tests (Figure 3a).[33]

The baseline test (June-July, 2005) 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 with a gap of two weeks between them. The first test (the 'lower endline') tested the same set of skills from the baseline (competencies up to that of the previous school year) with an exact mapping of question type from the baseline to lower endline to enable creation of an 'absolute' measure of learning progress in addition to comparison across treatment and control schools. The second test (the 'higher endline') tested skills from the current school year's syllabus. The same procedure was repeated at the end of the second year, with two rounds of testing. Doing two rounds of testing at the end of each year not only allows the testing of more materials, but also reduces measurement errors specific to the day of testing by having multiple tests around two weeks apart.

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 first year of the program in March-April, 2007.

[33] The low level of learning meant that a substantial fraction of children in grade 4 and 5 would score zero on a grade-appropriate test. The test papers therefore had to sample from previous years' skills in order to obtain adequate discrimination on the test. The tests encompass a broad range of difficulty except in grade 1, where it was not possible to include questions from a lower grade (Figure 3a). Figure 3a is based on the test at the end of the first year, but the tests in year 0 (baseline) and year 2 (at the end of two years of the program) shared these properties.

4.2 Basic versus higher-order skills

As highlighted in section 2.2, it is possible that broader educational outcomes are no better (or even worse) under a system of teacher incentives based on test scores even if the test scores improve. A key empirical question, therefore, is whether additional efforts taken by teachers to improve test scores for primary school children in response to the incentives are also likely to lead to improvements in broader educational outcomes. We asked EI to design the tests to include both 'mechanical' and 'conceptual' questions within each skill category on the test. The distinction between these two categories is not constant, since a conceptual question that is repeatedly taught in class can become a mechanical one. Similarly a question that is conceptual in an early grade might become mechanical in a later grade, if students acclimatize to the idea over time. For this study, a mechanical question was considered to be one that conformed to the format 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.

As an example, consider the following pair of questions (which did not appear sequentially) from the 4th grade math test under the skill of 'multiplication and division'

Question 1:
$$\begin{array}{r} 34 \\ \times\ 5 \end{array}$$

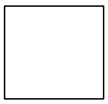
Question 2: Put the correct number in the empty box:

$8 + 8 + 8 + 8 + 8 + 8 = 8 \times \square$

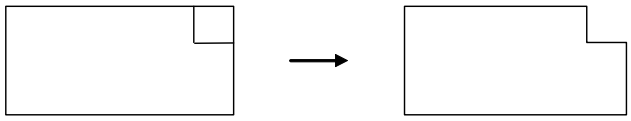
The first question follows the standard textbook format for asking multiplication questions and would be classified as "mechanical" while the second one requires the students to understand that the concept of multiplication is that of repeated addition, and would be classified as "conceptual." Note that conceptual questions are not more difficult per se. In this example, the conceptual question is arguably easier than the mechanical one because you only have to count that there are 6 '8's and enter the answer '6' as opposed to multiplying 2 numbers (with a carry over). But the conceptual question is unfamiliar and this is reflected in 43% of children getting Question 1 correct, while only 8% got Question 2 correct.

A second example is provided below from the fifth grade math test under the skill of 'Area, Volume, and Measurement'

Question 1: What is the area of the square below? _____



Question 2: A square of area 4 sq. cm is cut off from a rectangle of area 55 sq. cm.



What is the area of the remaining piece? _____ sq. cm

Again, the first question tests the idea of area straight from the textbook requiring the multiplication of the 2 sides, while the second requires an understanding of the concept of area as the magnitude of space within a closed perimeter. Of course, the distinction is not always so stark, and the classification into mechanical and conceptual is a discrete representation of a continuous scale between familiar and unfamiliar questions.[34]

4.3 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 including additional tests during the endline on science and social studies (referred to in AP collectively as Environmental Sciences or EVS) on which there was no incentive.[35] Since the subject is formally introduced only in grade 3 in the school curriculum, the EVS tests were administered in grades 3 to 5.

[34] 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.

[35] In the first year of the project, schools were not told about the EVS 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 EVS tests would also be conducted, but also knew from the first year that these tests would not be included in the bonus calculations.

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, 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). A more worrying possibility was that more teachers would try to transfer into the incentive schools in the second year of the project. 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 schools in the study for the duration of the study. As a result, the average teacher turnover in the second year was only 5%, and once again, there was no significant difference in teacher transfer rates across the various treatments (Table 1 – Panel B, rows 13 - 16).[36]

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 7.3% and 25% in year 1 and year 2 respectively, but there is no significant difference in attrition across the treatments (Table 1 – Panel B, rows 17 and 20). Beyond confirming sample balance, this is an interesting 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). Students with lower baseline scores were more likely to not take the endline in all schools, but we find no difference in mean baseline test score across treatment categories among the students who drop out (Table 1 – Panel B, rows 18, 19, 20, and 22).

[36] There was also a court order to restrict teacher transfers in response to complaints that teacher transfers during the school year were disruptive to students. This may have also helped to reduce teacher transfers during the second year of the project.

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:

$$T_{ijkm}(EL) = \alpha + \gamma \cdot T_{ijkm}(BL) + \delta \cdot Incentives + \beta \cdot Z_m + \varepsilon_k + \varepsilon_{jk} + \varepsilon_{ijk} \quad (5.1)$$

The main dependent variable of interest is T_{ijkm} which is the normalized test score on the specific test (normalized with respect to the distribution of the control schools), where *i, j, k, m* denote the student, grade, school, and mandal respectively. EL and BL indicate the endline and the baseline tests. Including the normalized baseline test score improves efficiency due to the autocorrelation between test-scores across multiple periods.[37] All regressions include a set of mandal-level dummies (Z_m) and the standard errors are clustered at the school level. Since the treatments are stratified (and balanced) by mandal, including mandal fixed effects increases the efficiency of the estimate. 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 if it was in the incentive treatment, and the parameter of interest is δ, which is the effect on the normalized test scores of being in an incentive school. The random assignment of treatment ensures that the 'Incentives' variable in the equation above is not correlated with the error term, and the estimate is therefore unbiased. The estimate of γ does not matter for the estimate of δ in specifications with Y0 scores on the right-hand side (since Y0 scores are balanced across treatments at the point of randomization), but it does matter in specifications with Y1 scores on the right hand side (since Y1 scores are a post-treatment outcome).

In other words, the extent of depreciation of test scores over time (implied by any value γ less than 1) impacts the estimate of the "gross" treatment effect beyond the first year of the program, because some of the gains in the first year are depreciated during subsequent years of the program. However, γ cannot be consistently estimated in the above specification due to downward bias from measurement bias and upward bias from

[37] Since grade 1 children did not have a baseline test, we set the normalized baseline score to zero for these children. All results are robust to completely excluding grade 1 children as well.

omitted individual ability.[38] Thus, the specification in (5.1) can be used to consistently estimate the one-year and two-year effect of the program, but not the second year effect alone.

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 program, and 0.22 SD higher at the end of the second year (Table 2 – Panel A, columns 1 and 5). The impact of the incentives at the end of two years is greater in math (0.28 SD) than in language (0.16 SD) and this difference is significant (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 3 and 4 of Table 2 show the results of estimating equation (5.1) with Y2 scores on Y1 scores. While, this estimate is biased (as mentioned earlier), Andrabi et al (2008) show that the biases from measurement error and omitted heterogeneity roughly cancel each other out in a similar context (primary education in Pakistan), in which case, the OLS estimate of γ would not be too different from the bias-corrected estimate. Taking these numbers as illustrative suggests that the "gross" treatment effect of the incentive programs was comparable across both years (0.15 SD and 0.14 SD). However, the two-year treatment effect of 0.22 SD is not the sum of these two effects because of depreciation of prior gains and the difference between the two-year effects and the one-year effect (of 0.07 SD) should be interpreted as a "net" treatment effect.[39]

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. Both these cases were dropped from

[38] See Andrabi, Das, Khwaja, and Zajonc (2008) for a detailed exposition of the relevant issues in the estimation of the coefficient on lagged test scores in modeling educational attainment.

[39] Thus the 2-year treatment effect would be the sum of the gross treatment effects over the years less the amount of first year gains that are depreciated away. So in Table 2, 0.15 + 0.14 – (0.45 * 0.15) is roughly equal to 0.22. The issue of depreciation of learning has not received much attention in the literature on the effects of education interventions on test scores over multiple years, but the important point to note is that the net treatment effect will typically underestimate the impact of the treatment beyond the first year.

the analysis and the concerned schools/teachers were declared ineligible for bonuses. Appendix B describes both the testing procedure and robustness checks for cheating.

Figure 2a (left panel) plots the density of the gain in normalized test scores $(T_{ijkm}(EL) - T_{ijkm}(BL))$ over 2 years for control and incentive schools. Figure 2a (right panel) shows the cdf of the same distributions, and that the distribution of gains in the incentive schools first order stochastically dominates that of the control school distribution. Figure 2b plots the quantile treatment effects of the performance pay program on student test scores. Formally, the quantile treatment effect $\delta(\tau)$ in the binary treatment case is defined 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.[40] In the figure, this is the vertical gap between the cumulative distributions of test scores by treatment, which is positive at every percentile and increasing. In other words, test scores in incentive schools are higher at every percentile, but the incentive program also increased the variance of performance, with students at the high end of the test score distribution at the end of two years gaining the most.[41]

5.4 Robustness of results across sub-groups

In addition to the overall effects of the incentives, we check the robustness of the results by looking at various sub-groups and seeing if the effects are equally present across sub-groups or if they are concentrated among certain groups of students. The general specification used for testing treatment effects by sub-group was:

$$T_{ijkm}(EL) = \alpha + \gamma \cdot T_{ijkm}(BL) + \sum_{i=1}^{n} \delta_i \cdot (Incentives \times Category_i) + \beta \cdot Z_m + \varepsilon_k + \varepsilon_{jk} + \varepsilon_{ijk}$$

followed by an F-test of equality across the δ_i's. The estimate of δ is positive for every grade in both years (Table 3, columns 1 and 2). Similarly, the treatment effects are also positive for all five project districts.

[40] See Koenker (2000)

[41] Note that Figure 2b plots the endline score distributions and therefore does not say anything about how the treatment effects depended on initial baseline scores. Section 5.5 shows that the interaction between the incentive program and the baseline scores was not significant, suggesting that the program effectiveness did not vary by initial learning levels.

In addition to being significantly positive for both math and language, the gains in the incentive schools are robustly present across various sub-categories of the tests. Table 4 breaks down the results by 'lower endline' (which covered previous year competencies tested in the baseline) and 'higher endline' (which covered current school year competencies) and shows that the test score gains were significant in both rounds of testing. The gains in the higher endline are greater than those in the lower endline (though not significantly so), which is consistent with our finding from teacher interviews that teachers report spending over 80% of their time on covering the syllabus from the present school year and less than 20% reviewing material from previous years.

We also check for robustness of the gains across the range of difficulty of questions. The left panel of Figure 3b pools all 406 questions (across all tests from Y1) and sorts them by difficulty (as measured by the fraction correct in the control schools). We see that the questions covered a full range of difficulty, and also see that the incentive schools did better than the control schools in over 80% of the questions. The right panel plots the question-level difference between incentive and control schools against the difficulty of the question, and both the intercept and slope in that regression are positive and significant. So, incentive schools perform better on questions of all difficulty, and the performance difference relative to control schools increases with question difficulty. We find the same pattern in the second year as well.

5.5 Heterogeneity of Treatment Effects

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

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

Table 5 (Panel A) shows the results of these regressions on several school and household characteristics. 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. [42] We also find no evidence of a significant difference in the

[42] Given the presence of several covariates in Table 5, caution should be exercised to avoid data mining for differential treatment effects since a few significant coefficients are likely simply due to sampling

effect of the incentives on 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 more affluent families showing a higher response to the teacher incentive program at both the end of the first year and the end of the second year.

We find no evidence of differential impact of incentives by teacher characteristics such as gender, designation, education, or training. However, we find that teachers with higher base pay respond less well to the incentives (Table 5 – Panel B, column 4), which suggests that the magnitude of the incentive was salient because the potential incentive amount (for which all teachers had the same contract) would have been a larger share of base pay for lower paid teachers. However, teachers with higher base pay are typically more experienced and we also see that more experienced teachers respond less well to the incentives (column 3). So, while this evidence is suggestive that the magnitude of the bonus matters, it is also consistent with an interpretation that young teachers respond better to *any* new policy initiative (including performance pay), and so we cannot distinguish the impact of the incentive amount from that of other teacher characteristics that influence the base pay.

5.6 Mechanical versus Conceptual Learning and Non-Incentive Subjects

To test the impact of incentives on these two kinds of learning, we again use specification (5.1) but run separate regressions for the mechanical and conceptual parts of the test. Incentive schools do significantly better on both the mechanical and conceptual components of the test and the estimate of δ is almost identical across both components (Table 6).[43] The other interesting result is that the coefficient on the baseline score is significantly lower for the conceptual component than for the mechanical component (in both years), indicating that these questions were more unfamiliar relative to the mechanical questions. The relative unfamiliarity of these questions increases our

---

variability. Thus, consistent evidence of heterogeneous treatment effects across multiple years provides a better test.

[43] The score on each component is normalized by the mean and standard deviation of the control school distribution for that component. Since the variance of the mechanical component is larger, normalizing by the standard deviation of the total score distribution would show that the magnitude of improvement due to incentives was larger on the mechanical component.

confidence that the gains in test scores represent genuine improvements in learning outcomes.

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)

$$T_{ijkm}(EL_{EVS}) = \alpha + \gamma_1 \cdot T_{ijkm}(BL_{Math}) + \gamma \cdot T_{ijkm}(BL_{Language}) + \delta \cdot Incentives + \beta \cdot Z_m + \varepsilon_k + \varepsilon_{jk} + \varepsilon_{ijk}$$

where lagged scores on both math and language are included, and again the parameter of interest is δ (EVS stands for "Environmental Sciences" which include both science and social studies). Students in incentive schools performed significantly better on non-incentive subjects as well scoring 0.11 and 0.18 standard deviations higher than students in control schools in science and social studies respectively at the end of two years (Table 7). The coefficients on the lagged baseline math and language scores here are much lower than those in Tables 2 and 6, confirming that the domain of these tests was substantially different from that of the tests on which incentives were paid.

These results do not imply that no diversion of effort away from EVS or conceptual thinking took place, but rather that in the context of primary education, teacher efforts aimed at increasing test scores in math and language also contribute to superior performance on broader educational outcomes suggesting complementarity in the various measures and positive spillover effects (though the result could also be due to an improvement in test-taking skills that transfer across subjects).

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 8, columns 1 and 7).[44] 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. However, by the end of the second year, students in individual incentive schools scored 0.27 SD higher than those in comparison schools, while those in group incentive schools scored 0.16 SD higher, with this difference being close to significant at the 10% level

[44] Table 8 is estimated with specification (5.1) but separating out the "incentive" treatment into group and individual incentives respectively.

(column 7). Estimates of the treatment effect in the second year alone (column 4) suggest that individual incentive schools significantly outperformed group incentive schools in the second year.

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 mean number of teachers across the 100 schools was 3.28, the median was 3, and the mode was 2). The limited range of school size makes it difficult to precisely estimate the impact of group size on the effectiveness of group incentives.

Our results are relevant not just for AP's rural schools but for rural schools throughout India because the government's access policy makes small schools common. Since educationists emphasize the value of cooperation within the school and the harmful effects of within-school competition, it is useful to know that both the group and the individual incentive programs had significant positive impacts in this context of rural schools with a small number of teachers. They were also equally cost effective (see section 7). However, the result should not be extrapolated to large schools in which the group would comprise many teachers.

6. Teacher Behavior and Classroom Processes

As described in section 3.3, the APF mandal coordinators (MCs) conducted several rounds of unannounced tracking surveys during the two school years to all 500 schools To code classroom processes, an MC 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. The MCs 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 endline 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. This is similar to the results in Glewwe et al (2003) who find no difference in teacher behavior 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 endline results. 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 9). While self-reported measures of teacher activity might be considered less credible than observations, we find a positive and significant correlation between nearly all the reported activities of teachers and the performance of their students (Table 9 – 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 9).

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.

Our use of both direct observations and interviews might help in reconciling the difference between the findings of Glewwe et al. (2003) and Lavy (2007) with respect to teacher behavior. Glewwe et al. use direct observation and report that there was no significant difference in teacher actions between incentive and comparison schools; Lavy

uses phone interviews with teachers and reports that teachers in incentive schools were significantly more likely to conduct extra classes, stream students by ability, and provide extra help to weak students. While both methods are imperfect, our results suggest that the difference between the studies could partly be due to the different methodologies used for measuring classroom process variables. 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 presence/absence margin on a given day during the school year, but that it probably made them exert more effort when present.

7. Comparison with Input Treatments & Cost-Benefit Analysis

To compare the effects across treatment types, we pool the 2 incentive treatments, the 2 input treatments shown in Table 3.1, and the control schools and run the regression:

$$T_{ijkm}(EL) = \alpha + \gamma \cdot T_{ijkm}(BL) + \delta_1 \cdot Incentives + \delta_2 \cdot Inputs + \beta \cdot Z_m + \varepsilon_k + \varepsilon_{jk} + \varepsilon_{ijk}$$

using the full sample of 500 schools. While both categories of treatments had a positive and significant impact on learning outcomes at the end of the first year, the incentive schools performed 0.06 standard deviations better than the input schools and this difference is significant at the 10 percent level (Table 10 - Column 1). At the end of two years, the difference is more pronounced with the incentive schools scoring 0.13 SD higher and this difference is significant at the 1% level (Table 10 – Column 7). 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 significantly lower amounts per school. The average annual spending 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. 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 gains below than 5% (below 0% in the second year) 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 (i.e. even conditional on the same distribution of

scores, the individual incentive pay out would be higher as long as there are some classes with negative gains because of truncation of teacher-level bonuses at zero in the individual incentive calculations).[45]

Thus, 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 (0.27 SD vs. 0.09 SD). The group incentive program had a smaller treatment effect than the individual incentive program (0.16 SD vs 0.27 SD), but on a cost effectiveness basis the group and individual incentive programs were almost identical in their effectiveness (0.16 SD for Rs. 6,000 in the GI schools and 0.27 SD for Rs. 10,000 in the II schools). Thus, both the incentive programs significantly outperformed the input programs and were roughly equal to each other in cost effectiveness.

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 component with an expected payout of 3% of base pay in lieu of a standard increase across the board. 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.[46] This is less than 10% of the mean incentive payment (3% of base pay) and thus, the long-run cost of the incentive program can be substantially lower than the full cost of the bonuses paid in the short

[45] The negative gains at the student-level were capped at -5% for the calculation of teacher bonuses. Recall that the penalty for a student dropping out was -5%. 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.

[46] 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.

run.[47] 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 (which would need to be compensated for by paying a risk-premium), and more due to heterogeneity in ability, which would attract higher-ability teachers into the profession, and reduce the rents paid to less effective teachers (see next section).

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 administering and grading the tests. The approximate cost of each 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.[48] This is probably an over estimate because the small size and geographical dispersion of the program was not conducive to exploiting economies of scale. 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 the context of a scheduled across the board increase in pay.

8. Teacher Opinions on Performance Pay

Nearly 75% of teachers in incentive schools report that their motivation levels went up as a result of the program (with the other 25% reporting no change); over 95% had a favorable opinion about the program; over 85% had a favorable opinion regarding the idea of providing bonus payments to teachers on the basis of performance; and over two thirds of teachers felt that the government should consider implementing a system of bonus payments on the basis of performance.

Of course, it is easy to support a program when it only offers rewards and no penalties, and so we also asked the teachers their opinion regarding performance-pay in a *wage-neutral* way. Teachers were asked their preference regarding how they would

[47] This would not be true for the current teachers in the system who are used to low levels of effort, and who would need to be compensated not only for the risk of variable pay but for the extra effort that they may need to exert under a performance-pay system. However, new teachers who are not accustomed to the rents of the civil-service job would only need to be compensated for the risk premium.

[48] The first year of the program required two rounds of testing, but subsequent years only require one round since the endline test in a given year will serve as the baseline for the next year.

allocate a hypothetical budgetary allocation for a 15% pay increase between an across-the-board increase for all teachers, and a performance-based component. Over 75% of teachers supported the idea of at least some performance-based pay, with over 20% in favor of having 20% or more of annual pay determined by performance.[49]

The longer-term benefits to performance pay include not only greater teacher effort, but also potentially the entry of better teachers into the profession.[50] We regress the extent of teachers' preference for performance pay holding expected pay constant (reported before they knew their outcomes) on the average test score gains of their students and find a positive and significant correlation between teacher performance and the extent of performance pay they desire. This suggests that effective teachers know who they are and that there are likely to be sorting benefits from performance pay. If the teaching community is interested in improving the quality of teachers entering the profession, this might be another reason to support performance pay.[51]

9. Conclusion

Performance pay for teachers is an idea with strong proponents as well as opponents and the empirical evidence to date on its effectiveness has been mixed. 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. The gains were spread out evenly across grades, districts, skills and competencies, and question difficulty. We detect no adverse consequences of the program with student performance improving on mechanical as well as conceptual questions and on incentive as well as non-incentive subjects. There was no difference

[49] If teachers are risk-averse and have rational expectations about the distribution of their abilities, we would expect less than 50% to support revenue-neutral performance pay since there is no risk premium being offered in the set of options. The 75% positive response could reflect several factors including over optimism about their own abilities, a belief that it will be politically more feasible to secure funds for salary increases if these are linked to performance, or a sense that such a system could bring more professional respect to teachers and enhance motivation across the board.

[50] See Lazear (2000) and (2003), and Hoxby and Leigh (2005)

[51] Ballou and Podgursky (1993) show that teachers' attitude towards merit pay in the US is more positive than is supported by conventional wisdom and argue that the dichotomy may be due to divergence between the interests of union leadership and members. There is some evidence that this might be the case here as well. Older teachers are significantly less likely to support the idea of performance pay in our data, but they are also much more likely to be active in the teacher union.

between the effectiveness of group versus individual incentives in the first year of the programs, but the individual incentive schools showed higher gains in the second year. Teacher absence did not differ across treatments, but teachers in incentive schools appear to teach more effectively when present.

The additional inputs we studied (para-teachers and school grants for spending on student-level inputs) were also effective in raising test scores. However, the incentive program spent the same amount of money on bonus payments and achieved significantly better outcomes. Since the long-term cost of performance pay is only the risk premium associated with variable pay and the administrative cost, teacher incentive programs may be a highly cost effective way of improving learning outcomes. Teachers supported the idea of performance pay (even holding expected pay constant) with over 85% of them being in favor of the idea of bonus payments on the basis of test score improvements.

The significant effect of teacher performance pay on learning outcomes over both the one-year and two-year horizon of the program suggests that the program effects are unlikely to be due to its novelty. The continued gains on both mechanical and conceptual test questions as well as on non-incentive subjects indicate that the distortions from multi-tasking are less of a concern at very low levels of learning. Finally, the cost effectiveness of teacher performance pay relative to the other input-based interventions makes teacher performance pay a highly promising policy option for improving education outcomes.

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. It is possible that the first two years of the program saw large performance effects from relatively small bonus payments because teachers have consumption commitments and so even small bonuses represent large increases in utility. However, if an expected bonus gets factored into annual expected income, the impact on teacher effort and student performance might be smaller in future years. An expectation of bonuses may also lead to political pressure on

the part of teachers to convert bonuses into raises, which would defeat the point of performance-based pay.

We have also not devised or tested the optimal long-term formula for teacher incentive payments. While basing bonuses on average test score gains is an improvement on simply using levels and avoids the problems associated with level targets, the current formula is not optimal. For instance, the fact that the coefficient on the lagged score (γ) is estimated as close to 0.5 as opposed to 1, suggests that the naïve 'average gain' formulation penalized teachers who happened to be in classes that had high baseline scores. There are also other determinants of student performance such as class size, school infrastructure, household inputs, and peer effects. An optimal formula for teacher bonuses would net out these factors to estimate a more precise measure of teachers' value addition. A related concern is measurement error and the potential lack of reliability of test scores at the class and school levels.[52]

The incentive formula can be improved with data over multiple years and by drawing on the growing literature on estimating teacher value added in developed countries.[53] However, there is a practical trade off between the accuracy and precision of the value-added measurement on one hand and the transparency of the system to teachers on the other. Teachers accepted the intuitive 'average gain' formula used in the first two years of the program. We expect that teachers will start getting more sophisticated about the formula in future years, 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.[54]

---

[52] Kane and Staiger (2002) show that measurement error in class-level and school-level averages can lead to rankings based on these averages being volatile. However, as Rogosa (2005) points out, mean test-scores can be quite *precise* (in the sense of accurately estimating levels of learning) even while not being very *reliable* (in the sense of accurately ranking schools). This might be a reason to prefer contracts over tournaments.

[53] See the excellent collection of essays in Haertel and Herman (2005) for instance.

[54] 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.

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 (in a representative sample of schools in the Indian state of Andhra Pradesh) that even modest amounts of performance-based pay for teachers can lead to substantial improvements in student learning outcomes, with limited negative consequences. As school systems around the world consider adopting various forms of performance pay, attempts should be made to build in rigorous impact evaluations of these programs during a process of slowly expanding coverage. This will also allow experimentation with variations such as basing bonuses on both subjective and objective measures of performance, and putting weight on both group and individual-level performance. Studies should 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.

References:

ANDRABI, T., J. DAS, A. KHWAJA, and T. ZAJONC (2008): "Do Value Added Measures Add Value? Accounting for Learning Dynamics," Harvard University.

ATKINSON, A., S. BURGESS, B. CROXSON, P. GREGG, C. PROPPER, H. SLATER, and D. WILSON (2004): "Evaluating the Impact of Performance-Related Pay for Teachers in England," Department of Economics, University of Bristol, UK, The Centre for Market and Public Organisation, 60.

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.

BALLOU, D., and M. PODGURSKY (1993): "Teachers' Attitudes toward Merit Pay: Examining Conventional Wisdom," *Industrial and Labor Relations Review*, 47, 50-61.

BANDIERA, O., I. BARANKAY, and I. RASUL (2006): "Incentives for Managers and Inequality among Workers: Evidence from a Firm Level Experiment," Center for Economic Policy Research.

BANERJEE, A., S. COLE, E. DUFLO, and L. LINDEN (2005): "Remedying Education: Evidence from Two Randomized Experiments in India," National Bureau of Economic Research Inc NBER Working Papers: 11904.

BARON, J. N., and D. M. KREPS (1999): *Strategic Human Resources: Frameworks for General Managers*. New York: John Wiley.

BRENNAN, G., and P. PETTIT (2005): *The Economy of Esteem : An Essay on Civil and Political Society*. Oxford ; New York: Oxford University Press.

CHIAPPORI, P.-A., and B. SALANIÉ (2003): "Testing Contract Theory: A Survey of Some Recent Work," in *Advances in Economics and Econometrics*, ed. by M. Dewatripont, L. P. Hansen, and S. J. Turnovsky. Cambridge, UK: Cambridge University Press.

DE LAAT, J., and E. VEGAS (2005): "Do Differences in Teacher Contracts Affect Student Performance? Evidence from Togo," World Bank.

DECI, E. L., and R. M. RYAN (1985): *Intrinsic Motivation and Self-Determination in Human Behavior*. New York: Plenum.

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

DONNER, A., and N. KLAR (2000): *Design and Analysis of Cluster Randomization Trials in Health Research*. London, New York: Arnold; Co-published by the Oxford University Press.

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 (2006): "Individual Teacher Incentives and Student Performance," Cambridge, MA: National Bureau of Economic Research Inc

FIGLIO, D. N., and J. WINICKI (2005): "Food for Thought: The Effects of School Accountability Plans on School Nutrition," *Journal of Public Economics*, 89, 381-94.

FREY, B. S., and F. OBERHOLZER-GEE (1997): "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out," *American Economic Review*, 87, 746-55.
GIBBONS, R. (1998): "Incentives in Organizations," *Journal of Economic Perspectives*, 12, 115-32.
GLEWWE, P., N. ILIAS, and M. KREMER (2003): "Teacher Incentives," Cambridge, MA: National Bureau of Economic Research.
GNEEZY, U., and A. RUSTICHINI (2000): "Pay Enough or Don't Pay at All," *Quarterly Journal of Economics* 115, 791-810.
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.
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.
— (1990): "Regulating Trade among Agents," *Journal of Institutional and Theoretical Economics*, 146, 85-105.
— (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. (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.
— (2006): "Teacher Absence in India," Harvard University.
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.
— (2007): "Performance Pay and Teachers' Effort, Productivity and Grading Ethics," Hebrew University
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.
MULLAINATHAN, S. (2006): "Development Economics through the Lens of Psychology," Harvard University.
MURALIDHARAN, K., and V. SUNDARARAMAN (2008): "Contract Teachers," Harvard.
MURALIDHARAN, K., and V. SUNDARARAMAN (2008): "The Impact of School Block Grants on Student Learning Outcomes," Harvard.
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., and D. SCHANZENBACH (2008): "Left Behind by Design: Proficiency Counts and Test-Based Accountability" University of Chicago.
OYER, P. (1998): "Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality," *Quarterly Journal of Economics*, 113, 149-85.
PRATHAM (2008): *Annual Status of Education Report*.
PRENDERGAST, C. (1999): "The Provision of Incentives in Firms," *Journal of Economic Literature*, 37, 7-63.
PRITCHETT, L., and D. FILMER (1999): "What Education Production Functions Really Show: A Positive Theory of Education Expenditures," *Economics of Education Review*, 18, 223-39.
PRITCHETT, L., and V. PANDE (2006): "Making Primary Education Work for India's Rural Poor: A Proposal for Effective Decentralization," New Delhi: World Bank.
ROGOSA, D. (2005): "Statistical Misunderstandings of the Properties of School Scores and School Accountability," in *Uses and Misuses of Data for Educational Accountability and Improvement*, ed. by J. L. Herman, and E. H. Haertel: Blackwell Synergy, 147-174.
TODD, P. E., and K. I. WOLPIN (2003): "On the Specification and Estimation of the Production Function for Cognitive Achievement," *Economic Journal*, 113, F3-33.

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.

Appendix A: Project Timeline and Activities

The broad timeline of AP RESt was as follows:

January 2004 – October 2004:	Planning, Permissions, Partner selection, Funding
November 2004 – April 2005:	Pilot
April 2005 – June 2006:	First full year of main interventions
June 2006 – June 2007:	Second full year of interventions

Main Project (Timeline of Key Activities)

April – June 2005

- Random sampling of the 500 schools to comprise the universe of the study
- Communication of the details of baseline testing to the various district-level officials in the selected districts (*only communicated about the baseline tests and not about the inputs and incentives at this point*)

Late June – July 2005

- Baseline tests conducted in all 500 project schools in a 2-week span in early July
- Scoring of tests and preparation of school and class performance reports
- Stratified random allocation of schools to treatments groups

August 2005

- Distribution of test results, diagnostics, and announcement of relevant incentive schemes in selected schools
- Treatment status and details communicated to schools verbally and in writing

September 2005

- Placement of extra teacher in the relevant randomly selected schools
- Provision of block grants to the relevant randomly selected schools, procurement of materials and audit of procurement

September 2005 – February 2006

- Unannounced tracking surveys of all 500 schools on average once a month
- From December, similar visits were made to additional "pure control" schools

March – April 2006

- Lower and higher endline assessments conducted in 500 schools plus a 100 extra schools that constitute the pure control category

August 2006

- Interviews with teachers on teaching activities in the previous school year and on their opinion about performance pay

September 2006

- Provision of school and class level performance reports
- Provision of incentive payments to qualified schools and teachers
- Communication letters about the second year of the program and repeat of above processes for the second year of the program

Appendix B: Project Team, Test Administration, and Robustness to Cheating

The project team from the Azim Premji Foundation consisted of around 30 full time staff and 250 to 300 evaluators hired for the period of the baseline and endline testing. The team was led by a project manager, and had 5 district coordinators (DCs) and 25 mandal coordinators (MCs). Each MC was responsible for project administration, supervision of independent tests, communications to schools, and conducting tracking surveys in 2 mandals (20 schools). The MCs were the 'face' of the project to the schools, while each DC was responsible for overall project implementation at the district level.

Teams of evaluators were hired and trained specially for the baseline and endline assessments. Evaluators were typically college graduates who were obtaining a certificate or degree in teaching. The tests were externally administered with teams of 5 evaluators conducting the assessments in each school (1 for each grade). For the baseline there were 50 teams of 5 evaluators with each team covering a school in a day. The 500 schools were tested in 10 working days over 2 weeks. For the end of year tests, we had 60 teams of 5 evaluators each and 600 schools (including 100 'pure control' schools) were tested in 2 rounds over 4 weeks at the end of the school year. The 'lower endline' was conducted in the first 2 weeks and the 'higher endline' was conducted in the last 2 weeks. Schools were told that they could be tested anytime in a 2-week window and the order of testing schools was also balanced across treatment categories.

Identities of children taking the test were verified by asking them for their father's name, which was verified against a master list of student data. Standard exam procedures of adequate distance between students and continuous proctoring were followed. The teachers were not allowed in the classes while the tests were being given. The tests (and all unused papers) were collected at the end of the testing session and brought back to a central location at the end of the school day. The evaluation of the papers, and the transcription to the 'top sheet' (that was used for data entry) was done in this central location under supervision and with cross checking across evaluators to ensure accuracy.

No cases of cheating were observed during the first year of the programs, but two cases of cheating were detected in the second year (one classroom and one entire school). These cases were reported to the project management team by the field enumerators, and the schools were subsequently disqualified from receiving any bonus for the second year. These cases are not included in the analysis presented in the paper.

Table 1: Sample Balance Across Treatments

	Panel A (Means of Baseline Variables)			
	[1]	[2]	[3]	**[4]**
	Control	Group Incentives	Individual Incentives	**P-value (Equality of all groups)**
School-level Variables				
Total Enrollment (Baseline: Grades 1-5)	113.2	111.3	112.6	**0.82**
Total Test-takers (Baseline: Grades 2-5)	64.9	62.0	66.5	**0.89**
Number of Teachers	3.07	3.12	3.14	**0.58**
Pupil-Teacher Ratio	39.5	40.6	37.5	**0.66**
Infrastructure Index (0-6)	3.19	3.14	3.26	**0.84**
Proximity to Facilities Index (8-24)	14.65	14.66	14.72	**0.98**
Baseline Test Performance				
Math (Raw %)	18.4	17.8	17.4	**0.72**
Math (Normalized - in Std. deviations)	0.022	-0.003	-0.019	**0.74**
Telugu (Raw %)	35.0	34.8	33.4	**0.54**
Telugu (Normalized - in Std. deviations)	0.019	0.014	-0.032	**0.52**
	Panel B (Means of Endline Variables)			
Teacher Turnover and Attrition				
Year 1 on Year 0				
Teachers Who Stayed the Full Year/ Total in School Beginning of School Year (%)	0.70	0.66	0.69	**0.63**
Teachers Who Stayed the Full Year/ Total in School at End of School Year (%)	0.66	0.67	0.68	**0.90**
Year 2 on Year 1				
Teachers Who Stayed the Full Year/ Total in School Beginning of School Year (%)	0.96	0.94	0.94	**0.53**
Teachers Who Stayed the Full Year/ Total in School at End of School Year (%)	0.95	0.96	0.97	**0.37**
Year 2 on Year 0				
Teachers Who Stayed the Full Year/ Total in School Beginning of School Year (%)	0.68	0.63	0.66	**0.47**
Teachers Who Stayed the Full Year/ Total in School at End of School Year (%)	0.63	0.64	0.67	**0.62**
Student Turnover and Attrition				
Year 1 on Year 0				
Student Attrition (Students who did not take an endline test as a fraction of those who took a baseline test)	0.082	0.066	0.071	**0.20**
Baseline Maths test score of attritors (Normalized)	-0.16	-0.15	-0.19	**0.95**
Baseline Telugu test score of attritors (Normalized)	-0.26	-0.19	-0.25	**0.81**
Year 2 on Year 0				
Student Attrition (Students who did not take an endline test as a fraction of those who took a baseline test)	0.26	0.24	0.25	**0.70**
Baseline Maths test score of attritors (Normalized)	-0.14	-0.07	-0.09	**0.73**
Baseline Telugu test score of attritors (Normalized)	-0.19	-0.14	-0.20	**0.78**

Notes:

1. The infrastructure index sums 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 sums 8 variables (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. The t-statistics 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 t-statistics are computed treating each school as an observation.

Table 2: Impact of Incentives on Student Test Scores

	Year 1 on Year 0		Year 2 on Year 1		Year 2 on Year 0	
Panel A: Combined Across Subjects						
Dependent Variable = Normalized End of Year Test Score						
	[1]	[2]	[3]	[4]	[5]	[6]
Normalized Lagged Test Score	0.5	0.5	0.553	0.572	0.45	0.45
	(0.013)***	(0.013)***	(0.016)***	(0.018)***	(0.015)***	(0.015)***
Incentive School	0.153	0.175	0.143	0.124	0.217	0.226
	(0.042)***	(0.042)***	(0.035)***	(0.042)***	(0.047)***	(0.048)***
School and Household Controls	No	Yes	No	Yes	No	Yes
Observations	68702	64364	78613	48074	49516	45556
R-squared	0.29	0.31	0.29	0.36	0.23	0.24
Panel B: Math						
Dependent Variable = Normalized End of Year Test Score						
	[1]	[2]	[3]	[4]	[5]	[6]
Normalized Lagged Test Score	0.49	0.495	0.496	0.512	0.418	0.417
	(0.017)***	(0.017)***	(0.021)***	(0.025)***	(0.022)***	(0.023)***
Incentive School	0.188	0.211	0.197	0.179	0.277	0.286
	(0.049)***	(0.050)***	(0.042)***	(0.052)***	(0.055)***	(0.056)***
School and Household Controls	No	Yes	No	Yes	No	Yes
Observations	34121	31970	39238	24000	24592	22621
R-squared	0.28	0.3	0.27	0.33	0.22	0.24
Panel C: Telugu						
Dependent Variable = Normalized End of Year Test Score						
	[1]	[2]	[3]	[4]	[5]	[6]
Normalized Lagged Test Score	0.516	0.513	0.615	0.637	0.484	0.482
	(0.014)***	(0.014)***	(0.012)***	(0.014)***	(0.014)***	(0.014)***
Incentive School	0.119	0.14	0.09	0.072	0.158	0.167
	(0.038)***	(0.038)***	(0.032)***	(0.037)*	(0.043)***	(0.044)***
School and Household Controls	No	Yes	No	Yes	No	Yes
Observations	34581	32394	39375	24074	24924	22935
R-squared	0.32	0.34	0.33	0.42	0.25	0.26

Notes:

1. All regressions include mandal (sub-district) fixed effects and standard errors clustered at the school level.
2. Constants are insignificant in all specifications and are not shown.

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

Table 3: Impact of Incentives by Grade

	Dependent Variable = Normalized Endline Test Score					
	Combined		**Math**		**Telugu (Language)**	
	Y1 on Y0	**Y2 on Y0**	**Y1 on Y0**	**Y2 on Y0**	**Y1 on Y0**	**Y2 on Y0**
	[1]	[3]	[4]	[6]	[7]	[9]
Incentives * Grade 1	0.102		0.106		0.098	
	(0.06)		(0.07)		(0.07)	
Incentives * Grade 2	0.107	0.14	0.12	0.166	0.095	0.115
	(0.054)**	(0.057)**	(0.058)**	(0.068)**	(0.06)	(0.053)**
Incentives * Grade 3	0.173	0.171	0.211	0.222	0.136	0.121
	(0.055)***	(0.056)***	(0.062)***	(0.068)***	(0.053)**	(0.053)**
Incentives * Grade 4	0.182	0.181	0.245	0.23	0.121	0.134
	(0.054)***	(0.061)***	(0.067)***	(0.070)***	(0.048)**	(0.057)**
Incentives * Grade 5	0.153	0.342	0.184	0.448	0.123	0.237
	(0.051)***	(0.065)***	(0.063)***	(0.081)***	(0.048)**	(0.058)***
Observations	68275	49516	33908	24592	34367	24924
F-Test p-value (Equality Across Grades)	0.679	0.011	0.303	0.005	0.971	0.119
R-squared	0.29	0.23	0.28	0.23	0.32	0.25

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 4: Impact of Incentives by Testing Round

	Dependent Variable = Normalized Endline Test Score					
	Combined		**Math**		**Telugu (Language)**	
	Y1 on Y0	**Y2 on Y0**	**Y1 on Y0**	**Y2 on Y0**	**Y1 on Y0**	**Y2 on Y0**
	[1]	[2]	[3]	[4]	[5]	[6]
Lower endline	0.003	0.029	-0.01	0.035	-0.003	-0.039
	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)
Higher endline	-0.009	-0.057	-0.013	-0.053	0.007	-0.023
	(0.04)	(0.04)	(0.03)	(0.05)	(0.03)	(0.03)
Incentive School*Lower endline	0.126	0.195	0.154	0.263	0.099	0.128
	(0.044)***	(0.053)***	(0.050)***	(0.065)***	(0.042)**	(0.049)***
Incentive School*Higher endline	0.180	0.238	0.221	0.290	0.141	0.187
	(0.050)***	(0.053)***	(0.062)***	(0.065)***	(0.043)***	(0.049)***
Chow Test (equality of interactions)	0.187	0.396	0.211	0.701	0.256	0.184
Observations	68275	49516	33908	24592	34367	24924
R-squared	0.29	0.23	0.28	0.22	0.32	0.25

Notes:

1. All regressions include mandal (sub-district) fixed effects and standard errors clustered at the school level.
Constants are insignificant in all specifications and not shown

2. The Lower Endline test covered competencies tested in the baseline (corresponding to the previous school year's materials), while the Higher Endline covered the materials taught in the current school year.
* significant at 10%; ** significant at 5%; *** significant at 1%

Table 5: Heterogenous Treatment Effects

Panel A: Household and School Characteristics

	[1] Number of Students in School	[2] Proximity (8 - 24)	[3] Infrastructure (0 - 6)	[4] Household Affluence (0 - 7)	[5] Parental Literacy	[6] SC/ST	[7] Gender	[8] Baseline Score
	Year 2 on Year 0							
Incentive	-0.188	-0.129	0.2	0.105	0.215	0.231	0.252	0.217
	(0.37)	(0.22)	(0.14)	(0.07)	(0.060)***	(0.048)***	(0.046)***	(0.047)***
Covariate	-0.075	-0.006	0.016	0.009	0.027	-0.054	0.015	0.453
	(0.05)	(0.01)	(0.04)	(0.01)	(0.006)***	(0.04)	(0.03)	(0.025)***
Interaction	**0.09**	**0.025**	**0.006**	**0.034**	**0.001**	**-0.013**	**-0.011**	**-0.005**
	(0.07)	**(0.015)***	**(0.04)**	**(0.017)****	**(0.01)**	**(0.06)**	**(0.03)**	**(0.03)**
Observations	49752	49516	49516	46596	46596	46584	46458	49516
R-squared	0.22	0.23	0.23	0.23	0.24	0.23	0.23	0.23
	Year 1 on Year 0							
Incentive	-0.401	-0.033	0.068	0.034	0.153	0.176	0.175	0.15
	(0.39)	(0.16)	(0.11)	(0.06)	(0.053)***	(0.045)***	(0.047)***	(0.042)***
Covariate	-0.115	-0.013	0.004	0.011	0.028	-0.006	0.021	0.502
	(0.059)*	(0.01)	(0.02)	(0.01)	(0.005)***	(0.03)	(0.02)	(0.021)***
Interaction	**0.106**	**0.014**	**0.03**	**0.035**	**0.001**	**-0.067**	**-0.005**	**-0.005**
	(0.08)	**(0.01)**	**(0.03)**	**(0.016)****	**(0.01)**	**(0.05)**	**(0.03)**	**(0.03)**
Observations	68438	66680	66680	65465	65465	65449	41232	68275
R-squared	0.29	0.3	0.3	0.31	0.31	0.3	0.31	0.29

Panel B: Teacher Characteristics

	[1] Education	[2] Training	[3] Experience	[4] Salary (log)	[5] Gender	[6] Absence	[7] Active Teaching	[8] Active or Passive Teaching
	Stacked Regression							
Incentive	0.04	0.005	0.284	1.715	0.204	0.111	0.064	0.015
	(0.16)	(0.16)	(0.062)***	(0.767)**	(0.057)***	(0.051)**	(0.05)	(0.08)
Covariate	0.038	0.026	-0.001	0.044	0.05	-0.114	0.057	0.039
	(0.03)	(0.04)	(0.00)	(0.05)	(0.05)	(0.12)	(0.08)	(0.08)
Interaction	**0.041**	**0.059**	**-0.008**	**-0.169**	**-0.035**	**0.131**	**0.167**	**0.21**
	(0.05)	**(0.06)**	**(0.004)****	**(0.085)****	**(0.07)**	**(0.17)**	**(0.099)***	**(0.125)***
Observations	76698	76698	77115	75615	77337	73523	73523	73523
R-squared	0.28	0.28	0.28	0.28	0.28	0.3	0.3	0.3

Notes:

1. All regressions include mandal (sub-district) fixed effects and standard errors clustered at the school level.
2. Each column in each panel reports one regression testing for heterogeneous treatment effects along the variable indicated in the column
3. For Panel B, teacher characteristics are linked to student test scores in each year and we run a stacked regression using both years' data (Y1 on Y0 and Y2 on Y1) to estimate heterogeneous treatment effects by teacher characteristics

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

Table 6: Impact of Incentives on Mechanical Versus Conceptual Learning

	Dependent Variable = Endline Test Score by Mechanical/Conceptual (Normalized by Mechanical/Conceptual Distribution in Control Schools)			
	Year 1 on Year 0		**Year 2 on Year 0**	
	Mechanical	**Conceptual**	**Mechanical**	**Conceptual**
	[1]	[2]	[3]	[4]
Normalized Baseline Score	0.481	0.336	0.446	0.306
	(0.012)***	(0.011)***	(0.013)***	(0.013)***
Incentive School	0.135	0.135	0.167	0.178
	(0.038)***	(0.043)***	(0.041)***	(0.046)***
Observations	68289	68289	42884	42884
R-squared	0.28	0.17	0.24	0.15

Notes:
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: Impact of Incentives on Non-Incentive Subjects

	Normalized Endline Score			
	Year 1 on Year 0		**Year 2 on Year 0**	
	Science	**Social Studies**	**Science**	**Social Studies**
	[1]	[2]	[3]	[4]
Normalized Baseline Math Score	0.214	0.222	0.155	0.166
	(0.019)***	(0.018)***	(0.023)***	(0.023)***
Normalized Baseline Language Score	0.206	0.287	0.214	0.182
	(0.019)***	(0.019)***	(0.024)***	(0.024)***
Incentive School	0.107	0.135	0.112	0.177
	(0.052)**	(0.047)***	(0.045)**	(0.049)***
Observations	12011	12011	9166	9166
R-squared	0.26	0.3	0.18	0.18

Notes:
Social Studies and Science tests were only administered to grades 3 to 5
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: Impact of Group Incentives versus Individual Incentives

	Dependent Variable = Normalized Endline Test Score								
	Year 1 on Year 0			**Year 2 on Year 1**			**Year 2 on Year 0**		
	Combined	Maths	Telugu	Combined	Maths	Telugu	Combined	Maths	Telugu
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Normalized Lagged Score	0.5	0.49	0.516	0.554	0.497	0.616	0.451	0.418	0.485
	(0.013)***	(0.017)***	(0.014)***	(0.016)***	(0.021)***	(0.012)***	(0.015)***	(0.022)***	(0.014)***
Individual Incentive School (II)	0.16	0.194	0.128	0.198	0.252	0.144	0.271	0.321	0.223
	(0.049)***	(0.060)***	(0.043)***	(0.044)***	(0.052)***	(0.041)***	(0.058)***	(0.068)***	(0.053)***
Group Incentive School (GI)	0.146	0.183	0.11	0.087	0.14	0.035	0.162	0.232	0.092
	(0.050)***	(0.058)***	(0.046)**	(0.045)*	(0.056)**	(0.04)	(0.058)***	(0.071)***	(0.052)*
Observations	68702	34121	34581	78613	39238	39375	49516	24592	24924
F-Stat p-value (Testiing GI = II)	0.78	0.87	0.68	0.05	0.10	0.03	0.12	0.29	0.03
R-squared	0.29	0.28	0.32	0.3	0.27	0.34	0.23	0.23	0.25

Notes:
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 9: Teacher Behavior (Observation and Interviews)

	Incentive versus Control Schools (All figures in %)			
Teacher Behavior	**Incentive Schools**	**Control Schools**	**p-Value of Difference**	**Coefficient of Teacher Behavior Indicator on Student Test Scores**
Based on Observation				
Teacher Absence (%)	0.24	0.24	**0.82**	**-0.110** **
Actively Teaching at Point of Observation (%)	0.44	0.42	**0.57**	**0.124** ***
Based on Interviews				
Did you do any special preparation for the end of year tests? (% Yes)	0.63	0.25	**0.000*** **	**0.102** ***
What kind of preparation did you do? (UNPROMPTED) (% Mentioning)				
Extra Homework	0.42	0.15	**0.000*** **	**0.085** **
Extra Classwork	0.46	0.17	**0.000*** **	**0.091** ***
Extra Classes/Teaching Beyond School Hours	0.16	0.04	**0.000*** **	**0.181** ***
Gave Practice Tests	0.31	0.10	**0.000*** **	**0.111** ***
Paid Special Attention to Weaker Children	0.21	0.05	**0.000*** **	**0.017**

Notes:
Each teacher-year combination is treated as one observation with t-tests clustered at the school
* significant at 10%; ** significant at 5%; *** significant at 1%

Table 10: Impact of Inputs versus Incentives on Learning Outcomes

	Dependent Variable = Normalized Endline Test Score								
	Year 1 on Year 0			**Year 2 on Year 1**			**Year 2 on Year 0**		
	Combined [1]	**Math** [2]	**Language** [3]	**Combined** [4]	**Math** [5]	**Language** [6]	**Combined** [7]	**Math** [8]	**Language** [9]
Normalized Lagged Score	0.511	0.492	0.535	0.552	0.495	0.614	0.461	0.423	0.497
	(0.010)***	(0.012)***	(0.011)***	(0.012)***	(0.016)***	(0.010)***	(0.012)***	(0.016)***	(0.012)***
Incentives	0.155	0.183	0.121	0.145	0.199	0.091	0.217	0.277	0.158
	(0.041)***	(0.049)***	(0.038)***	(0.036)***	(0.044)***	(0.033)***	(0.048)***	(0.056)***	(0.045)***
Inputs	0.096	0.110	0.082	0.047	0.047	0.047	0.084	0.092	0.076
	(0.037)***	(0.042)***	(0.036)**	(0.03)	(0.04)	(0.03)	(0.043)*	(0.049)*	(0.042)*
Difference (Incentives - Inputs)	0.06	0.07	0.04	0.10	0.15	0.04	0.13	0.19	0.08
F-Stat p-value (Inputs = Incentives)	**0.09**	**0.08**	**0.23**	**0.01**	**0.00**	**0.17**	**0.00**	**0.00**	**0.04**
Observations	112238	55542	56269	119836	59820	60016	82596	41053	41543
R-squared	0.29	0.27	0.32	0.29	0.26	0.33	0.21	0.21	0.23

Notes:

1. These regressions pool data from all 500 schools: 'Group' and 'Individual' incentive treatments are pooled together as "Incentives", and the 'extra para-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)

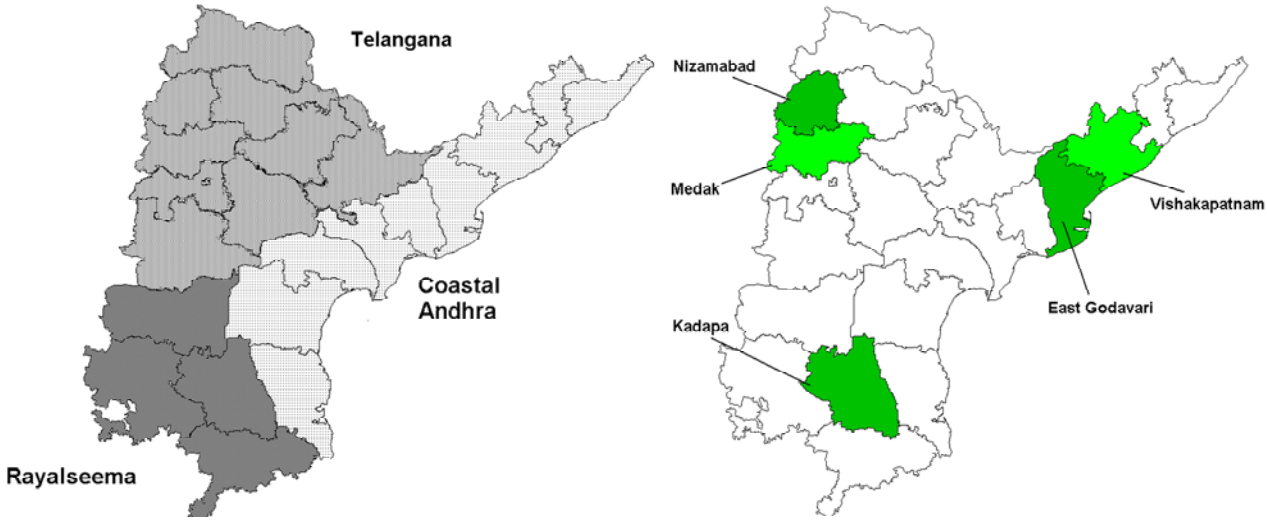


Figure 2a: Density/CDF of Normalized Test Score Gains after 2 Years

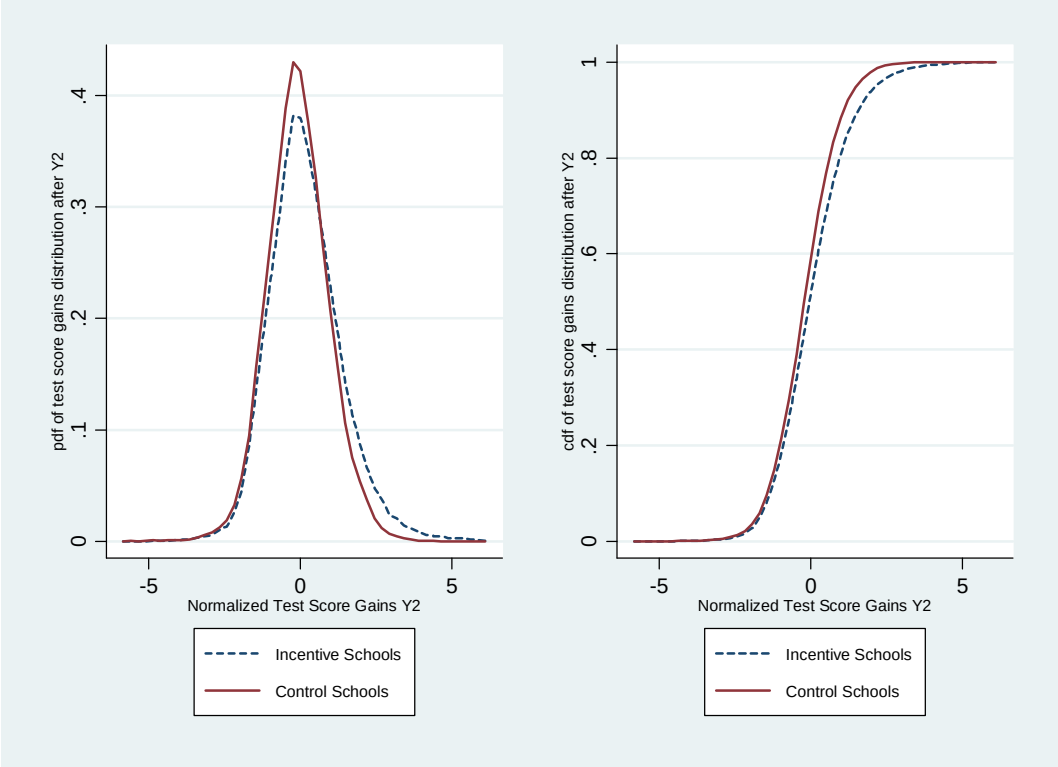


Figure 2b: Quantile Treatment Effects after 2 Years

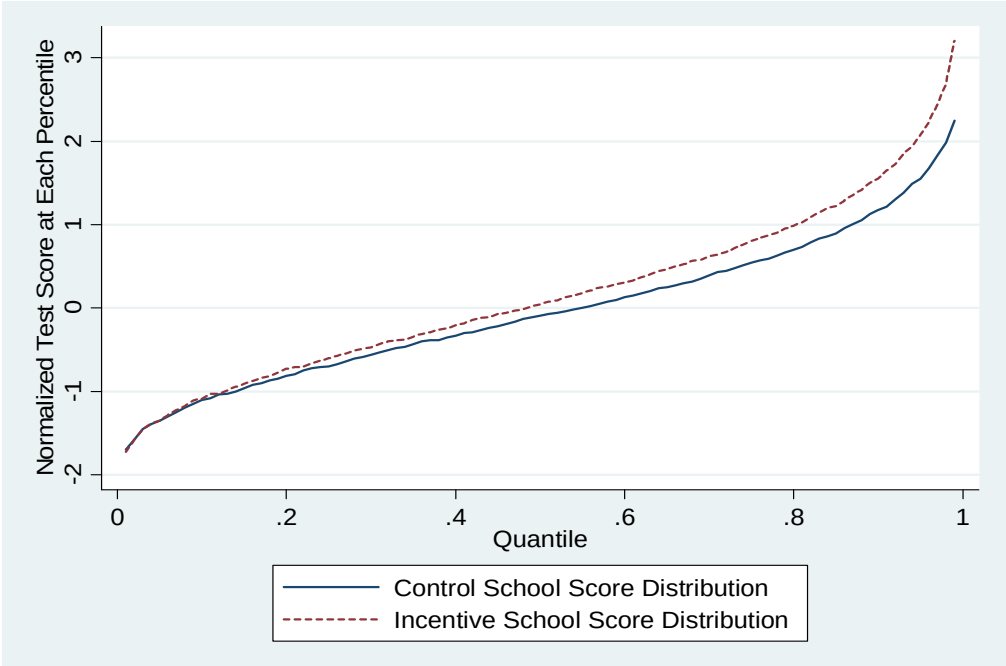


Figure 3a: Test Calibration – Range of item difficulty

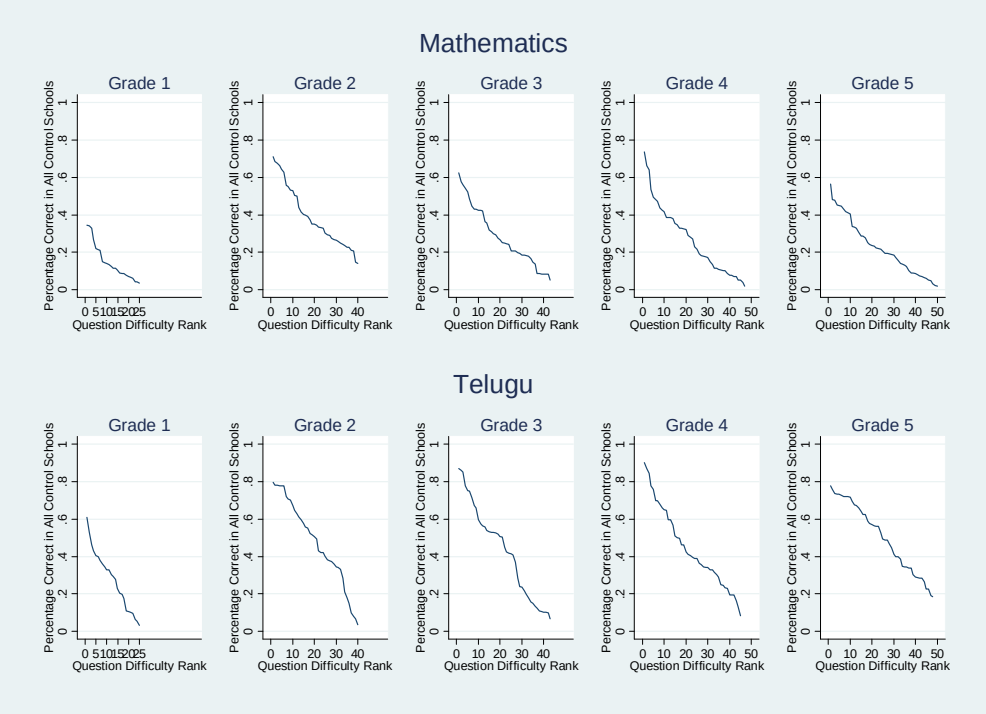


Figure 3b: Incentive versus Control School Performance – By Question Difficulty