Teacher Incentives in Developing Countries: Experimental Evidence from India

Karthik Muralidharan[†]

Venkatesh Sundararaman[‡]

15 November 2006[*]

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). Students in incentive schools performed significantly better than those in control schools by 0.19 and 0.12 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. We find no significant difference in the effectiveness of group versus individual teacher incentives. 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

---

[†] Department of Economics, Harvard University; E-mail: muralidh@fas.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. 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). Karthik Muralidharan thanks the Bradley and Spencer Foundations for fellowship support.

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 recently published *Annual Status of Education Report* found that 52% of children aged 7 to 14 in an all-India sample of nearly 200,000 rural households could not read a simple paragraph of second-grade difficulty, though over 93% of them were enrolled in school (Pratham, 2005).

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) that considered two alternative approaches to improving primary education in the Indian state of Andhra Pradesh (AP).[2] The first was to provide schools with additional "smart inputs" that were believed to be more cost effective than the status quo, and the second was to provide performance-based bonuses to teachers on the basis of the average improvement in test scores of their students. We studied two types of smart inputs (an additional para-teacher; and cash block grants to schools) and two types of incentives (group incentives based on school performance; and individual incentives based on teacher performance), and the additional spending in each of the four programs was calibrated to be slightly over 3% of a typical school's annual budget. The study was

[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.

conducted by randomly allocating the programs across a representative sample of 500 government-run schools in rural AP with 100 schools in each of the four treatment groups and 100 control schools serving as the comparison group. This paper presents results from the first year (2005 – 06) of all four interventions, but focuses on the two teacher incentive programs. Detailed results of the input experiments and estimates of the education production function are presented in a companion paper.

This paper attempts to answer the following questions in the context of primary education in a developing country: (i) Can teacher incentives based on test scores improve student achievement? (ii) What, if any, are the negative consequences of teacher incentives? (iii) How do school-level group incentives perform relative to teacher-level individual incentives? (iv) What is the impact of simply monitoring schools and measuring students' achievement without attaching incentives? (v) How does teacher behavior change in response to incentives? (vi) How cost effective are teacher incentives relative to other uses for the same money? and (vii) Will teachers support the idea?

We find that offering performance-based bonuses to teachers had a significant positive impact on test scores, with students in incentive schools scoring 0.19 and 0.12 standard deviations higher than students in control schools in math and language tests respectively. The mean treatment effect of 0.15 standard deviations is equal to 6 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. We find no evidence of heterogeneous treatment effects across any of these categories.

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.

We find no significant difference between the effectiveness of school-level group incentives and teacher-level individual incentives. We cannot reject equality of either mean or variance of student test scores across group and individual incentive schools. Because the average government-run school in rural AP is quite small with only 3 teachers, the results probably reflect a context of relatively easy peer monitoring.

To evaluate the impact of measurement without incentives, we also conduct endline assessments in an additional 100 schools (referred to as "pure control" schools) that were not formally a part of the study. We find no significant difference in the endline test scores between the regular "control" schools (who had a baseline test, were presented with feedback on their baseline performance, were subject to regular tracking surveys by enumerators, and knew about the endline tests in advance) and the "pure control" schools (who had none of these) suggesting that mere measurement and monitoring without any consequences for teachers had no impact on student learning outcomes.

Teacher absence did not differ across treatments. There was no difference in teacher behavior between incentive and control schools when measured by classroom observation of teaching processes, but 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. The similarity in observed classroom processes between incentive and control schools could be the result of teachers conforming to a behavior norm when they are subjected to repeated observation. We find that teachers in control schools (who were observed once a month) were more likely to be engaging in teaching activity (when observed) than teachers in the sample of "pure control" schools (who were only observed once during the year and never revisited), though there was no difference in test scores between the control and pure control schools. This suggests that repeated monitoring affected teacher behavior while it was being observed but had no effect on student performance.

The two input programs (an extra para-teacher; and cash block grants) also had a positive and significant impact on test scores. Student test scores in input schools were 0.09 standard deviations higher than those in control schools. The input programs were around five times more cost effective than the status quo, confirming the hypothesis that they were "smart" inputs. However, the incentive programs cost the same amount in

bonuses paid and students in incentive schools scored 0.06 standard deviations higher than students in the input schools, with the difference being significant at the 10% level. Thus, performance-based bonus payments were more cost effective even when compared with smart inputs, and substantially more so when compared with the status quo.

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% 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 (presented as a mean-preserving spread 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.[3]

Our results contribute to a small but growing literature on the effectiveness of performance-based pay for teachers.[4] 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 (2004), and Glewwe, Ilias, and Kremer (2003), but their evidence is mixed. Lavy uses regression discontinuity and matching methods to show

[3] 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.
[4]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.

that both group and individual incentives for high school teachers in Israel led to improvements in student outcomes. 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 incentives in a representative sample of schools.[5] 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 learning outcomes. We study group (school-level) and individual (teacher-level) incentives in the same field experiment. We isolate the impact of measurement and monitoring from that of incentives by including a set of "pure control" schools whose outcomes are compared to those in traditional "control" schools, and thereby contribute to the methodology of field experiments. 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 the first year of the program, but before they know their own performance, 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.[6] 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

[5] 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.
[6] 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.

the possibility that the change in the incentive structure was endogenously determined by unmeasured factors that directly impact the measured outcomes. True experiments in compensation structure with identical contemporaneous control groups are rare,[7] and our results can help to answer broader questions regarding performance pay in organizations, and group versus individual incentives in small group situations.[8]

An important caveat to the results presented here is that they are based on data from just the first year of the program. These results reflect only the announcement of the incentives, with no bonuses having been paid at the time of the endline tests. It is possible that the impact of the incentives will be larger in subsequent years, once the program's credibility is established; it is also possible that the gains in test scores may not persist in future years, once the novelty of the program wears off. We also do not know whether new and unanticipated dimensions of gaming will emerge as teachers become more familiar with the program. AP RESt is expected to continue until 2011, and we hope to answer these and other questions in the coming years, by continuing the experiment to study long-term outcomes.

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 shows results of the input interventions, compares them with the incentive treatments, and does a cost-benefit analysis relative to the status quo. Section 8 discusses stakeholder opinions and policy implications. Section 9 concludes.

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

[7] 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).
[8] 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.

(especially of small amounts) can sometimes crowd out intrinsic motivation and lead to inferior outcomes.[9] 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.[10]

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).[11] Let $\mathbf{a}$ be an n-dimensional vector of 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 $\mathbf{f}$ is a vector of marginal products of each action on V, and ε is noise in V.

---

[9] 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.

[10] 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.

[11] 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.

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 $V(\mathbf{a}(b_p^*)) < V(\mathbf{a}(b_v^*))$.[12] 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.[13] 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 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

[12] Thus a key challenge is choosing the appropriate performance measure P. Duflo and Hanna (2005) 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.

[13] 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.

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.[14]

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.[15]

Finally, if there are gains to cooperation, then it is possible that group incentives might yield better results than individual incentives.[16] Consider a case where teachers 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

[14] See Holmstrom (1982) for a solution to the problem of moral hazard in teams.
[15] 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.
[16] 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.

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.[17]

Lavy (2004) reports that a high-school teacher incentive program in Israel at the individual level was more effective than one at the group level. However, the two programs in Israel were implemented at different (non-overlapping) times and the schools were chosen by different (non-random) eligibility criteria. We study both group and individual incentives in the same field experiment.

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.[18] One teacher typically teaches all subjects for a given grade (and often teaches more than one grade simultaneously). All regular teachers are employed by the state, and their salary is mostly determined by experience and rank,[19] with minor adjustments based on postings, but no component based on any measure of performance. The average salary of regular

[17] 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.
[18] 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.
[19] A regression of teacher salary on experience and rank (in our sample) has an R-squared of 0.8.

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 ≈ 45 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.[20]

3.2 Sampling

We sampled 5 districts across each of the 3 socio-cultural regions of AP in proportion to population (Figure 1b).[21] 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 the first year 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 first school year, while the incentive treatments consisted of an announcement that bonuses would be paid at the beginning of the second school year *conditional* on average improvements in test scores during the first year. No school received more than one treatment. The school year in AP starts on June 15, and the

[20] 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.
[21] Subject to the selected districts within a region being contiguous for ease of logistics and supervision.

baseline tests were conducted in the 500 sampled schools during late June and early July, 2005.[22] After the baseline tests were evaluated, 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 6 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. Column 7 shows the largest difference in each variable across treatment categories, and we cannot reject the null that the variables are equal across any pair of treatments at the 5% level (column 8).

After the randomization, mandal coordinators (MCs) from the Azim Premji Foundation (APF) personally went to each of the 500 schools in the first week of August 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 six rounds of unannounced tracking surveys to each of the 500 schools during September 2005 to February 2006 (averaging one visit/month) to collect data on process variables including student attendance, teacher attendance and activity, and classroom observation of teaching processes. The 500 schools operated under identical conditions of information and monitoring and only differed in the treatment that they received. This ensures that Hawthorne effects are minimized and that a comparison between treatment and control schools can accurately isolate the treatment effect.

To estimate the impact of measurement and monitoring without incentives, data was also collected on schools other than the 500 schools in the main study (we refer to these

[22] 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).

as "pure control" schools). We sampled 300 extra schools (6 in each mandal), and the MCs made *only one* unannounced visit to these schools during the school year and collected similar process data from these schools. We use this data to construct the process variables for the "pure control" category of schools (since these schools are not subject to the effects of being under regular observation).

We randomly sampled 100 out of these 300 extra schools (2 in each mandal) and administered the same endline tests in these 100 schools as in the main 500 schools. These schools were given only a week's notice before being tested (whereas the 500 schools knew about the tests from the beginning of the year and were reminded of it by the repeated tracking surveys). Comparing process and test score outcomes between "control" schools (that were fully in the study) and "pure control" schools allows us to estimate the impact of external measurement and monitoring on school outcomes (which can be considered analogous to estimating the magnitude of the Hawthorne effect).

3.4 Description of Input and Incentive Treatments

3.4.1 Extra Para-Teacher

Para-teachers (also known as contract teachers) are hired at the school level and have usually completed either high school or college but typically have no formal teacher training.[23] Their contracts are renewed annually and they are not protected by any civil-service rules. Their typical salary of around Rs. 1000/month is less than 15% of the average salary of regular government teachers. Para-teachers usually teach their own classes and are not 'teacher-aides' who support a regular teacher in the same classroom. The use of para-teachers has increased in developing countries as a response to fiscal pressures and to the perceived inadequacy of incentives for regular civil service teachers. There is some evidence that para-teachers are more cost effective than regular teachers but their use is a controversial issue. Proponents argue that para-teachers are a cost-effective way of reducing class size and multi-grade teaching; opponents argue that the use of untrained teachers will not improve learning.

Para-teachers are paid for 10 months/year, and so the extra spending on the para-teacher intervention was Rs. 10,000/school. A typical government school has a little over

[23] See the 3 case studies in Pritchett and Pande (2006) for a detailed discussion on para-teachers in India. See Banerjee et al (2005) and De Laat and Vegas (2005) for other studies on para-teachers.

3 regular teachers and average *variable* costs of Rs. 25,000/month. Salaries are paid for all 12 months and the annual variable cost of running a typical government school is around Rs. 300,000/year. Thus, the additional spending of Rs. 10,000/year is a little over 3% of the average annual variable cost of running a government school. Schools selected to receive the extra para-teacher took a few weeks to hire and appoint the para-teacher and typically had them in place by September, 2005. The appointment of the extra para-teacher was independent of the appointment, posting, and transfer of regular teachers.

3.4.2 Extra Block-Grant

The block grant intervention targeted non-teacher inputs directly used by students.[24] The schools had the freedom to decide how to spend the block grant, subject to guidelines that required the money to be spent on inputs directly used by children. The block grant amount was set so that the average additional spending per school was the same as that in the para-teacher treatment.[25] Schools receiving the block grant were given a few weeks to make a list of items they would like to procure. The list was approved by the project manager, and the materials were jointly procured by the teachers and the APF mandal coordinators and provided to the schools by September, 2005. The majority of the grant money was spent on notebooks, workbooks, exercise books, slates and chalk, writing materials, and other interactive materials such as charts, maps, and toys.

3.4.3 Group and Individual Incentives

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[26]

[24] See Pritchett and Filmer (1999) for a political economy model, in which the marginal return to additional spending on non-teacher inputs is higher than the returns to additional spending on regular teacher salaries.
[25] The block grant was set on a per child basis, and so the value of the grant scaled linearly with valid enrollment. Schools receiving the para-teacher treatment however received only one extra para-teacher regardless of enrollment and so the effective class-size reduction would vary across treatment schools.
[26] 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.

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).[27]

The 'slope' of Rs. 500 per percentage point gain in average scores was set so that the expected incentive payment per school would be approximately equal to the additional spending in the input treatments (based on calibrations from the project pilot).[28] 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 will be no improvement threshold in subsequent years of the program because each year's endline score will be used as the next year's baseline and the testing will be conducted at the same time of the school year on a 12-month cycle.[29]

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.[30] If the function transforming teacher effort into test-scores is concave (convex) in the baseline score, teachers have an incentive to

[27] 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.

[28] 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.

[29] The convexity in reward schedule due to the threshold could have induced some gaming, but the distribution of mean class and school-level gains does 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.

[30] 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.

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. To make cheating as difficult as possible, the tests were conducted by external teams of 5 evaluators in each school (1 for each grade), the identity of the students taking the test was verified, and the grading was done at a supervised central location at the end of each day's testing.

4. Test Design

4.1 Test Construction

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).[31]

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 comparison of progress between treatment and control schools. The second test (the 'higher endline') tested skills from the current school year's syllabus. The term 'endline' refers to both rounds of end of year tests (the 'lower endline' and the 'higher endline').

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

[31] 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).

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

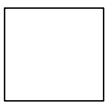
Question 2: Put the correct number in the empty box:

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

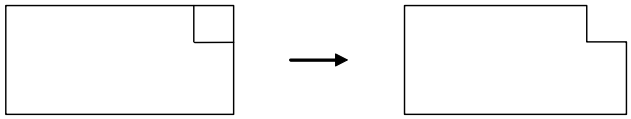
The first question follows the standard textbook format for asking multiplication questions and would be classified as "mechanical" while the second one requires the students to understand that the concept of multiplication is that of repeated addition, and would be classified as "conceptual." Note that 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.[32]

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. Since the subject is formally introduced only in grade 3 in the school curriculum, the EVS tests were administered in grades 3 to 5.

5. Results

5.1 Teacher Turnover and Student Attrition

Regular civil-service teachers in AP are transferred once every three years on average. While this could potentially bias our results if more teachers chose to stay in or tried to transfer into the incentive schools, it is unlikely that this was the case since the treatments were announced in August, 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 (Table 1 – Panel B).

[32] 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.

The average student attrition rate in the sample is 14.2%, and there is no significant difference in attrition across the treatments (Table 1 – Panel B). 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.

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:

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

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.[33] 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.

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.

An alternate specification stacks all the data from every test round and uses a difference in difference approach (DID) to estimate δ as follows:

[33] 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.

$$T_{ijkm} = \alpha + \phi \cdot Incentives + \gamma \cdot EL + \delta \cdot (Incentives \times EL) + \beta \cdot Z_m + \varepsilon_k + \varepsilon_{jk} + \varepsilon_{ijk} \quad (5.2)$$

The DID specification constrains the coefficient on the lagged test score (γ in equation 5.1) to be equal to one and is more efficient if this restriction is true. However, if the restriction is not valid, then specification (5.1) allows for this. We also run both specifications with controls for household and school variables.[34]

5.3 Impact of Incentives on Test Scores

Table 2 reports the results of estimating each of these specifications. Panel A shows that the estimate of δ from specification (5.1) is 0.154 standard deviations (SD). Panel B shows the difference-in-difference estimate of δ to be 0.169, which is not significantly different from 0.154. However, the data strongly rejects 'γ = 1' and so we prefer specification (5.1) because it does not constrain the coefficient on the lagged score to be equal to one.[35] The impact of the incentives is greater in math (0.19 SD) than in language (0.12 SD) and a Chow test shows this difference to be significant. The addition of school and household variables does not change the estimated value of δ in any of the regressions, confirming the validity of the randomization. We verify that teacher transfers do not affect the results by estimating equation (5.1) on the sample of teachers who were not transferred during the entire period, and the estimate of δ is 0.18. We have no reason to believe that cheating was a problem and Appendix B describes both the testing procedure and robustness checks for cheating.

Figure 2a (left panel) plots the density of the gain in test scores $(T_{ijkm}(EL) - T_{ijkm}(BL))$ 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 gain in normalized test scores by treatment at every percentile of the gain distribution (the x-axis is the gain percentile for each treatment). The vertical distance between the two

[34] The randomization implies that the estimates of δ will not change, but including the controls can decrease the clustered standard errors by absorbing common variation at the school level (though it can also increase it by absorbing within school variation of student demographic characteristics). See Donner and Klar (2000) for a discussion of the implications for power calculations and sample size determination.

[35] The estimate of γ is biased due to measurement error and omitted ability, but the data strongly rejects γ=1. Once we have data for more than 1 year, we can use the panel data to estimate γ consistently (see Todd and Wolpin, 2003 for details).

plots is positive at every percentile but increasing. In other words, gains are higher at every percentile and also have higher variance in incentive schools. Another way of looking at Figure 2b is that the horizontal gap represents the percentile-point gap in the gain distributions. Thus, the median score gain in an incentive school would be equal to the 57th percentile of score gains in a control school (as shown in Figure 2b).

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. Equivalently, the 'Incentives' variable can be included and one of the interactions can be omitted, followed by an F-test of the null that the δ_i's are jointly equal to zero. Table 3 presents the effect of incentives on performance by grade (1-5). The estimate of δ is positive for every grade and we cannot reject that the treatment effect is equal across the 5 grades. Similarly we cannot reject equality of treatment effects across the 5 project districts.

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) 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.

Finally, we aggregate the questions by 'skill/competency' as defined by EI and compare the difference in skill-level mean scores between incentive and control schools. Students in incentive schools outperformed those in control schools in 80% of the skills and significantly so in 40% of them. Out of the remaining 20% of skills where the incentive schools underperformed, only 2% were significant at the 10% level and none at the 5% level, which confirms that the gains in the incentive schools were broadly present in all parts of the curriculum.

5.5 Mechanical versus Conceptual Learning and Non-Incentive Subjects

Table 5 shows summary statistics of the fraction of questions answered correctly on the endline by mechanical and conceptual type (for each grade and subject), and the gap between the two types of questions appears to increase with the grade. This is consistent with the idea that in lower grades, all questions are equally unfamiliar (at which point a conceptual question can actually be easier as indicated in section 4.2), but that as the grades progress the students' knowledge seems to comprise more of knowing patterns from the textbook as opposed to understanding concepts.

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).[36] The other interesting result is that the coefficient on the baseline score is significantly lower for the conceptual component than for the mechanical component (0.34 versus 0.48), indicating that these questions represented unfamiliar territory relative

[36] 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.

to the mechanical questions. The relative unfamiliarity of these questions increases our 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 0.14 standard deviations higher than students in control schools in science and social studies respectively (Table 7). The coefficients on the baseline math and language scores here are much lower than those in Table 2 (below 0.25 versus 0.5), 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).

5.6 Heterogeneity of Treatment Effects

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

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

We find no evidence of a significant difference in the effect of the incentives on any of the student demographic variables, including an index of household affluence, an index

of household literacy, the caste of the household, the student's gender, and the student's baseline score.[37]

Similarly, we find no evidence of differential impact of incentives by teacher characteristics including gender, designation, experience, and base pay. The last finding might suggest that the magnitude of the incentive was not 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 educated and experienced and so we cannot disentangle the impact of the incentive amount from that of teacher variables that influence the base pay.

The only evidence of heterogeneous treatment effects is at the school level, where schools with better infrastructure show a greater response to the incentives. Incentive schools score 0.06 SD higher in the endline tests for every additional point on a 6-point infrastructure index (as described in the notes to Table 1). The mean infrastructure index in the sample is 3.2, implying that a school having an index value of 0 or 1 showed no improvement relative to the control schools, while schools having index values of 5 or 6 improved by 0.3 to 0.35 standard deviations.

5.7 Group versus Individual Incentives

Point estimates suggest that students in individual incentive schools perform slightly better than those in group incentive schools, but the difference is not significant (Table 8). There was also no significant difference in the variances of the two gain distributions (though the variance of the gain distribution of the group incentive schools is slightly lower). As mentioned earlier, the average school in rural AP is quite small and has only 3 teachers on average. The results therefore probably reflect a context of relatively easy peer monitoring.

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

[37] The affluence index (0-4) assigns one point for each of owning a brick home, and having functioning water, electricity and a toilet, and the literacy index (0-4) assigns one point for each parent being literate and an additional point for each parent who has completing primary school

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. Also, field reports suggest that teachers in some individual incentive schools agreed ex ante to split their bonus amounts equally, potentially reducing the difference in treatments.

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 the gains from teacher incentives can be obtained just as effectively from group and individual incentives in this context of rural schools with a small number of teachers. However, the result should not be extrapolated to large schools in which the group would comprise many teachers.

Beyond the context of schools and teachers, our results indicating near equivalence between individual and group incentives may also be relevant for compensation design in small groups in general. In particular, the result suggests that the inability to provide individual incentives in cases of team production and unobservable individual output might not be a salient constraint to incentive provision in small groups if group incentives perform as well as the individual incentives even when the latter are feasible.[38]

5.8 Impact of Measurement

To isolate the impact of 'measurement' on test scores, we run the regression:

$$T_{ijkm}(EL) = \alpha \ + \delta \cdot Control + \beta \cdot Z_m + \varepsilon_k + \varepsilon_{jk} + \varepsilon_{ijk}$$

using only the 'control' and 'pure control' schools. Table 9 shows that there is no significant impact of the baseline test, detailed feedback, continuous tracking surveys, and advance announcement of the endline assessments on the test score outcomes of the control schools relative to the 100 randomly-sampled schools in the pure control category that were told about the testing only a week before the endline tests. We find this result surprising given that our initial hypothesis was that the 'announcement effect' of external testing would have a larger impact. However, all 500 schools in the main study were told

[38] Of course, if the group's ability to prevent free riding is improved by being able to observe individual output (as opposed to effort), then the equivalence between group and individual incentives might not hold in cases where individual outcomes are not measurable.

that information about a specific school or teacher's performance would not be shared outside the school on an identifiable basis (to limit confounding of monetary and non-monetary incentives). Since no consequences were attached to poor or good performance on the test for the control schools, the result reinforces the importance of incentives, as opposed to mere diagnostic information, in changing behavior.[39]

6. Teacher Behavior and Classroom Processes

As described in section 3.3, the APF mandal coordinators (MCs) conducted six rounds of unannounced tracking surveys during the school year to all 500 schools (and one similar visit to a sample of extra schools outside the study). 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.

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 (Table 10 – Panel A). 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 the processes between the schools. One explanation could be that teachers' actions converge to certain norms of behavior when under repeated observation.

We check for this by comparing the process variables in control schools with those in the pure control schools. While there is no difference in teacher attendance, the control schools do better on many of the recorded measures of classroom processes (Table 10 – Panel B). Since there is no significant difference in test score outcomes between the

[39] Of course, this result does not rule out the possibility that widely disseminating information and creating 'public rankings' etc. can induce better performance even without monetary rewards.

control and the pure control schools, it seems likely that the superior process measures in the control schools are a result of the repeated observation by (usually) the same person over the course of the year. This interpretation is supported by the fact that there is no difference in teacher absence, or classroom cleanliness (which cannot be affected after the MC arrives in the school), but there is a significant difference in actions that a teacher is likely to believe constitute good teaching (such as using the blackboard, reading from the textbook, making children read from the textbook, and assigning homework). If actions converge to a norm of behavior under observation,[40] we may not be able to identify process differences between control and incentive schools by observation even if the true processes are different.

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 2005 – 06 school year 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 11). While self-reported measures of teacher activity might not be considered very credible, we find a positive correlation between the reported activities of teachers and the performance of their students. The effect of assigning homework and class work is positive but not significant, while that of extra classes and practice tests are positive and strongly significant.

The interview responses suggest other 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.

Our use of both direct observations and interviews might help in reconciling the difference between the findings of Glewwe et al. (2003) and Lavy (2004) with respect to

[40] A striking example is provided in Brennan and Pettit (2005) who present experimental evidence that users of public rest rooms in New York were twice as likely to wash their hands (80% versus 40%) after using the toilet if there was someone else present in the rest room.

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 teach more effectively when present. Regular observation changed teacher behavior while they were observed, but had no impact on student learning outcomes.

7. Input Treatments & Cost-Benefit Analysis

7.1. Effectiveness of Para-teachers and Block grants

Table 12 shows the impact on test scores of providing additional inputs to primary schools using the same specification as before (5.1) and the sample of 300 control, para-teacher, and block-grant schools. Columns 1, 3, and 5 pool both the input treatments together, while columns 2, 4, and 6 separate the impact of the extra para-teacher and the block grant. The input treatments had a significant positive impact on student test scores, increasing math and language scores by 0.1 and 0.08 standard deviations respectively relative to the control schools. There was no significant difference between the effects of the block grant and the para teacher. Further results on the inputs and estimates of the education production function are presented in a companion paper.

7.2. Comparison of Input and Incentive Treatments

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

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

using the full sample of 500 schools. While both categories of treatments had a positive and significant impact on learning outcomes, the incentive schools performed 0.06

standard deviations better than the input schools and this difference is significant at the 10 percent level (Table 13 - Column 1). The incentive schools perform better than input schools in both math and language and in both rounds of testing, though not all of the differences are significant. The performance difference is greater (and significant at the 5% level) for the higher endline test than for the lower endline test (where the difference is not significant). It is possible that the increased intensity of effort reported in the incentive schools applied more to current-year materials (covered in the higher endline test) than in reviewing previous years' materials (covered in the lower endline test).

Table 14 summarizes the amount of money spent on each of the interventions, and the total amount spent on each intervention is roughly equal. The average annual spending on the incentives was around Rs. 9,000/school. The bonus payment in the group incentive schools was around 35% lower than that in the individual incentive schools (in spite of having only slightly lower performance) because classes with gains less than 5% brought down the average school gain in the group incentive schools, while gains below 5% in some classes did not hurt other classes in the individual incentive schools. Thus, while all four interventions had a significant positive impact, the incentives were more cost effective than the inputs, and the group incentive treatment was the most cost effective of the four interventions.

Table 15 shows the fraction of eligible schools and teachers who qualified for an incentive payment along with the distribution of bonuses. 61% of schools and 62% of teachers who were eligible for a group incentive qualified for one. 66% of teachers eligible for an individual incentive qualified for it, and 87% of the individual incentive schools had at least one teacher who received a bonus. The frequency of high payments is greater in the individual incentive schools, which is consistent with performance having a higher variance at the class-level than at the school-level.

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.[41] 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 run.

7.3. Comparison with Regular Spending Patterns

Since the inputs considered for the comparison above are non-typical ones (in that they do not represent a proportional increase in current spending), a second way of evaluating the cost-effectiveness of the interventions is to compare their outcomes to those from the 'business as usual' pattern of spending. Since the skills and competencies assessed on the lower endline test correspond exactly to those tested on the baseline, we can consider the gain in scores in the control schools between the baseline and lower endline (but not the 'higher endline') to represent the progress made in one school year in a typical government school.[42] The gains in the incentive and input schools can then be compared with a year's progress in the average government school.[43]

Table 16 shows this calculation. Columns 1, 2, and 3 present the mean gain in test scores for all schools within a treatment category by grade for control, input, and incentive treatments respectively. Columns 4 and 5 present the ratio of the gain in the input and incentive schools with respect to the control schools. Averaging across grades and subjects, the input schools added the equivalent of 116% of a regular year's learning while the incentive schools added 126%.

Both the input and incentive interventions were highly cost effective relative to the status quo. As mentioned earlier, the variable cost of running a typical government school is around Rs. 300,000/year. The input treatments added the equivalent of 16% of

[41] 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.

[42] The baseline and lower endline were of exactly the same length and had a 1 to 1 correspondence at the question level. So if the baseline had a question testing 'double digit addition with carry over', then the lower endline also had a similar question but with different numbers. This does not apply to the higher endline because there was no 'matched baseline' corresponding to it.

[43] This is an approximate calculation, because a precise comparison of 'changes' in learning across schools usually requires retesting of the *same* questions as opposed to *similar* questions as done here.

a normal year's learning for an additional cost of Rs. 10,000/year, and so 6.25 years of the programs would add an equivalent of a full year's learning at a cost of Rs. 62,500 (or around 21% of the current variable cost of a year's schooling). This confirms our initial hypothesis that the para-teacher and block grant interventions are "smarter inputs" that are substantially more cost effective than the status quo.

Since a year of the incentive program added the equivalent of 26% of a normal year's learning, four years of such a program would add the equivalent of a full year's learning at a cost of Rs. 36,000 (assuming the same program effects continue). The teacher incentive program could therefore add the equivalent of a full year of schooling at 12% of the current cost of a year's schooling, which would make it a very cost effective program relative to the status quo.[44]

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 testing and data entry but not the additional costs borne for research purposes.[45] 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 highly cost effective even after adding these costs (adding a full year's learning for Rs. 56,000 or 19% of the current cost of a year's schooling) and even more so in the long-run when the incentive cost is only the risk premium and not the amount of the incentive payments.

8. Stakeholder Reactions and Policy Implications

8.1 Stakeholder Reactions

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

[44] This estimate is likely to be a lower bound for 2 reasons. First, it does not use the 'higher endline' test data where the performance advantage of incentive schools is even more pronounced. Second, the baseline tests were in June and the lower endline tests were in March, and so the baseline score might be artificially low due to students forgetting material over the vacation. If the true absolute improvement is therefore a little lower for both treatment and control schools, the relative advantage of the treatments studied here would be even higher.

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

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 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.[46]

To obtain data on parental preferences, the APF mandal coordinators interviewed a small random sample of parents during each round of the tracking survey. The most important reasons that parents report for sending their children to school are "improving job prospects" (66%) and "building discipline" (55%). The most important traits that parents report looking for in their child's teachers are: "Can improve skills and learning" (74%) and "can build discipline" (73%). These objectives are likely to be well aligned with providing teachers bonuses based on improvements in student test scores.

8.2 Policy Implications

Since both the incentive and para-teacher policies are more cost effective than the status quo, it might be possible to combine them creatively to improve primary education in Andhra Pradesh (and in India). For instance, if all new teachers are hired as para-teachers, the money saved could be used to expand the incentive program to all teachers. Having an incentive program would also produce a track record of teacher performance, and this data could be used to promote para-teachers to regular teacher status after a period of time on the basis of a sustained track record of performance.[47] While the reactions of teachers' unions to such a proposal could be different from that of individual

[46] 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 many things 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.

[47] A similar proposal is made in Pritchett and Pande (2006).

teachers,[48] the overall positive response from teachers (and potentially parents) suggests that it might be possible to institute performance-based pay in government schools.

The longer-term benefits to performance pay include not only greater teacher effort, but also potentially the entry of better teachers into the profession.[49] 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.

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 the incentives led to significant improvements in both math and language 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 between the effectiveness of group versus individual incentives. 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 and appear to be

[48] 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.

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

substantially more cost effective than the status quo. 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 main caveat to the results presented here is that they only represent data from the first year of the program. The impact of the incentives may be larger in subsequent years, once the program's credibility is established; but it is also possible that the gains in test scores may not persist in future years, once the novelty of the program wears off. We also do not know how an understanding of the dynamics of student learning will affect a rational teacher's response to an incentive program in the long run, and if unanticipated dimensions of gaming will emerge as teachers become more familiar with the program.

Another 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 year 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 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.[50]

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.[51] 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 year. 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. However, it is possible that there may be no satisfactory resolution of the tension between accuracy and transparency.[52]

We expect AP RESt to continue until 2011 and hope to answer many of the questions raised in this section as part of an ongoing effort to study the long-term outcomes of the input and incentive programs. It may also be possible to continue random allocation of variants of the incentive program during a process of slowly expanding coverage. This will allow experimentation with variations such as tournaments, incentives on both teacher attendance and test scores, and interactions between inputs and incentives. It may also be possible 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.

[50] 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.

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

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

References:

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., and R. HANNA (2005): "Monitoring Works: Getting Teachers to Come to School," Cambridge, MA: National Bureau of Economic Research Inc

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.
— (2004): "Performance Pay and Teachers' Effort, Productivity and Grading Ethics," Cambridge: National Bureau of Economic Research.
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 (2006): "Teacher and Non-Teacher Inputs in the Education Production Function: Experimental Evidence from India," 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.
OYER, P. (1998): "Fiscal Year Ends and Nonlinear Incentive Contracts: The Effect on Business Seasonality," *Quarterly Journal of Economics*, 113, 149-85.
PRATHAM (2005): *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 – onwards:	Long-term outcomes and sustainability

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

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.

We looked carefully at the student-level records in the top 10% of the incentive schools, and found 2 classrooms (out of the 100 top-performing classrooms) in the incentive schools with suspicious answer patterns.[53] As Jacob and Levitt (2003) point out, it is quite difficult to isolate cheating from very good teaching with just one year's data, and so we cannot be sure that these were cases of cheating. However, the results on incentives are robust to excluding these 2 classrooms (and even to completely dropping the 2 schools) from the analysis.

[53] Since the evaluation of the papers was done at a central location, the most likely form of cheating would not be a teacher altering a few students' answer sheets (as found by Jacob and Levitt, 2003) but a teacher potentially giving out answers to a few questions in the class (though the teachers were typically not allowed in the class during testing). Thus, we identify suspicious classrooms by looking for cases where *all* students get a few difficult questions correct, while some of them get easier questions wrong.

Table 1: Sample Balance Across Treatments

	[1] Control	[2] Extra Para-teacher	[3] Block Grant	[4] Group Incentive	[5] Individual Incentive	**[6] P-value (Equality of all groups)**	[7] Largest Difference	**[8] P-value (Largest Difference=0)**
	Panel A (Means of Baseline Variables)							
School-level Variables								
Total Enrollment (Baseline: Grades 1-5)	113.2	104.6	104.2	111.3	112.6	**0.82**	9.0	**0.38**
Total Test-takers (Baseline: Grades 2-5)	64.9	62.0	62.3	62.0	66.5	**0.89**	4.5	**0.42**
Number of Teachers	3.07	2.83	3.03	3.12	3.14	**0.58**	0.31	**0.14**
Pupil-Teacher Ratio	39.5	39.8	34.6	40.6	37.5	**0.66**	6.0	**0.16**
Infrastructure Index (0-6)	3.19	3.13	3.33	3.14	3.26	**0.84**	0.21	**0.32**
Proximity to Facilities Index (8-24)	14.65	14.97	14.71	14.66	14.72	**0.98**	0.32	**0.55**
Baseline Test Performance								
Math (Raw %)	18.4	17.3	16.6	17.8	17.4	**0.56**	1.8	**0.10**
Math (Normalized - in Std. deviations)	0.057	-0.022	-0.059	0.020	-0.005	**0.57**	0.116	**0.11**
Telugu (Raw %)	35.0	34.2	33.7	34.8	33.4	**0.81**	1.6	**0.30**
Telugu (Normalized - in Std. deviations)	0.032	-0.003	-0.027	0.024	-0.039	**0.82**	0.071	**0.32**
	Panel B (Means of Endline Variables)							
Turnover and Attrition								
Teachers Who Stayed the Full Year/ Total in School Beginning of School Year (%)	69.8	68.8	65.9	66.4	69.3	**0.81**	3.8	**0.34**
Teachers Who Stayed the Full Year/ Total in School at End of School Year (%)	66.0	66.6	66.4	66.7	67.5	**1.00**	1.5	**0.69**
Student Attrition (Students who did not take an endline test as a fraction of those who took a baseline test)	14.0	14.9	14.3	13.2	14.6	**0.79**	1.7	**0.23**

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

Panel A (Coefficient on the lagged score is unconstrained)

	Dependent Variable = Normalized Endline Test Score					
	Combined		**Math**		**Telugu (Language)**	
	[1]	[2]	[3]	[4]	[5]	[6]
Normalized Baseline Score	0.499	0.494	0.49	0.481	0.516	0.515
	(0.013)***	(0.013)***	(0.017)***	(0.018)***	(0.014)***	(0.015)***
Incentive School	0.154	0.153	0.189	0.193	0.12	0.113
	(0.042)***	(0.045)***	(0.049)***	(0.054)***	(0.038)***	(0.040)***
School and Household Controls	No	Yes	No	Yes	No	Yes
Observations	68716	54910	34132	27278	34584	27632
R-squared	0.29	0.30	0.28	0.29	0.32	0.33

Panel B (Difference in Difference - Coefficient on lagged score constrained to equal one)

	Dependent Variable = Normalized Endline Test Score					
	Combined		**Math**		**Telugu (Language)**	
	[1]	[2]	[3]	[4]	[5]	[6]
Endline * Incentives	0.169	0.181	0.197	0.215	0.14	0.147
	(0.047)***	(0.049)***	(0.054)***	(0.058)***	(0.047)***	(0.048)***
School and Household Controls	No	Yes	No	Yes	No	Yes
Observations	137025	104345	68305	52005	68720	52340
R-squared	0.09	0.09	0.09	0.10	0.09	0.10

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)**	
	[1]	[2]	[3]	[4]	[5]	[6]
Incentives * Grade 1	0.105	-0.05	0.109	-0.076	0.101	-0.024
	(0.064)	(0.065)	(0.069)	(0.073)	(0.065)	(0.068)
Incentives * Grade 2	0.11	-0.045	0.124	-0.061	0.097	-0.028
	(0.054)**	(0.061)	(0.058)**	(0.069)	(0.058)*	(0.066)
Incentives * Grade 3	0.181	0.026	0.22	0.035	0.143	0.018
	(0.055)***	(0.045)	(0.062)***	(0.054)	(0.053)***	(0.049)
Incentives * Grade 4	0.188	0.033	0.257	0.072	0.121	-0.004
	(0.054)***	(0.046)	(0.067)***	(0.060)	(0.048)**	(0.044)
Incentives * Grade 5	0.155		0.185		0.125	
	(0.051)***		(0.063)***		(0.048)***	
Incentive School		0.155		0.185		0.125
		(0.051)***		(0.063)***		(0.048)***
Observations	68716	68716	34132	34132	34584	34584
F-Test p-value (Equality Across Grades)	0.6104	0.6104	0.2244	0.2244	0.9533	0.9533
R-squared	0.29	0.29	0.28	0.28	0.32	0.32

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)**	
	Lower Endline	Higher Endline	Lower Endline	Higher Endline	Lower Endline	Higher Endline
	[1]	[2]	[3]	[4]	[5]	[6]
Incentive School	0.124	0.193	0.153	0.236	0.096	0.153
	(0.042)***	(0.050)***	(0.048)***	(0.062)***	(0.041)**	(0.042)***
Observations	38802	29914	19274	14858	19528	15056
R-squared	0.30	0.30	0.29	0.29	0.31	0.34

Notes:

1. All regressions include mandal (sub-district) fixed effects and standard errors clustered at the school level.
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: Raw Scores (% Correct on Endline) by Mechanical and Conceptual Categories

		Grade				
		1	**2**	**3**	**4**	**5**
Math	**Mechanical**	15.4	38.6	30.4	28.4	25.2
	Conceptual	18.7	40.0	25.8	23.3	16.3
	Difference	-3.3	-1.4	4.6	5.1	8.9
Language	**Mechanical**	32.4	54.3	54.2	49.5	58.1
	Conceptual	24.4	40.4	22.0	32.2	39.5
	Difference	8.0	13.9	32.2	17.3	18.6

Note: All differences except grade 2 math are statistically significant at the 5% level.

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)					
	Combined		**Math**		**Telugu (Language)**	
	Mechanical	Conceptual	Mechanical	Conceptual	Mechanical	Conceptual
	[1]	[2]	[3]	[4]	[5]	[6]
Normalized Baseline Score	0.482	0.338	0.492	0.265	0.48	0.411
	(0.012)***	(0.011)***	(0.015)***	(0.015)***	(0.013)***	(0.013)***
Incentive School	0.134	0.135	0.168	0.165	0.101	0.106
	(0.038)***	(0.042)***	(0.045)***	(0.049)***	(0.036)***	(0.040)***
Observations	69310	69310	34428	34428	34882	34882
R-squared	0.28	0.17	0.28	0.14	0.29	0.23

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

	Dependent Variable = Normalized Endline Test Score	
	Science	**Social Studies**
	[1]	[2]
Normalized Baseline Math Score	0.214	0.222
	(0.019)***	(0.018)***
Normalized Baseline Language Score	0.206	0.287
	(0.019)***	(0.019)***
Incentive School	0.107	0.135
	(0.052)**	(0.047)***
Observations	12011	12011
R-squared	0.26	0.3

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 8: Impact of Group Incentives versus Individual Incentives

	Dependent Variable = Normalized Endline Test Score		
	Combined	**Math**	**Telugu (Language)**
	[1]	[2]	[3]
Normalized Baseline Score	0.499	0.49	0.516
	(0.013)***	(0.017)***	(0.014)***
Group Incentive School (GI)	0.146	0.183	0.11
	(0.050)***	(0.058)***	(0.046)**
Individual Incentive School (II)	0.162	0.195	0.13
	(0.049)***	(0.060)***	(0.043)***
Observations	68716	34132	34584
F-Stat p-value (Testiing GI = II)	0.76	0.84	0.66
R-squared	0.29	0.28	0.32

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: Impact of Measurement on Test Score Performance (Control versus Pure Control)

	Dependent Variable = Normalized Endline Test Score		
	Combined	**Math**	**Telugu (Language)**
	[1]	[2]	[3]
Control Schools	0.012	-0.011	0.034
	(0.045)	(0.048)	(0.044)
Observations	49248	24615	24633
R-squared	0.11	0.11	0.11

Notes:

1. The sample includes the "control" schools and the "pure control" schools. The former had a baseline test, feedback on the baseline test, continuous tracking surveys, and advance notice about the end of year assessments. The "pure controls" had none of these and were outside the main study.

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%

Table 10: Teacher Behavior (Measured by Classroom Observation)

	Panel A			Panel B		
	Incentive versus Control Schools (All figures in %)			Control versus PURE Control Schools (All figures in %)		
Process Variable (Activities performed by Teachers unless recorded otherwise)	**Incentive Schools**	**Control Schools**	**p-Value of Difference**	**Control Schools**	**Pure Control Schools**	**p-Value of Difference**
Student Attendance	62.9	64.3	0.54			
Teacher Absence	24.9	22.5	0.21	22.5	20.6	0.342
Actively Teaching	47.5	49.9	0.46	49.9	40.9	**0.012****
Clean & Orderly Classroom	60.5	59.5	0.772	59.5	53.5	0.124
Giving a Test	26.6	26.6	0.993	26.6	27.6	0.790
Calls Students by Name	78.5	78.1	0.878	78.1	78.6	0.865
Addresses Questions to Students	62.8	63.2	0.871	63.2	58.1	**0.087***
Provides Individual/Group Help	37.1	35.7	0.625	35.7	31.9	0.263
Encourages Participation	37.6	37.0	0.835	37.0	37.0	0.996
Reads from Textbook	52.8	56.1	0.299	56.1	41.9	**0.000****
Makes Children Read From Textbook	57.8	60.0	0.43	60.0	45.6	**0.000****
Active Blackboard Usage	50.0	49.1	0.764	49.1	40.9	**0.014****
Assigned Homework	39.5	37.2	0.518	37.2	29.2	**0.034****
Provided Homework Guidance	33.6	32.9	0.849	32.9	18.0	**0.000****
Provided Feedback on Homework	24.7	27.0	0.478	27.0	13.1	**0.000****
Children were Using a Textbook	66.0	67.4	0.559	67.4	60.8	**0.026****
Children Asked Questions in Class	37.1	37.0	0.958	37.0	42.6	**0.069***
Teacher Was in Control of the Class	51.2	52.4	0.694	52.4	51.2	0.706

Notes:

1. The control and incentive schools were each visited by a mandal coordinator approximately once a month for a total of 6 visits between September 05 and March 06. The "pure control" schools consisted of around 300 randomly sampled additional schools (6 in each mandal) that were outside the main study. Each of these schools was visited only once (at an unannounced date) during the period from December 05 to February 06 and the mandal coordinators collected data similar to that collected in the schools in the main study. Student attendance data was not collected in the pure control schools.

2. Each round of classroom observation is treated as one observation and the standard errors for the t-tests are clustered at the school level (i.e. correlations across visits and classrooms are accounted for in the standard errors).

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

Table 11: Teacher Behavior (Measured by Teacher Interviews)

	Incentive versus Control Schools (All figures in %)		
Process Variable	**Incentive Schools**	**Control Schools**	**p-Value of Difference**
Did you do any special preparation for the end of year tests? (% Yes)	74.6	52.7	**0.000*** **
What kind of preparation did you do? (UNPROMPTED) (% Mentioning)			
Extra Homework	57.2	35.8	**0.000*** **
Extra Classwork	58.7	41.2	**0.000*** **
Extra Classes/Teaching Beyond School Hours	22.5	11.1	**0.000*** **
Gave Practice Tests	36.0	21.8	**0.000*** **
Paid Special Attention to Weaker Children	24.9	10.7	**0.000*** **

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

Table 12: Impact of Inputs on Learning Outcomes

	Dependent Variable = Normalized Endline Test Score					
	Combined		**Math**		**Telugu (Language)**	
	[1]	[2]	[3]	[4]	[5]	[6]
Normalized Baseline Score	0.516 (0.012)***	0.516 (0.012)***	0.487 (0.014)***	0.487 (0.014)***	0.553 (0.014)***	0.553 (0.014)***
Inputs (Pooled)	0.09 (0.033)***		0.101 (0.038)***		0.079 (0.032)**	
Extra Para Teacher (PT)		0.089 (0.038)**		0.104 (0.043)**		0.074 (0.038)*
Block Grant (BG)		0.091 (0.039)**		0.098 (0.046)**		0.084 (0.038)**
Observations	66246	66246	32926	32926	33320	33320
F-Stat p-value (Testing PT = BG)		0.96		0.90		0.81
R-squared	0.31	0.31	0.29	0.29	0.34	0.34

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 13: Impact of Inputs versus Incentives on Learning Outcomes

	Dependent Variable = Normalized Endline Test Score								
	Both Test Rounds			**Lower Endline Only**			**Higher Endline Only**		
	Combined	**Math**	**Language**	**Combined**	**Math**	**Language**	**Combined**	**Math**	**Language**
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Incentives	0.156 (0.041)***	0.189 (0.049)***	0.124 (0.038)***	0.127 (0.041)***	0.154 (0.047)***	0.100 (0.041)**	0.195 (0.050)***	0.235 (0.064)***	0.156 (0.042)***
Inputs	0.096 (0.037)***	0.110 (0.043)**	0.083 (0.036)**	0.097 (0.038)**	0.111 (0.042)***	0.083 (0.039)**	0.095 (0.044)**	0.107 (0.056)*	0.083 (0.039)**
Difference (Incentives - Inputs)	0.06	0.08	0.04	0.03	0.04	0.02	0.10	0.13	0.07
F-Stat p-value (Inputs = Incentives)	**0.09***	**0.06***	0.20	0.40	0.29	0.62	**0.02** **	**0.03** **	**0.03** **
Observations	112078	55681	56397	63338	31464	31874	48740	24217	24523
R-squared	0.29	0.27	0.32	0.29	0.28	0.31	0.3	0.28	0.34

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", and they are compared relative to the "Control" group.
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.
3. All regressions include mandal (sub-district) fixed effects and standard errors clustered at the school level.
* significant at 10%; ** significant at 5%; *** significant at 1%

Table 14: Spending by Treatment

	Extra Para Teacher	Block Grant	Group Incentives	Individual Incentives
Annual Spending by Treatment (Rs/School)	10,000	9,960	7,114	11,074

Table 15: Distribution of Incentive Payments

	Group Incentives	Individual Incentives
Number of Eligible Schools (Teachers)	100 (328)	100 (342)
Number of Schools (Teachers) Qualifying for a Bonus	61 (203)	87 (226)
Bonus Range (Rs. 0 - 5,000)	45 (158)	140
Bonus Range (Rs. 5,000 - 10,000)	12 (31)	61
Bonus Range (Rs. 10,000+)	4 (14)	25
Average Bonus (among all eligible teachers) (Rs.)	2142	3238
Average Bonus (among those who receive a bonus) (Rs.)	3461	4900
Maximum Bonus Received by a Teacher (Rs.)	11474	26448

Note: When there are 2 numbers together, the first one is the number of schools, and the number in parentheses is the number of teachers

Table 16: Comparison with Regular Pattern of Spending

	Control	Inputs	Incentives	Input Gain/ Control Gain	Incentive Gain/ Control Gain
	[1]	[2]	[3]	[4]	[5]
Mean Math Gain by Grade (in % points over Baseline Score)					
Grade 1 (Assuming Baseline = 0)	14.4%	17.6%	16.4%	1.22	1.13
Grade 2	22.6%	28.1%	26.3%	1.24	1.16
Grade 3	9.3%	10.3%	13.7%	1.11	1.48
Grade 4	16.2%	19.4%	20.2%	1.19	1.24
Grade 5	5.8%	6.6%	9.0%	1.15	1.56
Mean Math Gain Across Grades				**1.18**	**1.32**
Mean Telugu Gain by Grade (in % points over Baseline Score)					
Grade 1 (Assuming Baseline = 0)	27.3%	30.9%	29.8%	1.13	1.09
Grade 2	13.5%	17.6%	16.6%	1.30	1.23
Grade 3	10.2%	10.4%	14.1%	1.02	1.39
Grade 4	14.0%	15.3%	16.0%	1.09	1.14
Grade 5	16.7%	18.7%	20.2%	1.12	1.21
Mean Telugu Gain Across Grades				**1.13**	**1.21**
Mean Gain Across Subjects				**1.16**	**1.26**

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)

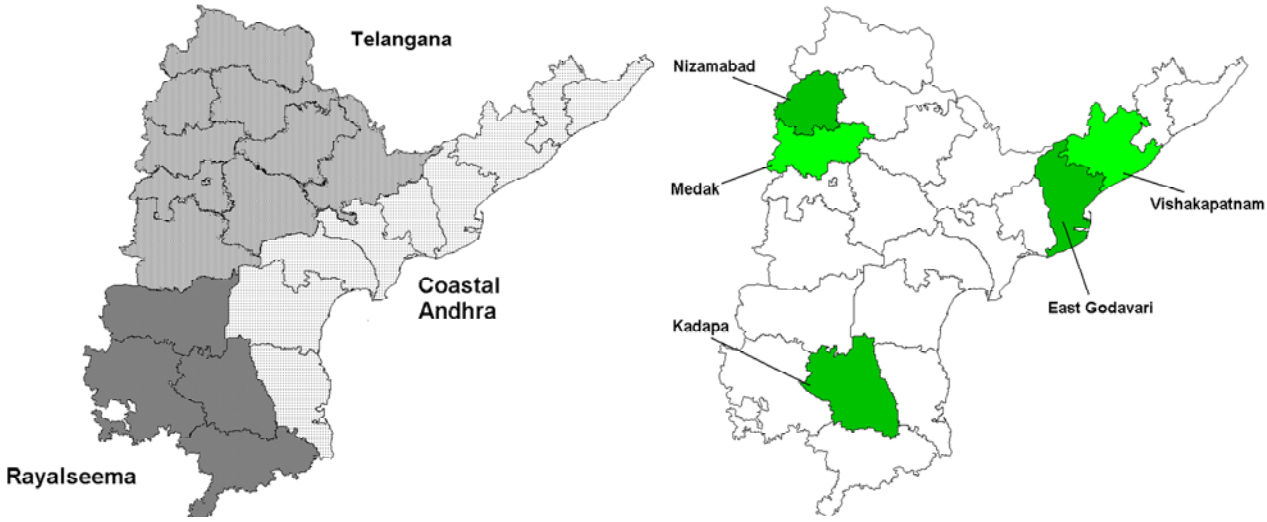


Figure 2a: Density/CDF of Normalized Test Score Gains by Treatment

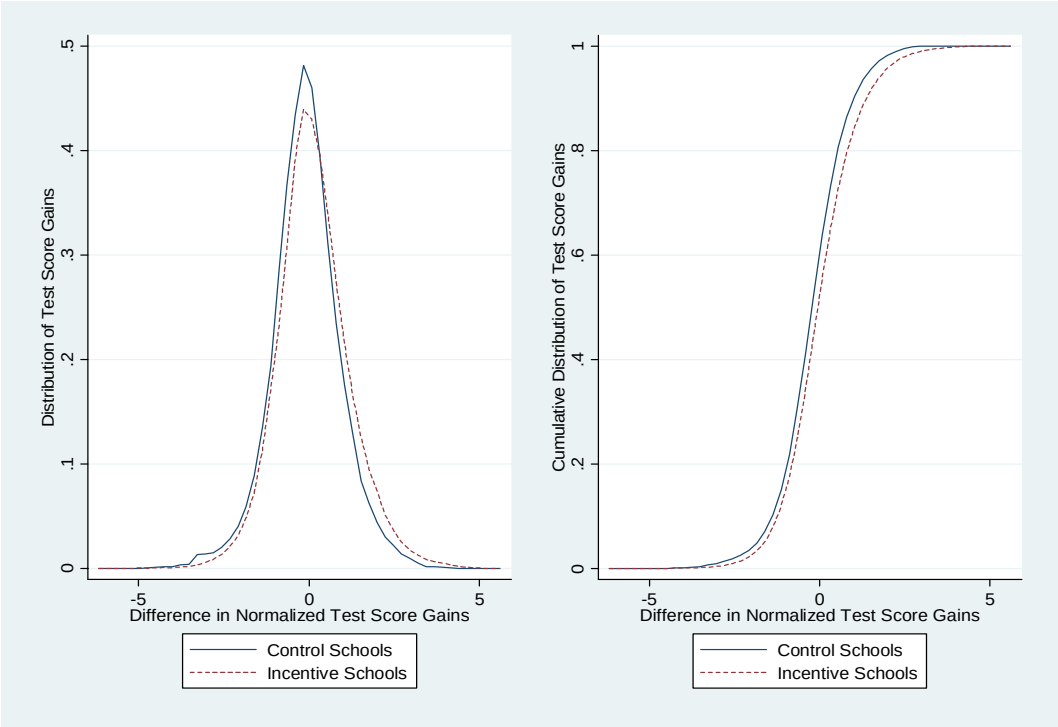


Figure 2b: Incentive Treatment Effect by Percentile of Endline Test Scores

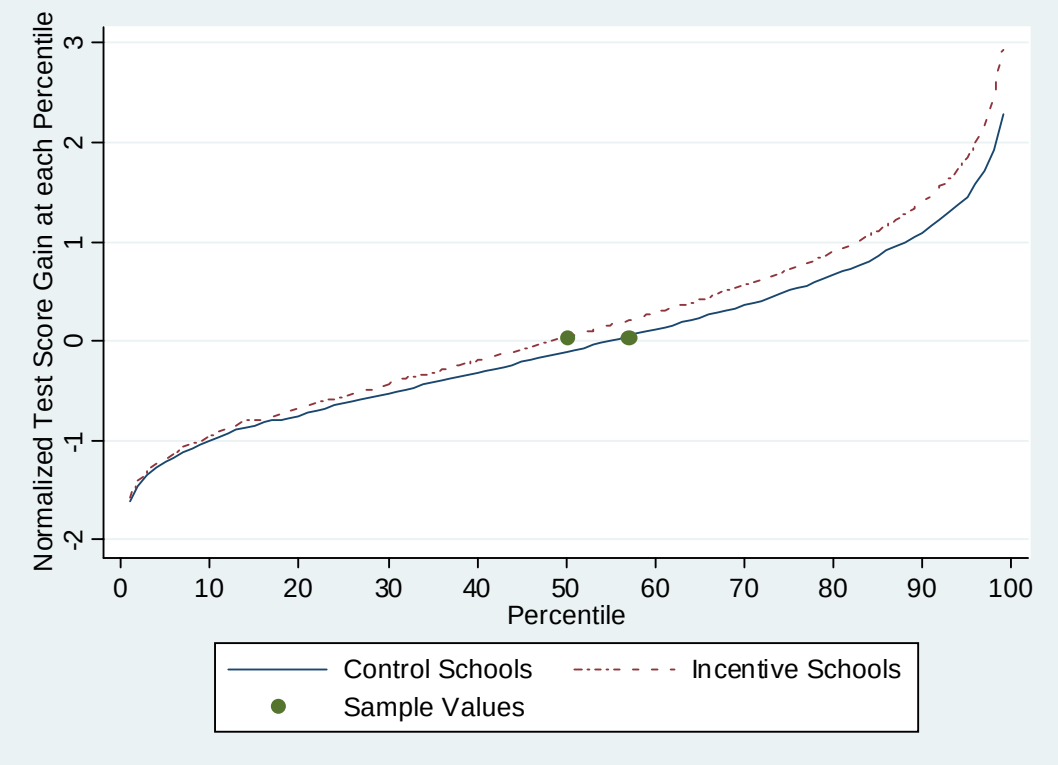


Figure 3a: Test Calibration – Range of item difficulty

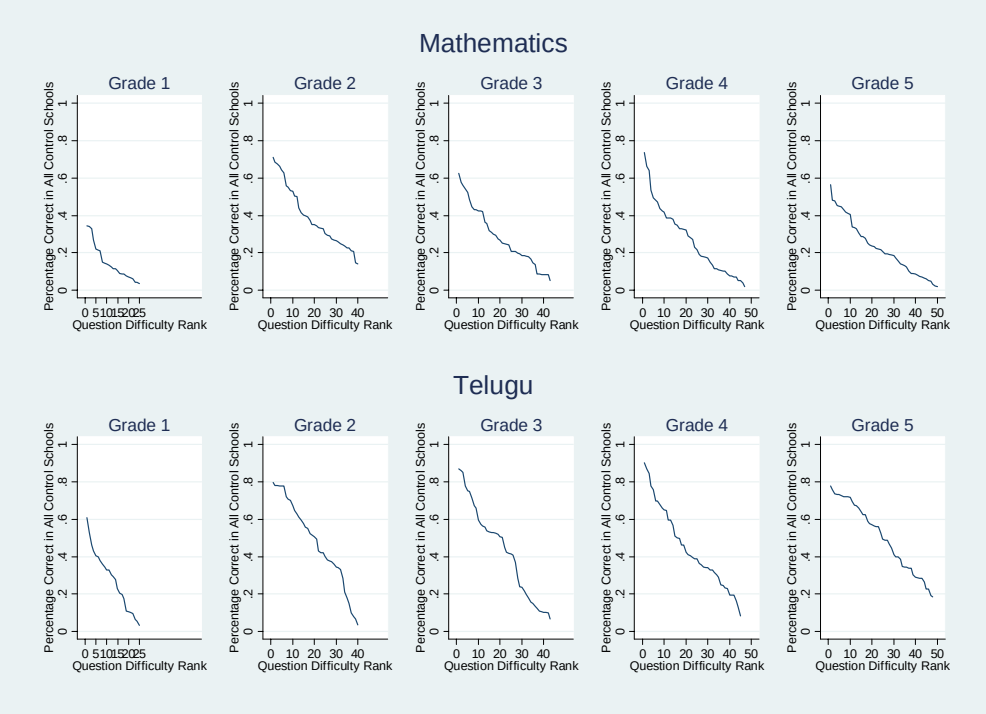


Figure 3b: Incentive versus Control School Performance – By Question Difficulty