Teacher Incentives in Developing Countries: 
Experimental Evidence from India

Karthik Muralidharan
Department of Economics
Harvard University

Venkatesh Sundararaman
South Asia Human Development
The World Bank

September 2007

Document of The World Bank
## ABBREVIATIONS AND ACRONYMS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP</td>
<td>Andhra Pradesh (AP)</td>
</tr>
<tr>
<td>APF</td>
<td>Azim Premji Foundation</td>
</tr>
<tr>
<td>AP REST</td>
<td>Andhra Pradesh Randomized Evaluation Study</td>
</tr>
<tr>
<td>CARA</td>
<td>Coefficient of Absolute Risk Aversion</td>
</tr>
<tr>
<td>CDF</td>
<td>Cumulative Density Function</td>
</tr>
<tr>
<td>EI</td>
<td>Educational Initiatives</td>
</tr>
<tr>
<td>EVS</td>
<td>Environmental Sciences</td>
</tr>
<tr>
<td>MCs</td>
<td>Mandal Coordinators.</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

We are grateful to Caroline Hoxby, Michael Kremer, and Michelle Riboud for their support, advice, and encouragement at all stages of this project and for the invaluable counsel from Michael Carter and Julian Schweitzer. We are grateful to the various officers of the Governments of Andhra Pradesh and India, for their continuous support and long-term vision for this research effort. In particular, we would like to thank S. Balasubramanyam, K. Chandramouli, A. Giridhar, P. Krishnaiah, I.V. Subba Rao, Vrinda Sarup, Ajay Seth, Prashant Singh, Anirudh Tewari, and C.B.S. Venkataramana.

This research effort would not have been possible without the Azim Premji Foundation as the main implementing partner, and we are particularly grateful to D.D. Karopady, M. Srinivasa Rao, Dilip K. Ranjrekar, and project staff for their outstanding work in managing all practical aspects of the project. Educational Initiatives Inc. played a critical role by taking a lead in the design of tests and we are grateful to Sridhar Rajagopalan, Vyjyanthi Shankar, and staff of Education Initiatives for their contributions. Richman Dzene, Gokul Madhavan and Palak Sikri provided excellent research assistance.

We thank George Baker, Efraim Benmelech, Nazmul Chaudhury, Roger Cunningham, Amit Dar, Jishnu Das, Shanta Devarajan, Martin Feldstein, Richard Freeman, Sangeeta Goyal, Robert Gibbons, Edward Glaeser, Richard Holden, Asim Khwaja, Venita Kaul, Benjamin Loevinsohn, Sendhil Mullainathan, Lipika Nanda, Ben Olken, Priyanka Pandey, Lant Pritchett, Halsey Rogers, Nistha Sinha, A.B.L. Srivastava, Tara Vishwanath, Michael Ward, Jeff Williamson, and various seminar participants for useful comments and discussions. Financial support for the project was provided by the British Department for International Development (DFID), the Government of Andhra Pradesh, and the World Bank. Karthik Muralidharan thanks the Bradley and Spencer Foundations for fellowship support.
## CONTENTS

1. Introduction ............................................................................................................................. - 2 -
2. Theoretical Framework .......................................................................................................... - 6 -
   2.1 Incentives and intrinsic motivation ........................................................................... - 6 -
   2.2 Multi-task moral hazard ........................................................................................... - 7 -
   2.3 Group versus Individual Incentives ............................................................................ - 9 -
3. Experimental Design ........................................................................................................... - 11 -
   3.1 Context ....................................................................................................................... - 11 -
   3.2 Sampling .................................................................................................................... - 11 -
   3.3 AP RESt Design Overview ....................................................................................... - 12 -
   3.4 Implementation of Treatments ................................................................................ - 16 -
4. Test Design and Process ...................................................................................................... - 16 -
   4.1 Test Construction ....................................................................................................... - 16 -
   4.2 Basic versus higher-order skills ............................................................................. - 17 -
   4.3 Incentive versus non-incentive subjects ................................................................... - 18 -
5. Results .................................................................................................................................. - 18 -
   5.1 Sample Balance, Attrition, and Validity of Randomization ...................................... - 18 -
   5.2 Specification .............................................................................................................. - 19 -
   5.3 Impact of Incentives on Test Scores ......................................................................... - 20 -
   5.4 Robustness of results across sub-groups ................................................................. - 21 -
   5.5 Mechanical versus Conceptual Learning and Non-Incentive Subjects ...................... - 22 -
   5.6 Heterogeneity of Treatment Effects ......................................................................... - 23 -
   5.7 Group versus Individual Incentives .......................................................................... - 24 -
6. Process Variables .................................................................................................................. - 25 -
7. Inputs Treatments & Cost-Benefit Analysis ......................................................................... - 28 -
   7.1 Effectiveness of Para-teachers and Block grant ....................................................... - 28 -
   7.2 Comparison of Input and Incentive Treatments ....................................................... - 28 -
   7.3 Comparison with Regular Spending Patterns ......................................................... - 29 -
8. Stakeholder Reactions and Policy Implications ................................................................... - 31 -
   8.1 Stakeholder Reactions ............................................................................................... - 31 -
   8.2 Policy Implications .................................................................................................... - 32 -
9. Conclusion ............................................................................................................................ - 32 -
References................................................................................................................................. - 36 -
Appendix A: Project Timeline and Activities......................................................................... - 39 -
Appendix B: Project Team, Test Administration, and Robustness to Cheating ............... - 40 -
Tables

Table 1: Sample Balance Across Treatments ................................................................. - 41 -
Table 2: Impact of Incentives on Student Test Scores................................................... - 42 -
Table 3: Impact of Incentives by Grade ..................................................................... - 43 -
Table 4: Impact of Incentives by Testing Round ............................................................ - 43 -
Table 5: Raw Scores (% Correct) by Mechanical and Conceptual .................................. - 44 -
Table 6: Impact of Incentives on Mechanical Versus Conceptual Learning ............... - 44 -
Table 7: Impact of Incentives on Non-Incentive Subjects ........................................... - 44 -
Table 8: Impact of Group Incentives versus Individual Incentives .............................. - 45 -
Table 9: Impact of Measurement on Test Score Performance (Control versus Pure Control) - 45 -
Table 10: Process Variables (Based on Classroom Observation) .................................. - 46 -
Table 11: Process Variables (Based on Teacher Interviews) ....................................... - 47 -
Table 12: Impact of Inputs on Learning Outcomes ..................................................... - 47 -
Table 13: Comparing Inputs and Incentives on Learning Outcomes ......................... - 48 -
Table 14: Spending by Treatment ............................................................................. - 48 -
Table 15: Distribution of Incentive Payments ............................................................... - 48 -
Table 16: Comparison with Regular Pattern of Spending .......................................... - 49 -

Figures

Figure 1a: Andhra Pradesh .................................................................................. - 49 -
Figure 1b: District Sampling (Stratified by Socio-cultural Region of AP) ................. - 49 -
Figure 2a: Density/CDF of Normalized Test Score Gains by Treatment ................... - 50 -
Figure 2b: Incentive Treatment Effect by Percentile of Endline Test Scores .............. - 50 -
Figure 3a: Test Calibration – Range of item difficulty ............................................. - 51 -
Figure 3b: Incentive versus Control School Performance – By Question Difficulty ...... - 51 -
Teacher Incentives in Developing Countries: Experimental Evidence from India

Karthik Muralidharan
Department of Economics
Harvard University

Venkatesh Sundararaman
South Asia Human Development
The World Bank

Abstract

We report results from the first year of 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. 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. Students in incentive schools scored significantly higher on “conceptual” as well as “mechanical” questions 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. Teacher absence did not differ across treatments, but teachers in incentive schools appear more likely to have engaged in various measures of teaching activity conditional on being present. We do not find a significant difference in the effectiveness of “group” versus “individual” incentives. Incentive schools performed significantly better than other randomly-chosen schools that received additional schooling inputs of a similar value.
1. Introduction

While the focus of primary education policy in developing countries such as India has typically been on access, enrollment, and retention; much less attention has been paid to the quality of learning. The recently published *Annual Status of Education Report* (Pratham, 2005) showed 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.

The standard policy response in developing countries to try and improve education has been to provide more inputs – typically expanding 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 educational 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). The first was to provide 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 inputs (teacher and non-teacher) and two types of incentives (group and individual), 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 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 4 treatment groups and 100 control schools serving as the comparison group. This paper presents

---

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 Hanushek (2003) argues that the cross-country evidence does not show any systematic relationship between levels of school spending and educational outcomes.
3 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, and we have served as technical consultants in designing and evaluating the various interventions.
results from the first year (2005 – 06) of all four interventions, but focuses on the teacher incentive programs.  

We find that offering to pay bonuses to teachers (the mean payment was 3% of annual base pay) on the basis of the average improvement of test scores of their students had a significant positive impact, with incentive schools scoring 0.19 and 0.12 standard deviations higher than 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 through 5), across both previous year’s and current year’s competencies, 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.

Incentive schools do significantly better on both mechanical components of the test that reflect rote learning and conceptual components of the test that were designed to capture deeper understanding of the material, suggesting that the gains in test scores represent an actual increase in learning outcomes. Additional endline assessments were conducted for science and social studies (on which there were no incentives), and the incentive schools also performed significantly better on these subjects.

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

To isolate the impact of 'measurement' from that of the treatments, 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 "control" schools (who had a baseline test, were presented with feedback on their baseline performance, were subject to continuous tracking surveys by enumerators, and knew about the endline tests in advance) and the "pure control" schools (who receive none of these) suggesting that 'measurement without any consequences' had no impact on learning outcomes.

---

4 Detailed results of the input treatment effects are presented in a companion paper.
Teacher absence did not differ across treatments. Nor did teacher activity (measured by classroom observation) vary between incentive and control schools. However, we find that teachers in control schools (who were observed once a month) were more likely to be engaging in various measures of teaching activity than teachers in a sample of 'pure control' schools outside the study (who were only observed once during the year and never revisited), even though there is no difference in test scores between the two groups. This suggests that observation affected the teaching processes while teachers were being observed but had no effect on test scores. Thus, 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. The teacher interviews, however, 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 two input programs also had a positive and significant impact. Student test scores in input schools were 0.09 standard deviations higher than those in control schools. However, the incentive program cost the same amount in bonuses paid and students in incentive schools scored 0.06 standard deviations higher than students in input schools, with the difference being significant at the 10% level. Since bonuses are another way of paying a salary, the long-run cost of the incentive program is not the bonus payment itself, but the risk premium associated with variable pay. A conservative estimate of the risk premium is 10% of the mean bonus payment, suggesting that an incentive program could have substantially lower costs in the long run than those of making additional bonus payments 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 the total wage budget were to be held constant. We also find that the extent of teachers' support for performance pay is positively correlated with their ex post performance. This suggests that effective teachers (as measured by test scores) know who they are and that performance pay might not only increase effort among existing teachers, but systemically draw more effective teachers into the profession over time.5

5 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.
Our results contribute to a small but growing literature on the effectiveness of performance-based pay for teachers. Unfortunately, a majority of such programs have been implemented in ways that make it difficult to construct a statistically valid comparison group against which the impact of the schemes 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 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 incentives based on test scores and find that, while test scores go up in program schools in the short run, the effect did not remain 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.

This paper attempts to answer the following questions in the context of primary education in developing countries: (i) Can teacher incentives based on test scores improve student achievement? (ii) Do these gains in test scores represent genuine increases in learning outcomes? (iii) Should the incentives be implemented at the school level or at the teacher level? (iv) What is the impact of simply measuring students' achievement without attaching incentives? (v) How does teachers' behavior change? (vi) Are teacher incentive programs cost effective? and (vii) Will teachers support the idea?

We make original contributions in answering each of these questions. We present results from the first randomized evaluation of teacher incentives in a representative sample of schools. We take the test design seriously and including 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 from that of incentives by including a set of "pure control" schools whose outcomes are compared to those in "control" schools. We record differences in teacher behavior with both direct observations based on tracking surveys and teacher interviews after the first year of the program. We study both input and incentive based policies in the same field experiment and calibrate the spending

---

6 Previous studies include Ladd (1999) in Dallas, Dee and Keys (2004) in Tennessee, and Atkinson et al (2004) in the UK. See Umansky (2005) and Vegas and 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.

7 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.
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.

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; but it is also possible that the gains in test scores may not persist in future years. Nor do we know whether new and unanticipated dimensions of gaming will emerge as teachers become more familiar with the program. AP RESt will continue until 2011, and we hope to answer these and other questions in the coming years of continuing a similar experimental design 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 treatments and the experimental design, while section 4 discusses the test design. Sections 5 and 6 present results of the incentive programs on test score outcomes and school process variables respectively. Section 7 presents 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 (especially of small amounts) can sometimes crowd out intrinsic motivation and lead to inferior outcomes.8 Teaching is often thought to be especially susceptible to this concern since many teachers enter the profession due to strong intrinsic motivation. The AP context, however, suggested that an equally valid concern was

---

8 A classic reference in psychology is Deci and Ryan (1985). References in economics include Frey and Oberholzer-Gee (1997), and Gneezy and Rustichini (2000). Frey and Jegen (2001) provide a review of empirical demonstrations of motivation crowding out due to financial incentives. For an excellent discussion of the relevance of intrinsic motivation to practical incentive design and communications, see Chapter 5 of Baron and Kreps (1999).
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.9

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.10 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 \( \varepsilon \) is noise in \( V \).

If the principal can observe \( V \) (but not \( \mathbf{a} \)) and offers a linear wage contract of the form \( w = s + b_\varepsilon \cdot V \) and if the agent's expected utility is given by:

\[
E(s + b_\varepsilon \cdot V) - h \cdot \text{var}(s + b_\varepsilon \cdot V) - \sum_{i=1}^{n} a_i^2 / 2
\]

---

9 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.
10 See Gibbons (1998) and Prendergast (1999) for general overviews of the theory and empirics of incentives in organizations that cover many of the themes discussed in this section. Dixit (2002) provides a discussion of these themes as they apply to public organizations.
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.
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 ($b_i^*$) is given by:

$$ b_i^* = \frac{F^2}{F^2 + 2h\sigma_e^2} $$ (2.2.1)

where $F = \sqrt{\sum_{i=1}^n f_i^2}$, and the expression reflects the standard trade-off between risk and aligning of incentives, with the optimal slope $b_i^*$ being lower as $h$ and $\sigma_e^2$ 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:

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

Since $g \neq f$, $P$ 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:

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

where $G = \sqrt{\sum_{i=1}^n g_i^2}$, and $\theta$ is the angle between $f$ and $g$. $\cos \theta$ is a measure of how much $b_p^*$ needs to be reduced relative to $b_i^*$ due to the distortion arising from $g \neq f$. 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, $\cos \theta < 1$, 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 $V(a(b_p > 0)) > V(a(b_p = 0))$, even
if there is some deviation relative to the first-best action in the absence of distortion and 
\( V(a(b^*_y)) < V(a(b^*_y)) \).\(^{12}\)

There are several reasons for why test scores might be an adequate performance measure in this 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 the private schools have.\(^{13}\) 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 \( w = \text{wage}, P = \text{Performance Measure}, \) and \( c(a) = \text{cost of exerting effort} a \) with \( c'(a) > 0, c''(a) > 0, P'(a) > 0, \) and \( P''(a) < 0 \). Unlike typical cases of team production, individual teachers’ output (test scores of their students) is observable and individual incentive contracts are feasible. The optimal effort for a teacher facing individual incentives is to choose \( a_i \) so that:
\[
\frac{\partial w}{\partial P_i} \cdot \frac{\partial P}{\partial a_i} = c'(a_i) \quad (2.3.1)
\]

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:

\(^{12}\) 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 on teacher attendance instead of test scores and find a significant increase in both teacher attendance and student test scores though teachers don't teach more when they are in school. This is a promising option because the multi-tasking problem is 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 by enabling disadvantaged students to compete with privileged students for scarce slots in higher levels of education on a more even playing field.
If the same bonus is paid to a teacher for a 'unit' of performance under both group and individual incentives then
\[
\frac{\partial w_i}{\partial \left[\frac{P_i + \sum P_j}{n}\right]} \frac{\partial \left[\frac{P_i + \sum P_j}{n}\right]}{\partial a_i} = c'(a_i) \tag{2.3.2}
\]

Since \( c''(a) > 0 \), the equilibrium effort 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 function and this function does 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 perform better 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 this, it would be equivalent to a reduction in \( c'(a) \) relative to a situation where each teacher had to teach all students/subjects. Since \( P''(a) < 0 \), the equilibrium effort could be higher and the outcomes under group incentives might be superior to those under individual incentives. Lavy (2004) reports that the Israeli teacher incentive program at the individual level was more effective than the one at the group level. Since the context here is completely different, we decided to study both group and individual incentives.

\(^{14}\) Holmstrom (1982) is the seminal reference on 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.
3. Experimental Design

3.1 Context

Andhra Pradesh (AP) is the 5th largest 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 various 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 (Figure 1b) and a total of 23 districts. 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 and there are a total of over 60,000 such schools in AP. Over 80% of children in rural AP attend government-run schools.

The average rural primary school is quite small, with an enrollment of around 80 to 100 students and 3 teachers. 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 typically determined by experience and rank, with minor adjustments based on postings, but no component based on any measure of performance. The average salary of regular teachers is over Rs. 7,500/month and total compensation including benefits is close to Rs. 10,000/month (per capita income in AP is around Rs. 2,000/month). Regular teachers' salaries and benefits comprise over 95% of non-capital expenditure on primary education in AP. Teacher unions are strong and disciplinary action for non-performance is rare.

3.2 Sampling

We sampled 5 districts across the 3 socio-cultural regions of AP proportional to population (Figure 1b). 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

17 This is a consequence of the priority placed on providing access to primary school to all children within a walking distance of 1 kilometer.
18 A regression of teacher salary on experience and rank (in our sample) has an R-squared of 0.8
19 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.
20 Subject to the selected districts within a region being contiguous for ease of logistics and supervision.
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>
<thead>
<tr>
<th>INPUTS (Unconditional)</th>
<th>INCENTIVES (Conditional on Average Improvement in Student Learning Measured by Test Scores)</th>
<th>NONE</th>
<th>GROUP BONUS</th>
<th>INDIVIDUAL BONUS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without Baseline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>With Baseline</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NONE</td>
<td>PURE CONTROL (100 Schools)</td>
<td>100 Schools</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXTRA PARA TEACHER</td>
<td>CONTROL (100 Schools)</td>
<td>100 Schools</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXTRA BLOCK GRANT</td>
<td></td>
<td>100 Schools</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As the table shows, the inputs were provided *unconditionally* to the selected schools at the beginning of the first school year, while the incentive treatment 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 baseline tests were conducted in late June and early July, 2005. After the baseline tests, we randomly allocated 2 out of the 10 project schools in each mandal to one of five treatment cells. Since 50 mandals were chosen across 5 districts, there were a total of 100 schools (spread out across the state), in each treatment cell. The geographic stratification implies that every mandal was an exact microcosm of the overall study and allows us to estimate the treatment impact with mandal-level fixed effects and to thereby 'net out' any common factors at the lowest administrative level of government.

The 500 schools in the main study operated under the same conditions of information and monitoring except for the programs so that a comparison between treatment and control schools can

---

21 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).
accurately isolate the treatment effect. To isolate the impact of 'measurement' from those of the treatments, we planned in advance that the endline tests would also be conducted in an additional 100 schools (outside the main study) that we refer to as 'pure control' schools. In effect, these additional 100 schools constitute an extra treatment cell and comparing process and outcome variables between 'control' and 'pure control' schools allows us to estimate the impact of measurement on these variables.

The details of the input and incentive treatments are as follows:

3.3.1 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. 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. 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. They usually teach their own classes and are not 'teacher-aides' who support a regular teacher in the same classroom. There is some evidence that para-teachers are more cost effective than regular teachers but the use of para-teachers is a controversial issue. Proponents arguing that they 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 each of the input interventions was Rs. 10,000/school. A typical government school has a little over 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

---

22 The identity of these schools was unknown till a few weeks before the endline tests, which is when the sampling of additional schools was done (2 extra schools were sampled in each mandal).
23 See the 3 case studies in Pritchett and Pande (2006) for a detailed discussion on para-teachers in India.
24 See Banerjee et al (2005) for evidence in an urban setting in India. They attribute the cost effectiveness of the para-teacher intervention they studied to different pedagogy and not to class-size reductions. De Laat and Vegas (2005) present evidence from Togo showing that students of contract teachers do worse than students of regular civil service teachers. Of course para-teachers could be cost effective even if they perform worse than regular teachers because they cost much less, and it is possible to hire more of them.
spending of Rs. 10,000/year is a little over 3% of the typical annual variable cost of running a government school.

3.3.2 Extra Block-Grant

The block grant intervention targeted inputs directly used by students.\textsuperscript{25} The schools had the freedom to decide how to spend the block grants, 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.\textsuperscript{26} 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.

3.3.3 Group and Individual Incentives

Teachers in incentive schools were offered bonus payments on the basis of the average improvement in test scores\textsuperscript{27} of students taught by them subject to a minimum improvement of 5%. The exact bonus formula was:

\[
\text{Bonus} = \begin{cases} 
\text{Rs. 500} \times (\% \text{Gain in average test scores} - 5\%) & \text{if Gain} > 5\% \\
0 & \text{otherwise}\end{cases}
\]

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)

\textsuperscript{25} See Filmer and Pritchett (1999) for a political economy model that attempts to explain why education spending goes disproportionately towards inputs favored by teachers as opposed to inputs that yield the highest return in terms of education outcomes. An implication of this hypothesis is that the marginal returns to additional spending on student inputs could be higher than the returns to additional spending on regular teacher salaries.

\textsuperscript{26} 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.

\textsuperscript{27} This formula corresponds to a simple measure of the teacher's 'value-added'. There are more sophisticated formulae for estimating teachers' value-added which might be better for longer-term policy, but would not have been transparent to the teachers at the beginning of the program. By focusing on average 'improvement' of all the students, we use an intuitive measure of teacher performance that avoids the many pitfalls of focusing only on 'levels' of learning.

\textsuperscript{28} 1\textsuperscript{st} 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 2\textsuperscript{nd} grade children in the school, and the 'gain' for grade 1 was calculated relative to this. The 5\% threshold did not apply to the 1\textsuperscript{st} grade.
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). The 'slope' of Rs. 500 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 pilot phase of the project). 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.

We 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. If the function transforming teacher effort into test-scores is concave (convex) in the baseline score, teachers have an incentive to focus on weaker (stronger) students, but no student is likely to be wholly neglected since each contributes to the class average. In order to discourage teachers from excluding students with weak gains from the testing, we assigned a zero improvement score to any child who took the baseline test but not the endline test. We made cheating as difficult as possible by having the tests conducted by external teams of 5 evaluators in each school (1 for each of grades 1-5) and having the grading done at a supervised central location at the end of each day's testing.

---


30 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 Lazear and Rosen (1982) and Green and Stokey (1983) for a discussion of tournaments and when they dominate contracts.

31 The convexity in reward schedule due to the threshold could have induced some gaming, but we don't see any evidence that this happened (the distribution of mean class and school-level gains does not have a gap below the threshold). However, 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.

32 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.
3.4 Implementation of Treatments

After the randomization, members of the APF project team 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. Schools selected for the incentive programs were given detailed letters and verbal communications explaining the incentive formula. 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 team, and the materials were jointly procured by the teachers and the APF mandal coordinators and provided to the schools by September. Schools selected to receive the para-teacher took a few weeks to hire and appoint the para-teacher and typically had them in place by September. The appointment of the para-teacher was independent of the appointment, posting, and transfer of regular teachers, and was done after the 'rationalization' of teacher postings for the year was completed (usually determined by August) to ensure that there would be no retrospective adjustment of teacher postings which could compromise the randomization.

4. Test Design and Process

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 of the syllabus from the text books into skills, creating of a universe of questions to represent the skills, and calibrating the difficulty of the questions in a pilot exercise in 40 schools during the school year prior to the main project (2004-05). Item characteristic curves were created for each question and the final test comprised of questions of high discrimination at a wide range of difficulties.

The baseline test (June – July, 2005) covered competencies of the previous school year. At the end of the school year (March – April, 2006), schools had two rounds of tests with a

---

33 Sample communication letters are available from the authors on request.
34 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 in the test. Figure 3a shows the range of difficulty of test items (questions are sorted by difficulty, with the y-axis showing the fraction of students who got the question correct).
gap of two weeks between them. The first test (the 'endline') tested the same set of skills from the baseline (that is, the previous year's competencies) with an exact mapping of question type from the baseline to 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.

4.2 Basic versus higher-order skills

As highlighted in section 2, it is possible that even if it is true that $V(a(b_p > 0)) < V(a(b_p = 0))$ even if it is true that $P(a(b_p > 0)) > P(a(b_p = 0))$ and so a key empirical question is whether additional efforts taken by teachers to improve test scores for primary school children in response to the incentives are also likely to lead to improvements in broader educational outcomes. We asked EI to design the tests to include both 'mechanical' and 'conceptual' questions within each skill category on the test. The distinction between these two categories is not constant, since a conceptual question that is repeatedly taught in class can become a mechanical one. Similarly a question that is conceptual in an early grade might become mechanical in a later grade, if students have gotten acclimatized to the idea over time. For the purpose of 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: \[
\begin{align*}
34 
\times 5
\end{align*}
\]

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

\[
8 + 8 + 8 + 8 + 8 + 8 = 8 \times \_
\]

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. It is arguably 'easier' in this example 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 questions are '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? __________

![9cm x 9cm square](image)

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

![Rectangle with square cut off](image)

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.\(^\text{35}\)

### 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 Sample Balance, Attrition, and Validity of Randomization

\(^{35}\) 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 paper attempts to address this concern.
Table 1 shows summary statistics of baseline school and student performance variables by treatment. 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).

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 (rows 1 and 2 of bottom panel in Table 1).

Similarly, we also consider attrition of students and changes in class composition between the baseline and endline. The average attrition rate in the sample is 14.2%, and there is no significant difference in attrition across the treatments (bottom row of Table 1). 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 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:

\[ T_{ijkm}(EL) = \alpha + \gamma \cdot T_{ijkm}(BL) + \delta \cdot \text{Incentives} + \beta \cdot Z_m + \varepsilon_k + \varepsilon_{jk} + \varepsilon_{ijk} \]  \hspace{1cm} (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. 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 $\delta$ 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 differences approach (DID) to estimate $\delta$ as follows:

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

The DID specification constrains the coefficient on the lagged test score ($\gamma$) 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.

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 $\delta$ from specification (5.1) is 0.154 standard deviations (SD). Panel B shows the difference-in-difference estimate of $\delta$ to be 0.169, which is not significantly different from 0.154. However, the data strongly rejects $\gamma = 1'$ and so we prefer specification (5.1) because it does not

---

36 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.

37 The randomization implies that the estimates of $\delta$ 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.
constrain the coefficient on the lagged score to be equal to one. 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 \( \delta \) in any of the regressions as would be expected if the randomization were valid. 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 \( \delta \) 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 \( (T_{ijkl}(EL) - T_{ijkl}(BL)) \) 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 vertical distance between the two plots is positive at every percentile but increasing. In other words, gains are higher 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.

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:

\[
T_{ijkl}(EL) = \alpha + \gamma \cdot T_{ijkl}(BL) + \sum_{i=1}^{n} \delta_i \cdot (\text{Incentives} \times \text{Category}_i) + \beta \cdot Z_m + \varepsilon_k + \varepsilon_{jk} + \varepsilon_{ijk}
\]

followed by an F-test of equality across the \( \delta_i \)'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 \( \delta_i \)'s are jointly equal to 0. Table 3 presents the effect of incentives on performance by grade (1-5). The

38 The estimate of \( \gamma \) is biased due to measurement error and omitted ability, but the data strongly rejects \( \gamma=1 \). Once we have data for more than 1 year, we can use the panel data to estimate \( \gamma \) consistently (see Todd and Wolpin, 2002 for details).
estimate of $\delta$ 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 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 endline (though not significantly so), which is consistent with our finding from teacher interviews (see next section for details) 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 correct by grade, subject, and mechanical/conceptual, 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 consider the normalized test scores for the mechanical and conceptual component of each test separately. Incentive schools do significantly better on both the mechanical and conceptual components of the test and the estimate of $\delta$ is almost identical across both components (Table 6). The other interesting result is that the coefficient on the lagged score is significantly lower for the conceptual component, indicating that these questions represented 'unfamiliar' territory relative 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 separately, and once again the parameter of interest is $\delta$. 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). Again, the coefficients on the lagged scores here are much lower than those in Table 2 and also lower than those on the conceptual components in Table 6, confirming that the domain of these tests was substantially different from 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, efforts at increasing test scores in math and language also contribute to superior performance on broader educational outcomes suggesting complementarily in the various measures (though the result could also be due to an improvement in test-taking skills).

5.6 Heterogeneity of Treatment Effects

39 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.
We test for heterogeneity of the incentive treatment effect across student, teacher, and school characteristics by testing if $\delta_1$ is significantly different from zero in:

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

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 or the student's baseline score.\(^{40}\)

Similarly, we find no evidence of differential impact of incentives by teacher characteristics including gender, designation, experience, or 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 older and more 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 (the index is described in Table 1). The mean infrastructure index in the sample is 3.26, 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**

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). We find no significant impact of the size of the school on the

---

\(^{40}\) 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.
relative performance of group and individual incentives (both linear and quadratic interactions of school size with the group incentive treatment are insignificant). 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. Field reports suggest that teachers in some individual incentive schools agreed to split their bonus amounts equally ex post.

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, our results suggesting near equivalence between individual and group incentives should not be extrapolated to large schools in which the group would comprise many teachers.

5.8 Impact of measurement

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

\[ T_{ijk} (EL) = \alpha + \delta \cdot Control + \beta \cdot Z_{m} + \varepsilon_{i} + \epsilon_{jk} + \epsilon_{ijk} \]

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 a little surprising because our initial hypothesis was that the 'announcement effect' of external testing would have a larger impact. However, all 500 schools were told that information about a specific school or teacher's performance would not be shared outside the school on an identifiable basis. 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.\(^{41}\)

6. Process Variables

\(^{41}\) 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.
The mandal coordinators (MCs) from APF made unannounced visits to each of the 500 schools around once a month in the period from September 2005 to February 2006 to conduct 'tracking surveys' where they collected data on various process variables at the school level including child attendance, teacher attendance and activity, and classroom observation of teaching processes. To code classroom processes, an MC typically spent between 20 and 30 minutes at the back of a classroom without disturbing the class and coded whether specific actions took place during the period of observation. Similar unannounced visits were made to other randomly-sampled schools in each of the mandals that were not in the main study – with such schools being visited only once during the year.\(^{42}\) 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). In addition to the tracking surveys, the MCs also interviewed teachers about aspects of their teaching during the school year, 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 of the incentive program were announced.

There was no difference in either student or teacher attendance between control and incentive schools (Table 10 – Panel A). We also find no significant difference between incentive and control schools on any of the various indicators of classroom processes as measured by direct observation. This is similar to the results in Glewwe et al (2003) and Duflo & Hanna (2005) where they find no difference in process variables between treatment and control schools from similar surveys\(^{43}\) and raises the question of how the outcomes are significantly different when there don't appear to be any differences in the processes in the schools. One possible explanation is that teachers' actions converge to certain norms of behavior when under repeated observation.\(^{44}\)

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

\(^{42}\) 300 extra schools were sampled (6 in each mandal) and each school was visited once during the period from December 2005 to February 2006 and never visited again for tracking surveys. 100 of these schools were sampled in early March 2006 to be tested in the endline. Thus the test scores for the 'pure control' schools are for these 100 schools, while the process variables are measured from around 300 schools.

\(^{43}\) With the exception of teacher attendance in Duflo & Hanna (2005) where the incentive was based on teacher attendance verified with a camera.

\(^{44}\) Other explanations could be that the measures of 'teaching activity' we use are not capturing meaningful data on classroom processes, or that teachers in incentive schools only increase effort close to the end of the school year, which we do not observe during our monthly surveys between September and February.
seems likely that the superior process measures in the control schools relative to pure 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 shows up), 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,45 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 process differences, since the 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 how they had done. Table 11 summarizes the differences in self-reported process variables across both types of schools and we see that teachers in incentive schools are significantly more likely to have assigned more homework and more class work, conducted extra classes beyond regular school hours, given practice tests, and paid special attention to weaker children. While self-reported measures of teacher activity might be considered less 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.

It appears that the incentive program did not change the teachers' cost-benefit calculations with regard to the decision to be absent, but that it did make them more likely to be actively engaged in various teaching activities when present.46 Regular observation changed process variables but not outcomes. The results here combined with those in Duflo & Hanna (2005) support the theoretical prediction that behavior will shift along margins where there is an incentive and not on others.

45 A striking example is provided in Brennan and Pettit (2005) who present experimental evidence that women in 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, even though this person was typically a stranger who was not observing them carefully and who they would probably never encounter again.

46 A teacher's decision to be absent on a given day can be thought of as a binary indicator of a latent cost-benefit variable that is subject to daily shocks such as illness of self or family, problems with public transport, changing daily value of outside options etc. The results suggest that the incentives on the basis of test scores did not result in any significant difference on the margin of absence, but that teachers probably compensated by working harder conditional on being present. This is different from the pattern observed in Kremer et al (2006) who find that schools with high absence also have lower teaching activity conditional on a teacher being present – which is more consistent with a story of absence as shirking, as opposed to one of 'efficient absence' with compensatory effort conditional on being present.
7. Inputs Treatments & Cost-Benefit Analysis

7.1 Effectiveness of Para-teachers and Block grant

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 their relation to the education production function literature are presented in our companion paper (Muralidharan and Sundararaman, 2006b).

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:

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

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 SD 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 differential is greater (and significant at the 5% level) for the higher endline test than for the endline test (where the difference is not significant) and 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.

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.

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 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 of 0.22% of base pay even if the CARA is as high 5.47 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 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
these interventions is to compare their outcomes to those from the 'business as usual' pattern of
spending. Since the skills and competencies assessed on the endline test correspond exactly to those
tested on the baseline, we can consider the gain in scores in the control schools between the endline
and baseline (but not the 'higher endline') to represent the progress made in one school year in a

47 The risk premium here is the value of ε such that 0.5[μ(0.97w + ε) + u(1.03w + ε)] = 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.
typical government school. The gains in the incentive and input schools can then be compared with a year's progress in the average government school.

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 therefore around Rs. 300,000/year. The input treatments added the equivalent of 16% of a normal year's learning, and so 6.25 years of the programs would add an equivalent of a full year's learning at around 21% of the current variable cost of a year's schooling. 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.

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. This is probably an over estimate because the small size and geographical dispersion of the program

---

48 The baseline and 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 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.

49 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.

50 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 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.

51 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.
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 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 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 revenue-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.\(^{52}\)

To obtain data on parental preferences, APF team members 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

---

\(^{52}\) 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.
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 creatively combine them 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 will also produce a track record of teacher performance, and this data can 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.53 While the reactions of teacher unions to such proposals could be different from that of individual teachers,54 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 the potential entry of better teachers into the profession.55 We regress the extent of teachers' preference for performance pay (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

53 A similar proposal is made in Pritchett and Pande (2006).
54 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.
55 See Lazear (2000) and (2003), and Hoxby and Leigh (2005)
significant performance improvements in both math and language test scores. The gains were spread out evenly across grades, districts, test competency, 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 change in incentive schools, but teachers in incentive schools appear to teach more 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 substantially more cost efficient 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 incentives is only the risk premium 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% 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, and there are many questions that cannot be answered with just one year's data. We do not know at this point if the gains demonstrated in the incentive schools will be sustained in future years and if they will be retained if the incentive program is stopped. We also do not know how a complete understanding of the dynamics of student learning will affect a perfectly rational teacher's response to an incentive program in the long run, and if new 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 of 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 there may be political pressure on the part of teachers to convert bonuses into 'raises', which would defeat the point of performance-based pay.
We also do not claim to have 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 ($\gamma$) is estimated as 0.5 as opposed to 1, suggests that a naïve 'gain' formulation penalizes 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, which an optimal formula would 'net out' to estimate a more precise measure of teachers' 'value addition' to base bonuses on. A related concern is measurement error and the precision of test scores at the class and school levels.56

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.57 A practical trade off, however, is between the accuracy and precision of the value-added measurement on one hand and the transparency of the system to the teachers on the other. Teachers have accepted the intuitive 'average gain' formula used in the first year, and our expectation is that they will start getting more sophisticated about the formula in future years, and that the decision regarding where to locate on the accuracy-transparency frontier will be made in consultation with teachers. However, it is possible that there may be no satisfactory resolution of the tension between accuracy and transparency.58

AP REST is expected to continue till 2011 and hope to answer many of the questions raised in this section as part of an 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, to experiment with variations such as tournaments, incentives on both teacher attendance and test scores, interactions between inputs and incentives, and variations in incentive amounts to estimate outcome elasticity with respect to the incentive amount.

56 Kahne 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.

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

58 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. Similarly, Kremer and Chen (2001) show that performance incentives even for something as objective as teacher attendance did not work when implemented through headmasters in schools in Kenya. The headmaster 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:


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)
- Preparation of all logistics for the baseline tests

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

May – July 2006
- Data entry, cleaning, analysis

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 (M Srinivasa Rao), 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 '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 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 that looked a little suspicious. 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.

---

59 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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>113.2</td>
<td>104.6</td>
<td>104.2</td>
<td>111.3</td>
<td>112.6</td>
<td>0.82</td>
<td>9.0</td>
<td>0.38</td>
</tr>
<tr>
<td>Extra Parateacher</td>
<td>64.9</td>
<td>62.0</td>
<td>62.3</td>
<td>62.0</td>
<td>66.5</td>
<td>0.89</td>
<td>4.5</td>
<td>0.42</td>
</tr>
<tr>
<td>Block Grant</td>
<td>3.07</td>
<td>2.83</td>
<td>3.03</td>
<td>3.12</td>
<td>3.14</td>
<td>0.58</td>
<td>0.31</td>
<td>0.14</td>
</tr>
<tr>
<td>Group Incentive</td>
<td>39.5</td>
<td>39.8</td>
<td>34.6</td>
<td>40.6</td>
<td>37.5</td>
<td>0.66</td>
<td>6.0</td>
<td>0.16</td>
</tr>
<tr>
<td>Individual Incentive</td>
<td>3.19</td>
<td>3.13</td>
<td>3.33</td>
<td>3.14</td>
<td>3.26</td>
<td>0.84</td>
<td>0.21</td>
<td>0.32</td>
</tr>
<tr>
<td>Largest Difference</td>
<td>14.65</td>
<td>14.97</td>
<td>14.71</td>
<td>14.66</td>
<td>14.72</td>
<td>0.98</td>
<td>0.32</td>
<td>0.55</td>
</tr>
<tr>
<td>P-value (Equality of all groups)</td>
<td>0.56</td>
<td>1.8</td>
<td>0.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P-value (Largest Difference=0)</td>
<td>0.57</td>
<td>0.116</td>
<td>0.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 | 0.30 |
| 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 |

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

<table>
<thead>
<tr>
<th></th>
<th>Combined</th>
<th>Math</th>
<th>Telugu (Language)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized Baseline Score</td>
<td>0.499</td>
<td>0.494</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>(0.013)**</td>
<td>(0.013)**</td>
<td>(0.017)**</td>
</tr>
<tr>
<td>Incentive School</td>
<td>0.154</td>
<td>0.153</td>
<td>0.189</td>
</tr>
<tr>
<td></td>
<td>(0.042)**</td>
<td>(0.045)**</td>
<td>(0.049)**</td>
</tr>
<tr>
<td>School and Household Controls</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Observations</td>
<td>68716</td>
<td>54910</td>
<td>34132</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.29</td>
<td>0.30</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Panel A (No constraint on coefficient on the lagged score)

Dependent Variable = Endline Test Score (Normalized by Distribution of Control Schools)

<table>
<thead>
<tr>
<th></th>
<th>Combined</th>
<th>Math</th>
<th>Telugu (Language)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endline * Incentives</td>
<td>0.169</td>
<td>0.181</td>
<td>0.197</td>
</tr>
<tr>
<td></td>
<td>(0.047)**</td>
<td>(0.049)**</td>
<td>(0.054)**</td>
</tr>
<tr>
<td>School and Household Controls</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Observations</td>
<td>137025</td>
<td>104345</td>
<td>68305</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
</tr>
</tbody>
</table>

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

Dependent Variable = Endline Test Score (Normalized by Distribution of Control Schools)

Notes:
1. All regressions include mandal (sub-district) level fixed effects and standard errors clustered at the school level.
2. Constants are insignificant in all specifications and not shown
* significant at 10%; ** significant at 5%; *** significant at 1%
Table 3: Impact of Incentives by Grade

<table>
<thead>
<tr>
<th></th>
<th>Combined</th>
<th>Math</th>
<th>Telugu (Language)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[1]</td>
<td>[2]</td>
<td>[3]</td>
</tr>
<tr>
<td>Incentives * Grade 1</td>
<td>0.105</td>
<td>-0.05</td>
<td>0.109</td>
</tr>
<tr>
<td></td>
<td>(0.064)</td>
<td>(0.065)</td>
<td>(0.069)</td>
</tr>
<tr>
<td>Incentives * Grade 2</td>
<td>0.11</td>
<td>-0.045</td>
<td>0.124</td>
</tr>
<tr>
<td></td>
<td>(0.054)**</td>
<td>(0.061)</td>
<td>(0.058)**</td>
</tr>
<tr>
<td>Incentives * Grade 3</td>
<td>0.181</td>
<td>0.026</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>(0.055)***</td>
<td>(0.045)</td>
<td>(0.062)***</td>
</tr>
<tr>
<td>Incentives * Grade 4</td>
<td>0.188</td>
<td>0.033</td>
<td>0.257</td>
</tr>
<tr>
<td></td>
<td>(0.054)***</td>
<td>(0.046)</td>
<td>(0.067)***</td>
</tr>
<tr>
<td>Incentives * Grade 5</td>
<td>0.155</td>
<td>0.185</td>
<td>0.125</td>
</tr>
<tr>
<td></td>
<td>(0.051)***</td>
<td>(0.063)***</td>
<td>(0.063)***</td>
</tr>
</tbody>
</table>

|                  | [5]      | [6]  |                  |
| 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) | 0.61 | 0.61 | 0.22 | 0.22 | 0.95 | 0.95 |
R-squared | 0.29 | 0.29 | 0.28 | 0.28 | 0.32 | 0.32 |

Notes:
All regressions include mandal (sub-district) level 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

<table>
<thead>
<tr>
<th></th>
<th>Combined</th>
<th>Math</th>
<th>Telugu (Language)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[1]</td>
<td>[2]</td>
<td>[3]</td>
</tr>
<tr>
<td>Incentive School</td>
<td>0.124</td>
<td>0.193</td>
<td>0.153</td>
</tr>
<tr>
<td></td>
<td>(0.042)***</td>
<td>(0.050)***</td>
<td>(0.048)***</td>
</tr>
<tr>
<td>Observations</td>
<td>38802</td>
<td>29914</td>
<td>19274</td>
</tr>
</tbody>
</table>
R-squared | 0.30 | 0.30 | 0.29 | 0.29 | 0.31 | 0.34 |

Notes:
1. All regressions include mandal (sub-district) level fixed effects and standard errors clustered at the school level.
2. The 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) by Mechanical and Conceptual

<table>
<thead>
<tr>
<th>Grade</th>
<th>Mechanical</th>
<th>Conceptual</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15.4</td>
<td>18.7</td>
<td>-3.3</td>
</tr>
<tr>
<td>2</td>
<td>38.6</td>
<td>40.0</td>
<td>-1.4</td>
</tr>
<tr>
<td>3</td>
<td>30.4</td>
<td>25.8</td>
<td>4.6</td>
</tr>
<tr>
<td>4</td>
<td>28.4</td>
<td>23.3</td>
<td>5.1</td>
</tr>
<tr>
<td>5</td>
<td>25.2</td>
<td>16.3</td>
<td>8.9</td>
</tr>
</tbody>
</table>

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 M/C Distribution of Control Schools) |
|---------------------------------------------------------------|---------------------------------------------------------------|
| Combined | Math | Telugu (Language) |
| Mechanical | Conceptual | Mechanical | Conceptual | Mechanical | Conceptual |
| 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) level 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 = Endline Test Score (Normalized by Distribution of Control Schools) |
|---------------------------------------------------------------|---------------------------------------------------------------|
| Science | Social Studies |
| 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) level 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 = Endline Score (Normalized by Distribution of Control Schools)

<table>
<thead>
<tr>
<th></th>
<th>Combined</th>
<th>Math</th>
<th>Telugu (Language)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized Baseline Score</td>
<td>0.499</td>
<td>0.49</td>
<td>0.516</td>
</tr>
<tr>
<td></td>
<td>(0.013)**</td>
<td>(0.017)**</td>
<td>(0.014)**</td>
</tr>
<tr>
<td>Group Incentive School (GI)</td>
<td>0.146</td>
<td>0.183</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>(0.050)**</td>
<td>(0.058)**</td>
<td>(0.046)**</td>
</tr>
<tr>
<td>Individual Incentive School (II)</td>
<td>0.162</td>
<td>0.195</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>(0.049)**</td>
<td>(0.060)**</td>
<td>(0.043)**</td>
</tr>
<tr>
<td>Observations</td>
<td>68716</td>
<td>34132</td>
<td>34584</td>
</tr>
<tr>
<td>F-Stat p-value (Testing GI = II)</td>
<td>0.76</td>
<td>0.84</td>
<td>0.66</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.29</td>
<td>0.28</td>
<td>0.32</td>
</tr>
</tbody>
</table>

**Notes:**
All regressions include mandal (sub-district) level 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 = Endline Score (Normalized by Distribution of Control Schools)

<table>
<thead>
<tr>
<th></th>
<th>Combined</th>
<th>Math</th>
<th>Telugu (Language)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1]</td>
<td>0.012</td>
<td>-0.011</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td>(0.045)</td>
<td>(0.048)</td>
<td>(0.044)</td>
</tr>
<tr>
<td>Observations</td>
<td>49248</td>
<td>24615</td>
<td>24633</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
</tbody>
</table>

**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) level fixed effects and standard errors clustered at the school level.
* significant at 10%; ** significant at 5%; *** significant at 1%
### Table 10: Process Variables (Based on Classroom Observation)

<table>
<thead>
<tr>
<th>Process Variable (Activities performed by Teachers unless recorded otherwise)</th>
<th>Panel A</th>
<th>Panel B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Incentive Schools</td>
<td>Control Schools</td>
</tr>
<tr>
<td>Student Attendance</td>
<td>62.9</td>
<td>64.3</td>
</tr>
<tr>
<td>Teacher Absence</td>
<td>24.9</td>
<td>22.5</td>
</tr>
<tr>
<td>Actively Teaching</td>
<td>47.5</td>
<td>49.9</td>
</tr>
<tr>
<td>Clean &amp; Orderly Classroom</td>
<td>60.5</td>
<td>59.5</td>
</tr>
<tr>
<td>Giving a Test</td>
<td>26.6</td>
<td>26.6</td>
</tr>
<tr>
<td>Calls Students by Name</td>
<td>78.5</td>
<td>78.1</td>
</tr>
<tr>
<td>Addresses Questions to Students</td>
<td>62.8</td>
<td>63.2</td>
</tr>
<tr>
<td>Provides Individual/Group Help</td>
<td>37.1</td>
<td>35.7</td>
</tr>
<tr>
<td>Encourages Participation</td>
<td>37.6</td>
<td>37.0</td>
</tr>
<tr>
<td>Reads from Textbook</td>
<td>52.8</td>
<td>56.1</td>
</tr>
<tr>
<td>Makes Children Read From Textbook</td>
<td>57.8</td>
<td>60.0</td>
</tr>
<tr>
<td>Active Blackboard Usage</td>
<td>50.0</td>
<td>49.1</td>
</tr>
<tr>
<td>Assigned Homework</td>
<td>39.5</td>
<td>37.2</td>
</tr>
<tr>
<td>Provided Homework Guidance</td>
<td>33.6</td>
<td>32.9</td>
</tr>
<tr>
<td>Provided Feedback on Homework</td>
<td>24.7</td>
<td>27.0</td>
</tr>
<tr>
<td>Children were Using a Textbook</td>
<td>66.0</td>
<td>67.4</td>
</tr>
<tr>
<td>Children Asked Questions in Class</td>
<td>37.1</td>
<td>37.0</td>
</tr>
<tr>
<td>Teacher Was in Control of the Class</td>
<td>51.2</td>
<td>52.4</td>
</tr>
</tbody>
</table>

**Notes:**

1. The control and incentive schools were each visited by a mandal coordinator around once a month for a total of 6 visits between September 05 and March 06. The "pure control" schools were extra schools that were outside the main study. Around 300 extra schools were randomly sampled (around 6 in each mandal) and tracking surveys were conducted to measure processes in a "typical" school that was outside the study. Each of these schools was visited only once (at an unannounced date) during the entire year. Student attendance was not taken 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: Process Variables (Based on Teacher Interviews)

<table>
<thead>
<tr>
<th>Process Variable</th>
<th>Incentive Schools</th>
<th>Control Schools</th>
<th>p-Value of Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did you do any special preparation for the end of year tests? (% Yes)</td>
<td>74.6</td>
<td>52.7</td>
<td>0.000***</td>
</tr>
<tr>
<td>Extra Homework</td>
<td>57.2</td>
<td>35.8</td>
<td>0.000***</td>
</tr>
<tr>
<td>Extra Classwork</td>
<td>58.7</td>
<td>41.2</td>
<td>0.000***</td>
</tr>
<tr>
<td>Extra Classes/Teaching Beyond School Hours</td>
<td>22.5</td>
<td>11.1</td>
<td>0.000***</td>
</tr>
<tr>
<td>Gave Practice Tests</td>
<td>36.0</td>
<td>21.8</td>
<td>0.000***</td>
</tr>
<tr>
<td>Paid Special Attention to Weaker Children</td>
<td>24.9</td>
<td>10.7</td>
<td>0.000***</td>
</tr>
</tbody>
</table>

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

<p>| Dependent Variable = Test Score (Normalized by Test Score Gain Distribution of Control Schools) |
|-------------------------------------------------|----------------------------------|--------------------|</p>
<table>
<thead>
<tr>
<th>Combined</th>
<th>Math</th>
<th>Telugu (Language)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normalized Baseline Score</td>
<td>0.516</td>
<td>0.516</td>
</tr>
<tr>
<td></td>
<td>(0.012)***</td>
<td>(0.012)***</td>
</tr>
<tr>
<td>Inputs (Pooled)</td>
<td>0.09</td>
<td>0.101</td>
</tr>
<tr>
<td></td>
<td>(0.033)***</td>
<td>(0.038)***</td>
</tr>
<tr>
<td>Extra Para Teacher (PT)</td>
<td>0.089</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td>(0.038)**</td>
<td>(0.043)**</td>
</tr>
<tr>
<td>Block Grant (BG)</td>
<td>0.091</td>
<td>0.098</td>
</tr>
<tr>
<td></td>
<td>(0.039)**</td>
<td>(0.046)**</td>
</tr>
<tr>
<td>Observations</td>
<td>66246</td>
<td>66246</td>
</tr>
<tr>
<td>F-Stat p-value (Testing PT = BG)</td>
<td>0.06</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>(0.039)**</td>
<td>(0.046)**</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.31</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Notes:
- All regressions include mandal (sub-district) level fixed effects and standard errors clustered at the school level.
- * significant at 10%; ** significant at 5%; *** significant at 1%
Table 13: Comparing Inputs and Incentives on Learning Outcomes

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined Math Language</td>
<td>Incentives</td>
<td>0.156 (0.041)**</td>
<td>0.189 (0.049)**</td>
<td>0.124 (0.038)**</td>
<td>0.127 (0.041)**</td>
<td>0.154 (0.047)**</td>
<td>0.100 (0.041)**</td>
<td>0.195 (0.050)**</td>
<td>0.235 (0.064)**</td>
</tr>
<tr>
<td>Inputs</td>
<td>0.096 (0.037)**</td>
<td>0.110 (0.043)**</td>
<td>0.083 (0.036)**</td>
<td>0.097 (0.038)**</td>
<td>0.111 (0.042)**</td>
<td>0.083 (0.039)**</td>
<td>0.095 (0.044)**</td>
<td>0.107 (0.056)**</td>
<td>0.083 (0.039)**</td>
</tr>
<tr>
<td>Difference (Incentives - Inputs)</td>
<td>0.06</td>
<td>0.08</td>
<td>0.04</td>
<td>0.03</td>
<td>0.04</td>
<td>0.02</td>
<td>0.10</td>
<td>0.13</td>
<td>0.07</td>
</tr>
<tr>
<td>F-Stat p-value (Inputs = Incentives)</td>
<td>0.09*</td>
<td>0.06*</td>
<td>0.20</td>
<td>0.40</td>
<td>0.29</td>
<td>0.62</td>
<td>0.02**</td>
<td>0.03**</td>
<td>0.03**</td>
</tr>
<tr>
<td>Observations</td>
<td>112078</td>
<td>55681</td>
<td>56397</td>
<td>63338</td>
<td>31464</td>
<td>31874</td>
<td>48740</td>
<td>24217</td>
<td>24523</td>
</tr>
<tr>
<td>R-squared</td>
<td>0.29</td>
<td>0.27</td>
<td>0.32</td>
<td>0.29</td>
<td>0.28</td>
<td>0.31</td>
<td>0.3</td>
<td>0.28</td>
<td>0.34</td>
</tr>
</tbody>
</table>

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. All Regressions include mandal (sub-district) level fixed effects
* significant at 10%; ** significant at 5%; *** significant at 1%

---

Table 14: Spending by Treatment

<table>
<thead>
<tr>
<th>Table 14: Spending by Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extra Para Teacher</td>
</tr>
<tr>
<td>Annual Spending by Treatment (Rs/School)</td>
</tr>
<tr>
<td>R-squared</td>
</tr>
</tbody>
</table>

---

Table 15: Distribution of Incentive Payments

<table>
<thead>
<tr>
<th>Table 15: Distribution of Incentive Payments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Eligible Schools (Teachers)</td>
</tr>
<tr>
<td>100 (328)</td>
</tr>
<tr>
<td>Number of Schools (Teachers) Qualifying for a Bonus</td>
</tr>
<tr>
<td>Bonus Range (Rs. 0 - 5,000)</td>
</tr>
<tr>
<td>Bonus Range (Rs. 5,000 - 10,000)</td>
</tr>
<tr>
<td>Bonus Range (Rs. 10,000+)</td>
</tr>
<tr>
<td>Average Bonus (among all eligible teachers) (Rs.)</td>
</tr>
<tr>
<td>Average Bonus (among those who receive a bonus) (Rs.)</td>
</tr>
<tr>
<td>Maximum Bonus received by a Teacher (Rs.)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Inputs</th>
<th>Incentives</th>
<th>Input Gain/Control Gain</th>
<th>Incentive Gain/Control Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Math Gain by Grade (in % points over Baseline Score)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grade 1 (Assuming Baseline = 0)</td>
<td>14.4%</td>
<td>17.6%</td>
<td>16.4%</td>
<td>1.22</td>
<td>1.13</td>
</tr>
<tr>
<td>Grade 2</td>
<td>22.6%</td>
<td>28.1%</td>
<td>26.3%</td>
<td>1.24</td>
<td>1.16</td>
</tr>
<tr>
<td>Grade 3</td>
<td>9.3%</td>
<td>10.3%</td>
<td>13.7%</td>
<td>1.11</td>
<td>1.48</td>
</tr>
<tr>
<td>Grade 4</td>
<td>16.2%</td>
<td>19.4%</td>
<td>20.2%</td>
<td>1.19</td>
<td>1.24</td>
</tr>
<tr>
<td>Grade 5</td>
<td>5.8%</td>
<td>6.6%</td>
<td>9.0%</td>
<td>1.15</td>
<td>1.56</td>
</tr>
<tr>
<td>Mean Math Gain Across Grades</td>
<td><strong>1.18</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| 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** |        |            |                         |                             |

| Mean Gain Across Subjects | **1.16** |        |

- 49 -
Figure 1a: Andhra Pradesh (AP)

Figure 1b: District Sampling (Stratified by Socio-cultural Region of AP)
Figure 2a: Density/CDF of Normalized Test Score Gains by Treatment

![Density/CDF of Normalized Test Score Gains by Treatment](image)

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

![Incentive Treatment Effect by Percentile of Endline Test Scores](image)

Quantile Regression Estimates of Incentive Treatment Effect

- Control Schools
- Incentive Schools
- Sample Values
Figure 3a: Test Calibration – Range of item difficulty

Mathematics

Grade 1

Grade 2

Grade 3

Grade 4

Grade 5

Telugu

Grade 1

Grade 2

Grade 3

Grade 4

Grade 5

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