Andhra Pradesh Primary Education – Randomized Evaluation Study Proposal

I. Overview

This study aims to analyze the relative impact on student learning outcomes of input-based and incentive-based education policies using the methodology of a randomized evaluation. Through this analysis we hope to obtain a better understanding of the relative importance of improving the resource base of primary education in India (via provision of additional schooling inputs) versus changing the institutional setting under which the system functions (via provision of bonus payments to teachers based on improvements in student learning).

The immediate motivation for designing and conducting this study was the finding in a series of recently published papers that 25% of teachers in government-run schools in India are absent on any given day and that less than half of them are engaged in any teaching activity during unannounced visits made to a nationally representative sample of schools. Since nearly 90% of primary education spending in India goes towards teacher salaries, these findings suggest substantial inefficiency in the use of the public education budget. As governments in India and other developing countries raise their primary education budgets to meet the schooling targets of the Millennium Development Goals, it becomes especially important to understand the most efficient way to spend these scarce public resources.

In particular, the teacher absence results suggest that we need to be cautious about doing “more of the same” in a couple of important ways. First, in a context where teacher salaries account for the bulk of educational spending, there is very little left for

---

1 Karthik Muralidharan – Department of Economics, Harvard University, and Venkatesh Sundararaman – Economist, South Asia Human Development Unit, World Bank
2 See Kremer et. al (2005) and Chaudhury et. al. (2005) for details regarding teacher absence in India and in a cross-country sample respectively.
3 Teachers in government-run schools in the poorest performing Indian states were absent over 40% of the time and were engaged in teaching activity less than 25% of the time during surprise visits to these schools.
4 The Government of India recently imposed a 2% tax surcharge to finance universal primary education
child-level learning materials such as notebooks, exercise books, and writing materials.\(^5\)

Second, since teacher salaries are completely independent of any measure of performance (even a basic one like teacher attendance), there are potentially large gains to experimenting with new institutional settings including making bonus payments to high-performing teachers to better align the incentives of teachers with the underlying objective of improving learning outcomes.

This research project feeds directly into these first-order policy questions by evaluating the relative returns to additional spending on schooling inputs on one hand, and teacher incentives on the other. In addition, we will also dig deeper into the specifics of these options, by evaluating the effectiveness of both teacher and non-teacher inputs on the “input” side, and evaluating both individual (teacher-level) and group (school-level) teacher incentives on the “incentive” side.

Specifically, we will be testing the following hypotheses in rural government-run\(^6\) primary schools (grades 1-5) in the Indian state of Andhra Pradesh\(^7\):

1) At current levels and patterns of educational spending, the gain to additional spending on child-specific educational inputs (such as notebooks, exercise books, writing/coloring materials, etc) is likely to be greater than additional spending on teacher salaries.\(^8\)

2) Additional spending on teacher salaries can probably be more efficiently spent by using these funds for incentive payments conditional on learning improvement than by unconditionally hiring more teachers.

---

\(^5\) 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.

\(^6\) While many villages have private schools these days, government-run schools still account for the majority of enrollment, especially among the socially and economically disadvantaged population.

\(^7\) Andhra Pradesh is the fifth most populous Indian state and has a population of over 75 million people.

\(^8\) Increases in spending on teacher salaries can come about either by hiring more teachers or by increasing the salary of existing teachers. This project will focus on the former and place an extra teacher in the schools that are randomly selected for the “teacher” input (this also allows us to cleanly estimate a class-size effect).
3) Group incentives might be more effective than individual incentives in small schools with a few teachers, but individual incentives might yield better outcomes in larger schools.

While there have been a few randomized evaluation studies on education inputs in India\textsuperscript{9} there have been none so far in a representative sample of government-run schools. Besides, there has been no rigorous randomized evaluation on the effect of class-size or of providing child-specific learning materials on learning outcomes in India. As a result, even the “input” component of this project will answer questions that have not been addressed before in a rigorous way.

However, what is more unique is that this is the first attempt to systematically introduce and study a teacher incentive program in government-schools in India and is one of very few of its kind in all developing countries\textsuperscript{10}. It is also unique in studying both individual (teacher-level) and group (school-level) incentives in the same research study\textsuperscript{11}. Indeed, while issues of teacher incentives are at the forefront of education policy discussions in many countries, including the US, there are almost no rigorous randomized evaluations of the effectiveness of these incentive programs. Since the majority of teacher incentive programs in the US have been non-randomly introduced, these policy innovations lack statistically reliable comparison groups against which to evaluate the

---

\textsuperscript{9} Notable among these are Banerjee, Cole, Duflo, and Linden (2004), which studies the effectiveness of a remedial education program and a computer-assisted learning program in NGO-run schools in India, and Duflo and Hanna (2005), which studies the effectiveness of monitoring teacher attendance in NGO-run schools in India using cameras with time and date stamps.

\textsuperscript{10} Glewe, Ilias, and Kremer (2003) is the only other paper that studies the effectiveness of teacher incentive programs in a developing country (Kenya) using a randomized evaluation. Lavy (2002 and 2004) are the other main papers that study teacher incentives based on test performance outside the US (in Israel) but these papers do not use a randomized evaluation and are instead evaluated based on ex-post identification of a suitable comparison group of schools.

\textsuperscript{11} The theoretical prediction of the relative effectiveness of individual and group incentives is ambiguous, since the former would dominate if free-riding was a major problem, while the latter would be superior if there are significant gains to teacher cooperation and if peer pressure induces higher effort. While Lavy’s work cited above suggests that individual incentives are more effective, it is based on two different programs at different points in time. It is also the only work to date that addresses this very important issue in a rigorous way.
effectiveness of teacher incentive programs\textsuperscript{12}. Thus the component of this project that studies the impact of teacher incentives is relevant not only for developing countries, but also for the debate on teacher incentives in education policy in countries like the US.

II. Concerns with Teacher Incentive Programs based on Test Scores

While the idea of providing financial incentives to align the interests of teachers to those of parents, students, and educational administrators is fine in theory, there are also some important concerns about incentives based on test scores, which we need to address both in the context of this research project and from the perspective of eventual scalability.

First and foremost is the concern that such incentives will induce “teaching to the test” as opposed to focusing on multi-dimensional student development\textsuperscript{13}. While this is a valid concern, there are two responses in this context. First, the grades where we will implement the incentive program are from grades 1-5, where the levels of learning are so low to begin with\textsuperscript{14} that even teaching to the test would be a first order improvement. Second, and more important, is the fact that test design can get progressively more sophisticated so that it becomes difficult to do well on the test without deeper conceptual understanding of the materials covered. One of the partners in this project is Education Initiatives Ltd. (EI), an educational testing company that focuses on designing test questions that evaluate underlying competencies and skills in addition to mechanical

\textsuperscript{12} For instance the US Department of Education reports that a teacher incentive program introduced in Chattanooga, Tennessee (known as the Benwood Initiative) led to the affected schools improving their reading, language and math scores at twice the rate of the rest of the district (as reported in a research report prepared by the House Research Organization of the Texas House of Representatives in 2004). However, the problem is that the program was introduced in 9 of the lowest-performing schools in the state, which means that the results could be in part due to “mean reversion” and in part due to the fact that “rates” of improvement are typically higher when the base levels of achievement are low. This makes it difficult to ascribe causality to the program in question.

\textsuperscript{13} Holmstrom and Milgrom (1991) and Baker (1992, 2002) are the main references for the theory of “multi-task” moral hazard

\textsuperscript{14} Our baseline tests indicate that the typical child in grades 4 and 5 in these schools has not even mastered competencies at the second grade level.
learning. Specifically, the research design takes psychometrics seriously\textsuperscript{15} and one of our innovations is that the test papers will include "pairs" of questions where one question in the pair will test mechanical knowledge while the other question will test the same idea in a more conceptual way\textsuperscript{16}. Since these tests will be administered to all schools in the study, we will be able to identify the extent to which improvements in test scores are due to only “mechanical” learning as opposed to both “mechanical” and “conceptual” learning.

A second concern is with outright cheating on the exams by teachers and students (Jacob and Levitt (2002)). Reliability of the test scores is obviously critical here and we will ensure this by having a completely independent testing team that is employed by a prominent NGO with experience in conducting child learning assessments (the Azim Premji Foundation) and is not directly connected to any of the stakeholders in the project.

The third and final set of concerns relates to the fact that incentives for performance can often induce perverse behavioral responses from teachers such as encouraging weaker children to drop out of school altogether (this can easily happen if for instance the incentive is tied to the fraction of students who meet a pre-determined learning threshold). We have therefore designed the program so that the incentives are

\textsuperscript{15} Some of the features of the test include: Creation of questions that test various competencies (mapped by grade) and being able to have reliable estimates of class-level scores for various competencies (as opposed to just having a total score for math or language), creation of item-characteristic curves by question, and use of item-response theory to ensure greater discrimination of the test as a whole.

\textsuperscript{16} Of course if a “conceptual” question is repeatedly taught in class it becomes a “mechanical” one. For the purpose of this study, a “mechanical” question will be considered to be one that conforms to the format of the standard exercises in the text book, whereas a “conceptual” one will be one that takes the same idea but tests it in an unfamiliar way. A good example is provided by the following case described by the researchers at EI: Teachers and students in a grade six class were asked if the students understood the concept of "area", to which the answer was an unequivocal "Yes". It turned out however that the nature of the understanding was quite limited. EI tested this by asking a series of multiple choice questions (all of which require the students to calculate the area of the shape provided). The first shape is a rectangle, the second shape is a triangle, and the third shape is that of a "house" with the triangle from question 2 put on top of the rectangle from question 1. They find that less than 50% of the children who correctly answer the first two questions are able to answer the third one with the common student reaction being "We have not been taught the formula for this" even though all they have to do is to add the 2 numbers that they have just computed! Clearly, the understanding of the idea of "area" was limited to the regurgitation of a formula for each shape as opposed to an understanding of the concept. This would be an example of a “question pair” that we plan to use, and the questions within such a “pair” will be spread out across the test paper.
tied to average *improvements* in child learning outcomes at the class/school level as opposed to reaching absolute standards (i.e. a simple form of the “value-added” measure\(^\text{17}\)). To make sure that teachers minimize drop outs, we will assign a very low score in the final test for any child who took the baseline but does not take the end of the year test. There is therefore an incentive to ensure that children do not drop out.

**III. Methodology**

Typical characteristics of government-initiated policy innovations that make evaluations difficult include: (i) **universal coverage** of interventions, (ii) interventions through non-random **targeting** of certain segments of the population based on measurable criteria, and (iii) targeting based on **self-selection** into programs. While the details of the difficulties in each of these cases vary, the essential problem is the same, which is the lack of a statistically-valid comparison group that is identical to the “treatment” group in all respects except the receipt of the treatment.

For these reasons, matching schools on observable characteristics and then randomly assigning some of them to a program or combination of programs provides the most rigorous way of learning about the effectiveness of the program(s). Since the program is randomly assigned to a subset of the potential recipients, the remaining potential recipients provide a perfect control group as there is no selection bias in who received the treatment and who did not. Also, other extraneous factors should affect both groups similarly on average and their effect can be netted out by calculating a “difference

\(^{17}\) There are more sophisticated forms of value-added calculation such as those devised by William Sanders (currently at SAS Corporation), and the one used in Israel. The essence of these systems is that a predicted score is generated for each student based on previous scores (baseline in this case) and family and other characteristics. The teacher’s “value added” is then measured relative to this expected score in order to better isolate the specific teacher’s contribution to learning. While such a measure might be better for longer-term policy, the major drawback is that it is not transparent in the beginning – especially in a context where teacher incentives have not been seriously discussed at all. By focusing on average “improvement” of all the students, we hope to use an intuitive measure of teacher performance that avoids the many pitfalls of focusing only on “levels” of learning (which for instance would penalize teachers assigned to weaker schools).
in differences” estimate\textsuperscript{18}. The methodology of randomized evaluation is considered the “gold standard” in evaluating the causal impact of programs not only by economists and clinical scientists, but also in the “No Child Left Behind” Act which calls for education policy to rely on a foundation of scientifically-based research, with randomized evaluations being ranked at the top of various research methodologies (Angrist – 2003).

While randomizing the allocation of programs ensures internal validity of the study results, one limitation of this methodology is that external validity (i.e. validity of the results beyond the set of schools being studied) can be more contentious. We specifically aim to improve the external validity of this study in a couple of ways. First, the state of Andhra Pradesh where the study is being conducted is close to the all-India average in both teacher absence and teaching activity and in that sense is one of the most “representative” states of India from the perspective of efficiency of education service delivery. Second, the set of schools included in the study is itself randomly chosen across five districts representing the distinct geographic and socio-cultural regions of Andhra Pradesh, and is therefore a representative sample of schools in a state with a population of 75 million people.

IV. Details of Inputs and Incentives to be studied

Inputs: There are two main kinds of inputs whose effectiveness will be evaluated via randomly providing these to certain “treatment” schools. These include:

a) Provision of an extra teacher to schools

b) Provision of a block grant of funds (equal to the cost of an extra teacher) to school committees that will have the freedom to use the money to procure non-teacher inputs for the school within certain guidelines. The guidelines give preference to

\textsuperscript{18} See Duflo and Kremer (2003) for a detailed exposition on the methodology of randomized evaluations in development policy.
child-specific learning inputs (such as note books, exercise books, and writing materials) and explicitly prohibit the use of money for construction projects.\textsuperscript{19}

\textbf{Incentives:} Similarly there are two kinds of incentives that are being considered for evaluation. These include:

\begin{itemize}
  \item[a)] Provision of monetary incentives on a \textbf{group} basis -- whereby every teacher in the school receives bonus payments over and above the existing salary based on the extent of improvement of the average score of all children in the school, and
  \item[b)] Provision of monetary incentives on an \textbf{individual} basis -- whereby every teacher in the school receives a certain amount of bonus payment based on the extent of improvement of the average score of the children in the class/classes taught by that particular teacher.
\end{itemize}

Since the incentive scheme is the same for all teachers in a school regardless of their base pay, we will also be able to identify and estimate the effect of varying the amount of the incentive payment as a fraction of the base pay of the teachers. This is because while the same amount of average improvement will generate the same monetary bonus payment for all teachers, it will generate different “percentages” of bonus payments relative to base pay across teachers.

An overview of the research design is provided in the figure below. Each cell represents a unique “treatment” and we can estimate their effectiveness relative to the control group and also relative to each other\textsuperscript{20}.

\textsuperscript{19} While the impact of “construction” inputs such as buildings and classrooms would be interesting to study, it is not feasible here for 2 reasons. First, the process would be much more difficult to standardize and second, the “true return” of a construction project needs to be assessed over the depreciation-cycle of the investment, which would be around 15 to 20 years!

\textsuperscript{20} The hierarchical structure of the data (students within classrooms within schools) means that even though the unit of observation is the individual student, these observations are not statistically independent of each other. The effective number of observations therefore depends on the intra-cluster correlation ($\rho$) of these scores, and one important objective for the pilot phase we carried out last year (2004-05) in 40 schools was to calculate $\rho$ and to choose our sample size (number of schools in each cell and number of interventions) accordingly. Conservative calculations of statistical power based on the estimates of $\rho$ from the pilot phase suggest that the project design has a 94\% probability of detecting a “treatment effect” of 0.2 standard deviations (of the baseline test score distribution) at a 5\% confidence level and a 99\% probability of
We randomly chose 10 sub-districts in each of the 5 chosen districts and then randomly chose 10 schools in each sub-district to create the study universe of 500 schools across 5 districts of Andhra Pradesh. The average school has around 100 children, and the study sample is around 50,000 children. After conducting baseline tests of all children in all 500 schools, we randomly allocated 2 of the 10 schools in each sub-district to each of the 5 treatment “cells” above (i.e. the treatments were stratified by geography\textsuperscript{21}). Since the baseline testing and randomization have already been completed (please see the “Detailed Work Plan” in the attached Appendix for a full list of the activities conducted so far over the past 18 months and expected to be conducted over the next 18 months), we can also report at this stage that the treatment and control schools are statistically indistinguishable on various observable characteristics such as baseline scores in math and language, school enrollment, pupil-teacher ratio, etc. This “statistical balance” of the treatment and control schools is an essential component of randomized evaluations. We have also conducted household surveys to obtain household demographic data for all the 50,000 children in the study universe. While this is not essential to estimate the “average treatment effects”, it allows us to study the differential detection at a 10% confidence level. See Chapter 5 of Donner & Klar (2000) for details regarding the equations used in these power calculations.

\textsuperscript{21}This is important because the lowest level of education administration is at the sub-district level, and this stratification is essential to ensure that results are not invalidated by unique “sub-district level” features such as the characteristics of the sub-district level education officer.
impact of the treatments being studied by household characteristics, and we expect that having the detailed household data will greatly enrich the analysis.

The main outcome that we will track is the performance of the children on the independently administered tests (that measure both mechanical and conceptual learning and also measure learning by various competencies). However, we are also conducting tracking surveys to collect data on teacher and student attendance as well as classroom processes in both treatment and control schools. This will allow us to track differences in process variables as well as outcome variables and we hope to thereby not only measure differences in outcomes but to obtain some insight into the mechanisms by which the differences arise. Finally, we will also be interviewing a random sample of parents of the children in both treatment and control schools to understand if the various treatments generate any effects that are visible to the parents.

What makes this project particularly unique in a developing country context is the fact that we have worked closely with a government and designed this research study for implementation in government-run schools, which greatly increases the likelihood that the findings from this study will strongly permeate policy thinking. We hope that the results from this study will help inform discussions on education policy all over India and beyond.
References


Appendix A

Project Team

- Government of Andhra Pradesh, Department of School Education
  - Secretary, Department of School Education
  - State Project Director, Sarva Shiksha Abhiyan
  - Director, School Education

- Technical Consultants
  - Venkatesh Sundararaman, Economist, South Asia Human Development Unit, World Bank
  - Karthik Muralidharan, Department of Economics, Harvard University

- Azim Premji Foundation (Project Implementing Agency):
  - D D Karopady – Head of Research, APF
  - M Srinivasa Rao – Project Leader
  - Project Team
    - 5 District Coordinators
    - 25 Sub-district Coordinators
    - Evaluators

- Education Initiatives (Test Development and Psychometrics):
  - Sridhar Rajagopalan – Managing Director, EI
  - Vyjyanthi Shankar – Head of Test Development
  - Test development and analysis team
Appendix B

Detailed Work Plan and Schedule

The overall project has been carried out in 2 phases with an optional 3rd phase (depending on the government of Andhra Pradesh) according to the following time line:

Phase 0 (Planning): January 2004 – August 2004
Phase 1 (Pilot): September 2004 – April 2005
Phase 2 (Main): April 2005 – June 2006
Phase 3 (Sustainability): June 2006 – May 2007

The pilot phase of the project has been completed, and we spent most of the summer of 2005 in India setting up the main phase of the project. The main activities so far have included:

**Phase 1: Detailed process pilot in 30-40 schools in one district (COMPLETED)**

1) Understanding and mastering the logistics of the intervention.
   a. This included all the logistics of
      i. How we let schools know that they have been selected for the program
      ii. How the tests are conducted
      iii. Obtaining student household information
      iv. Data entry and quality control
   b. This also includes the process of tester selection, training, and supervision, as well as figuring out processes for coordination and quality control across 5 districts during main intervention.
   c. The end product was a set of written guidelines for all the people involved in implementing the main phase of the project

2) Calculating the intra-cluster (intra-school) correlations in test scores, and test score improvements. This is a critical input in calculating the minimum sample sizes required for each treatment cell in the main intervention, and was one of the main analytical reasons for the pilot.
   a. Since the intra-school correlation in test-score improvement (our main variable of interest) turned out to be higher than we had expected, we ended up reducing the number of interventions that we evaluated and increased the number of schools in each “treatment cell” in order to ensure adequate statistical power for the experiment.

3) Test refinement and calibration.
   a. The main objective here was to refine the tests and ensure that they were designed so as to allow us to capture fine distinctions in student achievement to the extent possible.
Phase 2: Main Evaluation of Inputs and Incentives in Rural Government Schools

April 2005 – June 2005 (COMPLETED)
- Prepared detailed process map for main intervention based on learning from Phase 1
- Decided the final set of interventions that will be conducted and determined sample size for each cell
- Hired the project manager (1), district coordinators (5), sub-district coordinators (25) and the testers (250) and conducted various training and quality control sessions for the different team-members (total team size of nearly 280 people).
- Conducted random sampling of the 500 schools that would comprise the universe of the study
- Worked with senior officials in the state education ministry to communicate the details of the main project 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).
- Iterated and confirmed the testing instruments
- Prepared all logistics for the baseline tests (printing test papers, organizing data entry, etc.)

July 2005 (COMPLETED)
- Conducted baseline tests in all 500 target schools in a 2-week span in early July. 50 teams of 5-testers each were employed, with each team conducting the testing in one school per day. Each tester was responsible for the testing in one grade.
- Data entry of baseline tests
- Collection of detailed household demographic data
- Evaluation of baseline tests, and production of diagnostic reports at the child, class, school, sub-district, and district levels
- Conducted the randomization exercise to allocate schools to the control and treatment groups.

Late July – early August 2005 (COMPLETED)
- Combined meeting with relevant district and sub-district level officials to communicate the details of the various interventions
- Distribution of test results, diagnostics, and announcement of relevant incentive schemes in selected schools
- Placement of extra teacher in the relevant randomly selected schools
- Provision of block grants to the relevant randomly selected schools

Current Activities:

We are currently conducting tracking surveys of the 500 schools in the study (visiting each school approximately once a month) to collect data on teacher and student
attendance, and classroom processes from both the treatment and control schools. These tracking surveys started in September 2005 and will go on till February 2006.

**Timeline of forthcoming activities:**

**January - April 2006:**
- Data entry, cleaning, and preliminary analysis of the household survey data, baseline test data, and tracking survey data
- Repeat of many of the processes involved in conducting the baseline tests
- Conducting of end of year testing in the 500 schools in the study

**April - May 2006:**
- Endline test data entry, cleaning, and preliminary analysis

**June 2006 onwards:**
- Awarding of incentives to teachers based on performance and announcement of continuation of program
- Intense work on analysis and writing of paper draft
- First draft of the paper should be ready by end of June 2006
- Feedback, comments, and iterations of drafts between July and August
- Target a finished draft by end of September 30th