How to Teach English in India: Testing the Relative Productivity of Instruction Methods within the Pratham English Language Education Program[1]

First Draft: October 31, 2006

Current Draft: July 1, 2008

Fang He

Leigh L. Linden

Margaret MacLeod

Abstract: Using a pair of randomized evaluations, we assess the relative productivity of several modes of implementing an Indian English education curriculum. Each consists of a specially designed machine or flash card based activities implemented either indirectly through a teacher training program or directly by externally supervised teaching assistants. The new methods are very effective and, on average, all implementation strategies yield gains of about 0.25-0.35 standard deviations in students' knowledge of English. Weaker students benefit more from interventions that include teacher directed activities while stronger students benefit more from the more self-paced machine-based implementation. Compared to an externally implemented version of the curriculum, the treatments implemented through the teacher training program improved students' math and English scores rather than just their English scores, a result that may be due to the fact that teachers implemented the interventions more efficiently.

[1] Like all project of this magnitude, this study would not have been successful with out the hard work and dedication of a large number of people. We are indebted to Madhav Chavan for suggesting this study, for continued collaboration in this line of work, and for his inspiration and support on this and many other projects. The study and intervention were funded by the Michael and Susan Dell Foundation and we thank Barun Mohanty for his understanding of the time it took to clean and process all of the data necessary to produce this final report. For the execution of the project, we are indebted to Madhuri Pai for running the intervention and overcoming what always seemed like insurmountable odds to make the project happen. None of the data would have been collected without the constant on-ground supervision of Pratima Bandekar. Nandit Bhatt and Rajashree Kabare provided invaluable assistance in the coordination of the research effort. The views expressed in this article are those of the authors, and do not necessarily reflect those of the U.S. Department of State or the U.S. Government. He: Department of Economics, Columbia University. Linden (corresponding author): Department of Economics and School of International and Public Affairs, Columbia University, 420 West 118th Street, Mail Code 3323, New York, NY 10027, leigh.linden@columbia.edu. MacLeod: School of International and Public Affairs, Columbia University and Department of State, United States Government.

I. Introduction

In many developing countries, differences in income are correlated with differences in language. Due to the legacies of colonial authorities, the official languages of government and business is often French or English rather than the native language of the population. Speaking these languages can confer significant economic benefits. In India, for example, speaking English is a particularly valuable skill – providing speakers with a premium of about 25 percent over non-English speaking workers with similar characteristics (Munshi and Rosenzweig, 2006). Angrist and Lavy (1997) analyze a curricular change in which Moroccan school began teaching Arabic rather than French. They conclude that the change caused a significant reduction in earnings primarily due to weaker French writing skills.

Unfortunately, while the official curriculum of many countries requires students to learn these languages, the quality of instruction is often particularly poor and students learn little. In the sample of our current study in India, only about 10 percent of second and third grade students can correctly identify the pictures of objects listed on the official curriculum when given the object's English name. This raises two fundamental questions about the nature of education production. First, can a better teaching method improve students' acquisition of a crucial second language? And, if so, what is the most effective technology to deliver the curriculum to children?

We conduct two randomized treatment-control trials to test a novel strategy for teaching English to children in grades 1-5. The program includes both a machine-based implementation strategy and a series of activities based on a specially designed set of flashcards and teacher manuals. The research design varies both the implementation technology and whether the intervention is delivered through externally hired tutors or the local public schools' own teachers and teaching assistants.

Methodologically, our study builds on a growing literature on the efficacy of educational interventions in developing countries. These evaluations include, among many others, assessments of classroom resources like textbooks (Glewwe, Kremer, and Moulin, 2003) and flipcharts (Glewwe et al., 2004), programs designed to reduce class sizes and restructure classrooms (Duflo, DuPas, and Kremer, 2007), and programs designed to change pedagogy used by teachers (Banerjee, Cole, Duflo, and Linden, 2007; Linden, 2008). While these studies provide evaluations of a wide variety of interventions, a major drawback to these evaluations is that they are implemented in very different contexts and are often used to deliver very different curricula. Our study allows us to assess not just the efficacy of a new method for teaching English, but to do so using different implementation technologies – allowing a direct comparison of their productivity by holding the curriculum and educational context constant.

In comparing these implementation strategies, we also build on two more specific but separate strands of the economics of education literature: evaluating the effectiveness of both computer-assisted learning programs and teacher training programs. Computers have provided an obvious potential solution for providing consistent instruction to students; the effects of various computer assisted learning programs have been extensively studied in the developed world with mixed results (Boozer, Krueger, and Wolkon, 1992; Goolsbee and Guryan, 2002; Kirkpatrick and Cuban, 1998; Wenglinsky, 1998; Machin, McNally, and Silva, 2006). The most rigorous of these yeild mixed conclusions. Angrist and Lavy (2002), Rouse and Kreuger (2004), and Dynarski (2007) evaluate a general computerization program in Isreal, an English program in the United States, and a series of math programs in the United States respecitvely all of which have limited effects on students' test scores. Machin, McNally, and Silva (2006) evaluate a general British program to put computers in schools and Barrow, Markman, and Rouse (2007) evaluate a math focused computer assisted learning program in several cities in the United States. Both find positive effects.

The main cause of these mixed results may be the relative quality of learning environment that computerized instruction replaces. In developing countries, the quality of instruction is significantly lower than in most developed countries. This raises the possibility that computers could provide a more productive relative learning experience than in the developed world. The only evaluations of programs in developing countries have shown that computers can make a significant difference in students' knowledge of subjects (Banerjee, Cole, Duflo, and Linden, 2007; Linden, 2008). However, in developing countries, computers are sufficiently expensive that despite these large gains, the programs are not always cost effective.

Building upon the literature evaluating teacher training programs, we also evaluate a version of the program that is implemented by teachers through a teacher training program. Teacher training programs seem to have attracted less attention than computer-assisted learning programs in developing countries, but mixed evidence exists for more developed countries. Kennedy (1998) conducts a meta-analysis of several studies, concluding that, while generally effective, the content of the training programs seems to matter more for changes in student performance than the structure of the program itself. Jacob and Lefgren (2004) and Angrist and Lavy (2001) conduct large scale evaluations of teacher training programs in Chicago and Israel respectively using quasi-experimental designs, but reach different conclusions. In Chicago, the programs seem ineffective, while in Israel, the programs implemented in secular schools improve test scores by 0.2 to 0.4 standard deviations while those in religious schools (which were poorly organized) seemed to have no effect.

In what follows, we test elements of a unique English language training program developed by Pratham, an Indian network of NGOs, that is designed to change the way that English is taught in Indian classrooms. The curriculum can be delivered through a series of educational activities and games that utilize a set of specially tailored flash cards or through specially designed software and booklets used with a uniquely designed electronic machine (called a PicTalk machine). The research design allows us to evaluate elements of the program in both a rural and urban location and allows us to vary the method of implementation to include direct administration of the program by an external agency and implementation of each element of the program by the localities' own teachers and teaching assistants. This allows us to assess both the effects of the new curriculum and also different methods for delivering the curriculum to students.

In general, the new methodology is very effective – increasing students' English test scores by an average of 0.25-0.35 standard deviations. In addition, each implementation method seems to be, on average, equally effective at teaching English regardless of the technology used by the instructor or whether the instructor is an external provider or the students' own teacher after having undergone Pratham training. The teacher implemented interventions, however, have the added benefit of also causing changes in students' math scores, which is a subject not covered by the intervention. Teachers used the interventions less frequently compared to the external team's implementation, suggesting that teachers may have efficiently substituted to other subjects once the students had grasped the English concepts.

The evaluation of the teacher implemented intervention was designed to allow for a direct comparison of the different implementation technologies. While all interventions are equally effective on average, we find that students benefit differentially from individual technologies. Lower performing students benefit more from interventions that include teacher implemented activities while higher performing students benefit more from the relatively self-paced machine only intervention. The gains among the lowest performing students under the activities intervention is a striking change from the normal trend in Indian classes that typically focus on the better performing students (Banerjee, Cole, Duflo, and Linden, 2007). These results also suggest that significant gains (up to 1/3 of the average treatment effect of 0.36 standard devaitions) could be made by targeting treatment towards individual types of children.

The remainder of the paper is organized as follows. Section 2 contains a brief introduction to the literature on second language acquisition and a description of the program. In Section 3, we characterize the design of the experiment, describe the data that was collected, and layout the plan for analyzing the data. Section 4 contains the results, and finally, we conclude in Section 5.

II. Second Language Acquisition and the Pratham PicTalk Program

A. Second Language Acquisition (2LA)

Developing strategies, like the PicTalk program, for improving language instruction in primary schools is particularly important because early instruction in a new language is critical for a child's eventual proficiency. Numerous empirical studies show a strong negative relationship between age of language acquisition and proficiency in the second language (see Kroll and de Groot (2005) for a review). For example, Kim (1996), McDonald (2000), and Shim (1993) all administered oral grammar tests to subjects and observed moderate to strong negative correlations between age of acquisition and proficiency. They also observe positive correlations between age of acquisition and reaction time to grammar questions. The same relationship exists for tests covering picture/sentence matching (Lee and Schallert, 1997), auditory comprehension (Oyama, 1978), and pronunciation (Oyama, 1976; Patkowski, 1980).

There is also neurological evidence that late learners of second languages may use entirely different brain processes than early learners. Ullman (2001) develops a model showing different degrees to which first language and second language acquisition are dependent on declarative memory (used in the learning and use of fact and event knowledge) which is based in temporal lobe structures in the brain and procedural, or implicit, memory (used in the acquisition and expression of motor and cognitive skills) which is based in the left frontal/basal-ganglia structures of the brain. He argues that increasing age leads to a larger shift of grammar in 2LA to declarative memory. In other words, as individuals age, more of 2LA is allocated to the memory system that is associated with fact and event acquisition than the memory system associated with acquiring skills. Perani and Abutalebi (2005) dispute the relative use of the neurological mechanisms, but nonetheless conclude that older bilinguals seem to experience greater processing requirements than younger ones when performing similar tasks.

Johnson and Newport (1989) further argue that there is actually a “critical period” in which an individual is able to acquire a second language up to native-like levels. The exact definition of this critical period has not been established, but it is generally claimed to be before the onset of puberty. While there is some disagreement on this (see for example, Hakuta, Bialystok, and Wiley (2003); Birdsong (1999)), the strength of Johnson and Newport’s argument emphasizes the general consensus on the importance of early introduction in the learning of a new language.

B. Program Components

A critical problem then arises when the Indian public primary school system is ineffective at teaching English, a second language that is not only crucial to a student's employment opportunities, but that is also a skill most easily learned at an early age. The PicTalk program is designed to help address this problem.[2] The PicTalk program consists of two distinct components. The first component is the actual PicTalk Machine itself. Based on Interactive Paper Technology (IPT) created by LeapFrog Enterprises, the PicTalk Machine resembles the LeapPad® and Leapster® interactive learning tools marketed in the US.[3] The machine is designed to be used individually by a child. Inserting a glossy spiral-ring book with a corresponding memory cartridge into the PicTalk machine, children can point to pictures with a stylus and hear the word pronounced aloud. Additionally there are touch points to receive instructions in the local language, Marathi,

[2] The PicTalk program was designed by Pratham, an India-wide network of regional NGOs, Trusts, and Foundations, has taken on the mission to get all Indian children "in school and learning well."

[3] There exists a small literature into the effectiveness of PicTalk-like technologies funded by LeapFrog. However, while these studies generally conclude that the technology is effective, almost all of them have severe methodological problems including small sample sizes and high attrition rates (for example, RMC Research, 2003a, 2003b; LeapFrog, 2004a and 2004b).

as well as quiz questions that ask the children to identify words by pointing and then give auditory feedback. Designed to weather the extra abuse of handling by children, the PicTalk machine resembles a small notebook with a hard plastic shell. Attached is a small plastic stylus on a cord that registers the location on the page to which a child is pointing. Pratham Mumbai developed five or more IPT "books" each for grades one through four, including picture dictionaries, stories, poems, and actions. For the first and second grade students for whom oral proficiency is the government-mandated objective, pronunciation drill books were also developed.

The second component is a set of interactive activities designed around sets of 440 flashcards, designed to teach the same competencies covered by the PicTalk machine booklets. To implement these activities in a consistent way, Pratham gives instructors manuals that include recommended drills for use with students. These activities are intended to promote oral communication through conversation and recitation. They include chants created independently by Pratham, and poems that are already a part of the school curriculum and separately available on government-issued audio tapes.

The general strategy of the intervention is to introducing syntax, pronunciation, and vocabulary simultaneously, a sharp contrast to existing pedagogical methods used in the government schools. Generally, in the government schools, teachers train children to recite the alphabet, then memorize vocabulary lists out of context. Only after the alphabet and words have been taught by rote are conversations introduced. Among the prescribed text books are some that transcribe English into Marathi script, thereby not requiring the teacher to read. In classes beyond the second grade, some teachers skip reading the source text in class, choosing to dictate and teach only the questions and answers in preparation for upcoming exams.

III. Research Design

A. Structure of Research Groups

To evaluate the efficacy of the PicTalk program, we undertook two randomized treatment-control experiments, the first one over the 2005-2006 academic year and the second one over the 2006-2007 academic year. Between the two experiments, we were able to vary the method of implementation of the program in order to better understand the functioning of the individual program components, including the method of implementation itself. First we evaluated both components in an integrated program administered by Pratham hired and managed assistants.[4] This allowed us to evaluate the efficacy of a version of the program that came as close as possible to the one originally

[4] The intervention specifically provided children with the PicTalk and activities on alternating days of the week.

designed by Pratham. In the second year, we divided the sample to decompose the effects of the entire program, and implemented three interventions: machines only, activities only, and machines with activities. These later schools also did not receive external implementation but rather the existing teaching assistants and teachers were trained to implement the program using the same materials as used in the first year of the study. As in all teacher training programs, Pratham could provide assistance to teachers, but had no ability to mandate that teachers implemented the curriculum as designed.

Table 1 describes the sample of students in both years of the study. In the first study year, the sample of students included second and third standard children enrolled in the Thane Municipal School District. In the Indian school system, this means children generally aged about 6-9 years. Within the Thane Municipal School District, 97 schools were included in the study. Schools were assigned to one of two research groups: 1) PicTalk class in second standard but not in third, and 2) Pictalk class in third standard but not in second. In effect, every school served as both a treatment and a control. The random assignment determines only which of the standards serves as which.[5] Because of the relatively small number of schools, we used a stratified randomization design in which schools were randomly assigned with equal probability between the two groups

[5] This is similar to the research designed used in Banerjee, Cole, Duflo, and Linden (2007)

based on their score on the baseline test and their location in a more rural area. Within these schools, we then randomly chose one division from each school-standard combination to track for the follow-up test. This left the final sample of 97 schools and 194 divisions equally split between standards two and three.

In the second study year, the sample of students included children in standards one through five enrolled in the Mangaon sub-district government schools. Within Mangaon, 242 schools were included in the study. Schools were assigned to one of four research groups: 1) PicTalk machine classes only, 2) activities classes only, 3) both PicTalk machine and activities classes (as in the first year), and 4) neither PicTalk machine nor activities classes. Unlike the first year study, each school is categorized into one of the four research groups. We used a stratified randomization design in which schools were randomly assigned with equal probability between the two groups based on their size. This research design resulted in 1,002 classes evenly distributed across the four research groups. Student-wise, there were 9,944 students evenly distributed across the four research groups.

Based on their school's assignment to the various research groups, teachers attended a five day Pratham training session teaching them how to implement the respective components of the PicTalk program.[6] Support of the actual PicTalk machine (replacement of defective units, technical assistance, etc.) was provided by the same private company that was used for the first year study – the firm simply supported the teachers rather than the Pratham staff members. Finally, all treatment schools also had regular access to Pratham monitors who circulated amongst the schools on a regular basis to assist teachers as questions arose.

B. Description of the Data

During both study years, the data was collected by a local group of assistants who operated independently of the group implementing the program in order to ensure objectivity. For both years, four types of data were ultimately collected. Baseline test scores, follow-up test scores, and attendance data were collected through direct surveying of the students by these assistants in both years. Demographic information was collected through the schools' administrative records in both years, and additional demographic information was collected during the second year through direct surveying.

Measuring baseline knowledge in a subject in which students have little skill

[6] Twelve schools did not attend the training, but both received materials from Pratham and received second hand instruction from a trained teacher in the same *kendra,* an administrative cluster of five to ten schools supervised by a single individual, called a *kendra pramok.*

presents a dilemma for the evaluator. Generally, a baseline test to measure students' knowledge of English prior to the start of the program would have a two-fold purpose. First, we need a baseline understanding of students’ English knowledge in order to understand how that knowledge changes over time. Second, and more importantly, the test must differentiate, as much as possible, students’ English skills so that we could control for these differences when comparing students’ performances on the follow-up exam. For this purpose, the most important function of the exam is to distinguish between students’ relative skill levels. An exam that demonstrated that students were equally ignorant of the surveyed material would be of little use.

During the first study year, piloting of different test instruments unfortunately suggested that students' knowledge of English was very poor and that students felt uncomfortable being quizzed on English one-on-one. As a result, we designed the test to include only very basic questions and instructed surveyors to probe as much as possible to identify whether or not students could provide a correct answer to a given question. Surveyors were instructed to provide children with every opportunity to answer the questions correctly, even prompting the child with additional information when necessary. As a result, the baseline test probably overestimates student's understanding of English. But in doing so, we generated additional variance in pre-test scores that within a regression model will provide more precise estimates of the difference in student's abilities during the follow-up exam.

During the first study year, the baseline test contained a single section with five subsections. The first subsection was designed to test children’s recognition of the alphabet and to put children at ease by starting with the simple task of identifying an individual word from a list in which all of the words were those whose pronunciations are the same in English and Marathi (“Familiar Words”). The subsections then required the children to follow instructions given in English (e.g. “Sit down,” “Touch your head,” etc.), identify pictures of objects, written verbs, and written nouns from a list of possible answers. Finally, because children would have a very difficult time completing the tasks in English if they did not understand the Marathi question, the surveyors asked the same questions in Marathi after asking them in English.

In the second year of the study, piloting indicated that students’ knowledge of English was not as limited as in the first year. This allowed us to use the same testing procedures and test format at both baseline and follow-up, allowing us to directly compare students between baseline and follow-up. Unlike the first year of the study, this allows us to not only assess the effect of the interventions on students’ understanding of the material in the post test but also to measure how that understanding changes over time.

The follow-up test contained three sections. The first section was the English section from the baseline assessment. The second section tested more difficult competencies in English. The first four subsections of the second section required students to identify words from a list of similarly structured words. The list of words was chosen to focus on differences in various phonemes within the words. Finally, students were asked to identify more complex words out of a list and then to identify a given sentence from a list of possible answers. The last section was a simple math exam designed to test the children's ability to identify numbers, add, subtract, and multiply. Because the purpose of the follow-up test was to understand the changes in students' learning levels generated by the treatments, students were never prompted during the testing process and the follow-up test was consistently administered in both years.

In both study years, we also collected two other types of information. First, schools in India maintain basic demographic information on their children. This information almost always includes the child's age, religion, caste, address, and parents' names. For most of the children, the school also collects each parent's occupation and education level. During the second study year, we also conducted direct surveys of the students and teachers to collect additional demographic information. After piloting different versions of the demographic surveys to ascertain what types of information a primary school child may be able to report accurately, we ended up collecting information on household infrastructure (electricity and water sources), household assets (appliances and livestock), family characteristics (number of siblings, number of adults in household, presence of parents).

Second, because of the popularity of using computers and learning English, we also collected information on student's attendance at school throughout the two study years. Surveyors were sent to each class in the study once a week on a randomly chosen day and called roll in the classroom. Finally, because we could not control the actual implementation of the treatment by the teachers in the second year, we also implemented a short survey of five randomly chosen children per class during each attendance monitoring visit on their experiences with English instruction during the previous day.[7] In this part of the survey, students were asked very basic questions including whether or not they (a.) studied English on that day, (b.) learned English from the chalkboard, (c.) studied English in their textbooks, (d.) participated in English games or activities, or (e.) learned English through a PicTalk machine.

[7] The surveying of the children was a compromise. We had originally planned to survey the teachers, but chose the children due to concern that the teachers would react negatively to having their activities directly monitored by the study.

C. Analytic Models

To evaluate the educational outcomes of children who had access to the PicTalk program relative to those who did not, we use three types of difference estimators that compare the pre and post test scores[8] across the relevant treatment and control groups. First we use a simple difference estimator of the following form:

$$Y_{ijk} = \beta_0 + \beta_1 Treat_{jk} + \varepsilon_{ijk} \tag{1}$$

where Y_{ijk} is a student characteristic (such as gender, normalized pre-test scores, etc.), and $Treat_{jk}$ is an indicator variable for whether or not the students standard-school combination was assigned to receive the treatment. The subscript *i* denotes the individual student, *j* denotes student’s standard, and *k* denotes the student's school. This estimator is primarily used for comparing the characteristics of the various research groups at baseline.

For estimating the possible effects of attrition, we compare the relative patterns of attrition in the treatment and control groups based on the information collected at baseline. The estimator takes the following form:

$$Y_{ijk} = \beta_0 + \beta_1 Treat_{jk} + \beta_2 Attrit_{ijk} + \beta_3 Treat_{jk} * Attrit_{ijk} + \varepsilon_{ijk} \tag{2}$$

where $Attrit_{ijk}$ is an indicator set to one if the student did not give a post-test. In this

[8] All test scores are normalized by grade relative to the distribution of scores in the control group.

specification, the coefficient β_3 is a difference in differences estimate of the relative differences between attritors and non attritors in the control and treatment groups.

To assess the results of the various programs, we use a similar difference estimator to equation 1 but also control for demographic characteristics. This specification takes the following form:

$$Y_{ijk} = \beta_0 + \beta_1 Treat_{jk} + X_{ijk} + \varepsilon_{ijk} \tag{3}$$

The only difference between equations 1 and 3 is the vector X_{ijk} which controls for differences in student and classroom characteristics at baseline. For each student, we control separately for their normalized pre-test scores for each section of the exam (linear and quadratic terms) as well as the student's age, grade, gender, mother's educational level, father's educational level, mother's employment status, and minority religious or caste affiliation. For estimates that only include the second year, we also include the additional demographic variables collected from the students and teachers. At the classroom level, we control for the student teacher ratio and the average and standard deviation of pre-test English scores.[9]

[9] One of the challenges of using administrative data is that one has little control over how the data is collected. In the case of the students' socioeconomic characteristics, information on the parents is usually only available for about 60 to 80 percent of the students. Dropping observations when the data is missing would mean varying the sample when making estimates with and without controls. To avoid this, we zeroed out missing observations and included a dummy variable for whether or not the each measure was provided by the school of each student.

In all of the data we collect, students within the same classroom share a number of experiences (same teacher, same classmates, same class resources, etc.) that cause their scores to be correlated. Without taking this into account, this correlation will cause us to overestimate the precision of the differences represented by β_1. We correct for this by allowing the standard errors to be correlated at the school level.

Finally, we had to modify these specifications slightly to take into account the slight idiosyncrasies of the two data sets. For year one, we randomly selected subsets of the classes (one division for each grade in each school) for inclusion in the follow up test. In order to make our estimates representative of all of the students in the city, we re-weight our outcome estimates by their probability of selection.

In the second year, Mangaon presented an unexpected challenge. Unlike the first year where all students seemed to consistently know very little, students in Mangaon varied significantly in their knowledge of English. The distribution for Mangaon is comparatively flat. This heterogeneity became a particular problem at the tails of the distribution because the lack of density resulted in dissimilar schools being spread across four research groups. Essentially, these are schools that are unique enough that there are too few schools with similar average scores to use as a comparison. As a result, the machine only group was assigned a disproportionate number of extremely high scoring schools. To remove this heterogeneity, we trim the overall sample, removing the top and bottom 10 percent of schools based on the average baseline English scores. The net effect of trimming is explained in detail in the appendix.

IV. Results

The results are divided into five sections. First, we evaluate the internal validity of the study by checking to make sure that the randomization generated comparable treatment and control groups and find that they are (see Table 2). To make sure that changes in the student population through attrition did not modify the relative composition of the various research groups, we conduct an analysis of attrition patterns in subsection B (Table 3). The similarity in attrition patterns then allows us to estimate the effects of the program by directly comparing the performance of students after receiving the intervention. Subsections C through F then assess the outcomes of the study, analyzing follow-up test scores (Subsection C, Table 4), utilization of the interventions (Subsection D, Table 5), effects on subsets of the sample (Subsection E, Tables 6-8), and attendance (Subsection F, Table 9).

A. Internal Validity

The purpose of the random assignment is to generate comparable research groups that differ only in that one group receives a particular treatment and others do not. While we cannot check the similarity of the research groups along every dimension, we can test for differences in characteristics observed within the two study years. We focus on three sets of characteristics: school characteristics, baseline test scores, and socioeconomic characteristics. Most observed differences are very small and it seems that the randomization succeeded in creating comparable research groups for both study years.

Table 2 shows the sets of differences between the research groups for both study years. For the first study year in Thane, the first column displays the mean for the control group. The second column then displays the average difference between the treatment and the control group. To help assess the magnitudes of these differences, the third column contains the results of regressing the post-test score on each of the characteristic using the schools in the control group. For the second study year in Mangaon, the first column displays the mean for the control group and the second column provides the difference between the average characteristics in all of the treatment groups and the control group. The third, fourth, and fifth columns then show the differences in average characteristics from the control group for the machine and activities, the machine only, and the activities only interventions respectively. As in year 1, the last column provides the marginal correlation between each of the control variables and the post-test score in Mangaon.

We present school level differences in panel A. We first compare the number of students within each school to gauge the population density. Secondly, we compare the average aggregate baseline test scores in order to measure the strength of the schools. Lastly, we compare the number of students per classroom as a proxy for the resources available to students. Along all three of these school level characteristics, the research groups in both study years are similar. The largest difference between the research groups is the 0.161 standard deviation difference in school average aggregate baseline test scores between the machine and the control group in year 2. Even so, that difference is not statistically significant and given the correlation between the school average score and the post-test, it is directly equivalent to only 0.046 standard deviations of the follow-up exam.

Table 2 panel B contains comparisons of the students by baseline test performance. In the first study year, the baseline assessment only consisted of English and Marathi sections. These score difference in the treatment and control groups are very small (-0.014 and -0.038 standard deviations). In the second year of the study, we used a baseline test that contained sections on English and mathematics. The characteristics for these scores are provided in rows 4 and 6. The differences here are larger than in the first year, but still small, especially when taking into account the coefficients in the regression in column 9.

Figures 1 and 2 compare the entire distribution of students' baseline English scores for the first and second year of the study respectively using a kernel density estimator. The distribution of the treatment and control students in Figure 1 are very similar reflecting the small differences in mean scores. The differences between the control groups and respective treatment groups are a bit larger in Figure 2, but again these differences are small given the correlation between baseline and follow-up tests.

The set of socioeconomic characteristics in Panel C also suggests that the different research groups are similar. These characteristics were chosen either because they are characteristics of common interest or because they are characteristics that may have obvious implications on student academic performance. However, because the data collection differed between the two study years, the set of available socioeconomic characteristics also differed. In the first study year, we have data on the gender, age, religion, caste, parents' education, and mother's occupation status. The minority religion variable groups Islam, Buddhism, and Christianity into a larger category and is relative to the majority Hindu religion. The lower caste variable groups scheduled tribes, scheduled castes, and other backward castes into a larger category. These particular groups may not only be economically disadvantaged but also academically disadvantaged compared to other groups. Nonetheless, the treatment and control groups in the first study year are nearly identical along these dimensions.

In the second study year, we have data on gender, age, religion, caste, mother's occupational status, household electricity availability, household running water availability, number of household assets (television, radio, and refrigerator), number of livestock (buffalo, oxen, cow, chicken, and goat), and whether either parent lives elsewhere (e.g. if a father is a migrant worker in Mumbai). Again, the minority religion and the lower caste variables are categorical variables that include the same groups as in the first study year. However, unlike the first study year, there are a few statistically significant differences among the research groups, but these differences are practically insignificant taking into account the regression in column 9. For example, many of the treatment groups are less likely to have household running water compared to the control group, but as shown in column 9, the relationship between running water and the follow-up score is statistically insignificant, and implies that the proportional effect on post-test scores is small.

B. Attrition Analysis

One danger in longitudinal studies is that the composition of the sample may change over time. If this change is correlated with the assignment of the treatment, then change in sample can bias the resulting estimates. This can happen, for example, if the treatment encourages particularly strong or weak students not to drop out. To check this possibility, Table 3 shows the patterns of attrition for both study years. An attritor was defined to be a student who took the baseline assessment but did not take the follow-up test. The first row shows the overall attrition rate. During the first study year, attrition is relatively low with the average under 8 percent for both groups. The difference, however, is extremely small – about eight hundredths of a percent.

During the second study year, attrition was even lower. Overall, the average attrition rate for the treatment groups was only 2 percentage points higher than for the control group. Breaking this difference up by individual treatment shows that the primary difference is the five percentage point difference in attrition rate for the machine only treatment group, a difference that, while still small, is entirely due to the refusal of a single school to participate in the study after the baseline assessment. The largest difference for the other two research groups is 1.1 percentage points in the difference for the group receiving both activities and machines.

These low attrition rates mean that it is very unlikely that changes in the sample could generate differences in the respective research groups. Any differences would have to be very large to change the composition of the sample given that only 404 students failed to take the follow-up test in the first study year and only 280 students failed to take the follow-up test in the second study year. But to check for such large differences, we estimate the relative attrition patterns between the treatment and control groups using equation (2).

The differences in baseline test scores and socioeconomic characteristics between those students who attrit and those who take the follow-up test in the control group are listed in the first column of the year 1 Thane study and the first column in the year 2 Mangaon study. In column 2 of the first study and in columns 2 through 4 of the second study year, we list the corresponding estimated differences between the control group difference and each of the treatment group differences.

In both years, the observed differences are small. In the first year, students dropping out of the control group tend to have higher pre-test scores on Marathi, are more often female, are more likely to come from the general caste category, and more likely to have parents with less education when compared to the students we observe at follow-up. On the other hand, dropouts from the treatment group are more likely to resemble those who we are able to test at follow-up. However, given the low number of students who atritt from the sample, these small differences could not significantly change the composition of the research groups.

In the second year study, the main difference between the groups is that students who drop out from the control group are more likely to have mothers with an occupation than those observed at follow-up. Again, the treatment groups differ (to the same degree) in that attritors in each group have about the same probability of having a mother with a registered occupation. The actual attrition rates, however, for the second year are even smaller than those in the first year, and it is unlikely that changes in the composition of the groups over time effected composition of the research groups.

C. Academic Treatment Effects

To assess the effects of each variation on the program, we estimate the average differences in post-test scores between the respective treatment and control groups first using equation 1, and then controlling for class characteristics, demographic characteristics, and baseline test scores using equation 3. On average, the treatments are similarly effective at improving students' English scores. The externally implemented program increases students' scores by 0.26 standard deviations and the teacher implemented interventions increased students' score by 0.36 standard deviations. When implemented by the local teachers and teaching assistants, however, the interventions increase not only English scores, but also math scores. As we show in the next section, this seems to be due to the discretion teachers have over the intensity with which the program is implemented.

Table 4 contains the results for each of the experiments. Panel A contains the results for the externally implemented treatment while Panel B contains the results for the teacher implemented treatment. The first column provides the relative change in test scores over the academic year for the control group for each section of the follow-up exam that was administered at both baseline and follow-up.[10] The following eight columns provide the estimated treatment effects first for the average of all of the treatments and then for the individual treatments. Columns 2 through 5 contain estimates without controls (equation 1) and columns 6 through 10 contain estimates with controls (equation 3). Each row then provides the results for a different section of the exam. The English competencies are grouped into Section 1 and Section 2. Section 1 focuses mainly on the identification of simple words ("cat," "hat," etc.), and identifying

[10] For the purposes of this comparison, both scores are normalized relative to the baseline control distribution.

the words that play different roles in a simple sentence while Section 2 contains more difficult questions that focus on subtle differences of pronunciation and sentence structure

Overall the results in Panel A suggest that the effect of the program on students' English skills is large and positive. The average treatment effect is about twenty six hundredths of a standard deviation and is fairly consistently distributed across the English competencies. The treatment effects are slightly higher for the easier competencies, but this difference is not statistically significant. Consistent with similarities of the research groups, the point estimates are also very similar regardless of whether or not the specification controls for the baseline characteristics of the students. Figure 3 contains a kernel density estimate of the distribution of follow-up scores for the English sections of first year of the study. Reflecting the mean differences in scores, the treatment density has significantly more mass above the control mean (zero) than the control distribution. Finally, the program seems to have no effect on students' math scores with an overall treatment effect of 0.052 standard deviations.

Panel B contains the results for the second year of the study in which the program was implemented through a teacher training program. Like the external implementation model, the results are remarkably consistent across the treatments – regardless of who implements the treatment or which tool is used, there seems to be an average change of English scores of about 0.36 standard deviations. Unlike the first year of the study, there are some minor differences in the baseline test scores of students in the treatment groups that received either both activities or just the machine activities relative to the control group. However, once we control for baseline characteristics, the estimated effects of each treatment on the English competencies are remarkably similar to the overall average treatment effect of 0.36 standard deviations. As in the first year, these results are very consistently distributed across the individual competencies. Figure 4 contains a kernel density estimate of the distribution of English scores for the respective research groups. Compared to Figure 2, the treatment groups have shifted significantly more mass to the right of the distribution, reflecting the mean differences in test scores.

The main difference between the treatments in each year of the study seems to be their effect on other subjects. Focusing on the results for the math section, the externally implemented program during the first year generated no change in math scores. The intervention implemented by teachers and teaching assistants, however, seems to have generated gains in math scores that are equivalent to effects on English scores.

D. Measure of Implementation

The main difference between the external and teacher implementation strategies, other than the actual person implementing the treatment, is that teachers may have had more discretion in how they used the program in the second year of the study. If, for example, these activities made them more productive at teaching their children English, then they may have been able to substitute some time away from the subject in order to spend more time on other competencies. This additional time on the other subject could have then produced the observed improvements in students' math scores.

To check for this, we monitored the activities of teachers in the second year of the study. The perceived sensitivity of teachers to having the behavior in the classroom challenged prevented us from collecting detailed information from the teachers themselves. However, we did arrive at a compromise in which we surveyed the children during the process of collecting attendance data and asked the children about their learning experiences in the previous day in which they had classes. While students can probably accurately perceive whether or not they used the PicTalk machine, the largest draw back to this strategy is that the students will be less likely to differentiate between the types of non-machine activities used in the classroom (e.g. Pratham activities versus general activities used by teachers) and that they may also, on occasion, try to respond with the answer they think that the surveyor wants to hear.

These results are presented in Table 5. The first panel shows the number of schools whose children report using the given activity at least 10 percent of the time over the course of the entire academic year. Not surprisingly, all schools teach English and all of them use the commonly available resources of government textbook and the chalkboard. The differences between utilization rates for the machine and activities are much starker. Only two control group schools report using machines and only nine report using activities or games. For the schools assigned to receive both treatments, almost all students report experiencing both of the treatments, and similarly for the machine only and activities only groups, schools are much more likely to use the treatment assigned to them.[11]

More importantly than whether or not the treatment was implemented is the relative frequency of implementation. Panel B provides the results of a regression comparing the average responses of students each day they were surveyed using equation 1. As expected, teachers almost always teach English using both the government

[11] It is important to note that any deviation from the treatment design is likely the result of incorrect responses. Because we were able to control directly the allocation of the Pictalk machines (and the support of the machines) and the cards that form the basis of the Pratham activities, we know that none of the research groups erroneously received either these resources or the associated training.

textbook and the blackboard, and do so at similarly high levels for each research group. The use of the interventions assigned to each group parallels the assignment of treatments. Looking at the averages for the control group, the machine is used only 1 percent of the time and activities/games are used only 17 percent of the time. In the treatment groups, however, the appropriate treatment is used 55 to 70 percentage points more than in the control group. And the use of activities in the machine only group is the same as that in the control group. This suggests that these activities are, in fact being used much more often in the treatment groups than in the control groups, but also that they are not being used on a daily basis as they were when implemented by externally hired assistants. It is thus possible that teachers may have used this additional time to focus on other subjects, such as math.

E. Heterogeneity of Effects

To gain a better understanding of both how the programs operate and the details of how they affect students in the classroom, we assess the effects of the program on subsets of the student population. We divide students by three characteristics in Tables 6 through 8. Given the salience of gender, we follow the majority of the literature and test for difference in effects between males and females. We also estimate differences by grade since material may be more or less appropriate for students of different ages, and we divide students by their baseline test performance.

The division by baseline score is particularly important because student performance on the baseline test is highly correlated with academic gains over the course of the academic year. For control group students in year 2, for example, one standard deviation increase in a student's baseline English score increases the students' performance on the post-test by 0.2 standard deviations – above the average gain of 0.47 standard deviations over the academic year. The result is that students at the top of the distribution (0.75 standard deviations and above) score above the follow-up mean by 0.28 standard deviations while those at the bottom (-0.75 standard deviations and below) score on average 0.62 standard deviations below the mean. The available data is insufficient to establish causality between baseline score and the observed treatment effect, but the correlation between baseline score and follow-up score indicates that we can use performance on the baseline test to classify students into those that are and are not thriving within the existing learning environment.

To graphically depict the relationship between baseline score and treatment effect, we begin by non-parametrically regressing students' follow-up English scores on their baseline English scores. Figures 5-7 provide these estimates using locally weighted polynomial estimators (Fan and Gijbels, 1996). Figure 5 contains the estimates for the externally implemented treatment using both machines and activities. The solid line provides the estimates for the control group and the dashed line shows the performance of the treatment group. The difference between these two lines is the estimated treatment effect for students with the given baseline score. For this treatment, the gains are relatively evenly distributed across all students.

For the teacher implemented interventions, gains are heterogeneously distributed. Figure 6 shows the results for the machine only and activities only research groups. Relative to the control group (solid line), students in the machine group (dotted line) experience an increasingly large treatment effect the higher their baseline test score – suggesting that stronger students benefited more from the intervention. The activities only intervention (dashed line), on the other hand, seems to provide larger treatment effects for weaker students. This is particularly clear when comparing the performance of students in the machine and activities group directly – lower performing students score higher in the activities only group while higher performing students score higher in the machine only group. Figure 7 shows the results for the teacher implemented activities and machine group which shows a similar distribution of gains as the activities only group except that students scoring above 0.75 standard deviations show no treatment effect while those in the activities only group do seem to show an effect.

Table 6 contains the estimated treatment effects for these sub-samples. The columns are organized as in the previous tables. Panel A contains the results for the externally implemented experiment and Panel B contains the results for the teacher implemented experiment. The first row of each panel contains the overall estimate for the sample as a reference and each subsequent row provides the respective treatment effect estimates for the specified sub-sample. All estimates are made controlling for baseline characteristics (equation 3).

Consistent with the estimates in Figure 5, the estimated treatment effects in Panel A are relatively consistently distributed. Compared to the average effect of 0.287, boys experience a slightly higher treatment effect and girls experience a slightly smaller effect, but in both cases, the differences are less than 0.025 standard deviations. Rows 4 through 6 divide the students by their baseline score on the English test. The cut-offs of -0.75 and 0.75 are motivated by Figures 6 and 7, but the results do not depend on the specific value of the cut-off – all students experience about the same gain in English scores due to the program. Finally, we also estimate treatment effects for two groups of students that generally underperform their peers – students who are older than their peers (over age for their grade) and students who hail from minority religions.

Panel B provides the results for the second experiment in which teachers implemented the interventions. These treatment effects are less evenly distributed. The effects for boys and girls are similar across all of the interventions and both girls and boys experience treatment effects of approximately 0.36 standard deviations. The effects for the student groups divided by baseline score show patterns similar to those depicted in Figures 6 and 7. The machine only treatment generates treatment effects of 0.65 standard deviations for the strongest students, but delivers smaller and statistically insignificant results for students who score below 0.75 standard deviations above the mean on the baseline English exam. The activities only and machines and activities treatments generate the opposite results. The activities only group generates gains of 0.52 standard deviations for the lowest performing students and statistically insignificant gains of 0.2 standard deviations for other students. The activities and machine group generates statistically significant gains for students who score less than 0.75 standard deviations below the mean on the baseline and for those scoring between 0.75 standard deviations below and 0.75 standard deviations above the mean. Like the activities only group, the activities and machine group does not provide statistically significant effects for the highest scoring students.

The heterogeneity in treatment effects raises the question of whether or not specific types of students would benefit more from one intervention than another. Comparing performance under the teacher and externally implemented interventions is difficult because the experiments were conducted in different localities with different samples of students. The experiment with the teacher implemented interventions, however, was explicitly set up to facilitate these comparisons.

The most direct comparison among the teacher implemented interventions can be made by simply estimating equation 3 for the activities only and activities and machine group using the machine only group as the reference group. The results of these comparisons are presented in Table 7. Panel A provides the results for activities and machine intervention while Panel B provides the results for the Activities only intervention.

Column 1 estimates the difference in treatment effect for the baseline score groups presented in Table 6. Relative to the machine only intervention, the weakest students perform significantly better in the interventions with activities than in the machine intervention (0.41 standard deviations for the activities and machine intervention and 0.57 standard deviations more in the activities only intervention). Conversely, the strongest students perform worse in these interventions than in the machine intervention. Students scoring over 0.75 standard deviations above the mean on the baseline test score

0.59 standard deviations worse in the activities and machine group and 0.4 standard deviations worse in the activities only group.

Column 2 provides a similar comparison using a slightly different specification. Rather than divide the sample into groups, we interact the baseline English score directly with the indicated treatment variable. For both interventions, the interaction term is negative, statistically significant at the one percent level, and of similar magnitudes. In the activities and machine group, students perform 0.43 standard deviations worse than in the machine group for each standard deviation they score in the baseline test. Students in the activities only group score 0.36 standard deviations less for each standard deviation scored in the baseline.

These results show that stronger and weaker students benefit differentially from the teacher implemented interventions. Column 3 tests whether or not these differences in treatment effects are statistically significant. In other words, do poorly performing students in the activities only group (or activities and machine group) experience proportionately smaller treatment effects relative to the machine group than their better performing peers. In these specifications, the interaction term for students scoring less than 0.75 standard deviations below the baseline mean is omitted and an indicator variable for the machine group is included so that the estimates on the interaction terms represent the relative difference compared to the lowest performing students. All of the differences for each group are large and statistically significant. The difference is largest for the highest and lowest performing students who experience almost a full standard deviation difference in the magnitude of the treatment effect between the respective interventions.

These differences in relative effects may be due to the structure of the interventions. Low performing students may benefit from additional interaction with the teachers through the activities interventions, learning material that may prove more difficult for them in the relatively self-paced machine intervention. Stronger students on the other hand, may find themselves relatively constrained by the pace of the teacher directed activities. The distribution of these results in the teacher implemented interventions with activities stands in stark contrast to the normal patterns of learning in Indian public schools. Incentivized by how quickly they cover the government textbook in a given academic year, teachers typically focus little on those students who lag behind their peers (Banerjee, Cole, Duflo, and Linden, 2007). Providing teachers with a different pedagogical model seems to allow them to refocus their efforts on these low performing students.

The results also suggest that there may be significant gains to targeting interventions at specific students. For example, an intervention that provided activities to low performing students and the machine intervention to high performing students may have outperformed either intervention separately. Any such change would necessarily require a reorganization of the classroom that could have ancillary effects, but based on the current estimates, providing the treatment that generates the largest treatment effect to each type of student would increase the average treatment effects by a third (0.13 standard deviations over the 0.36 standard deviation average effect).

Finally, Table 8 divides students by their grade. We divide students into two groups: those in grades 1 and 2 and those in grades 3, 4, and 5. Since the first year of the study only included students in grades 2 and 3, we separate out each year of the study into a separate column. The competencies listed in the first column are identical to those in Table 4. And as in Table 4 we present the normalized change in test scores between the pre and post-test for each control group as a reference.

For effects on English scores, the results across all groups are fairly evenly distributed. For the externally and teacher implemented interventions with both activities and machines, the are no major differences across either section of the English exams for either older or younger students and both groups of students seem to experience the same gains. For the machine only and activities only teacher groups, the treatment effects show a different pattern. Older students show similar gains in both sections of the English test while younger students only show gains on the easier competencies. It may be that for younger students, different methods of presenting the same information may be valuable for more difficult subjects.

The ancillary benefits of changes in math scores are uniform across both groups of students within the teacher and external delivery models. In both groups of students, the external treatment does not help students learn math while each of the teacher implemented interventions do.

F. Effects on Attendance

Table 9 provides the results from our directly observed attendance measures. The layout of the table is similar to that of Table 6 with each row providing the treatment effect estimate only for students in the specified groups. The groups are the same as those used in Table 6. The first column provides the average attendance rate of the control students and each column then provides an estimate of the difference in attendance rates for the respective treatment group using equation 3.

The evidence suggests that none of these interventions encourage higher attendance rates. None of the differences for the external implementation are statistically significant. For the teacher implemented version of the program, the only statistically significant effect is the 3.7 percentage point increase in attendance for low scoring students in the activities and machine group. These results are consistent with those of other studies of interventions designed to improve the learning environment (Banerjee, Cole, Duflo, and Linden, 2007; Muralidharan and Sundararaman, 2006). An improved learning environment does not seem to provide an incentive for higher levels of student participation.

V. Cost Effectiveness

While the various components of the program are effective, the relative costs of those gains are modest when compared to other interventions with similar goals. The comparison between the externally implemented treatment and the teacher implemented treatment can be made most directly by considering the teacher implemented treatment with both activities and machines. Including the full cost of the machines and material development in a one year program, we see that for the integrated machine and activity intervention there was a per student cost of $20.46 achieving 0.26 standard deviation increase in average test scores in Thane (with external instructors) and a $11.20 per student cost in Mangaon (with teacher implementation) achieving a 0.36 standard deviation increase in average test score. Measured by academic gains, PicTalk's cost per child per tenth standard deviation is $7.87 (external) and $3.11 (teacher). This, however, overestimates the true cost of the program because the machines and components are expected to last at least five years on average. Amortizing the cost of the hardware, books, and layout design over this period, the price drops to $3.19 (external) and $1.05 (internal) per child per tenth standard. The interventions tested in the second year of the study were cheaper because they relied on existing teachers and teaching assistants rather than a dedicated implementation team

The cost-effectiveness estimate for the machine only and activities only teacher implemented interventions are even lower. Amortizing the capital costs over five years yields the following costs per student per tenth of a standard deviation: $1.00 (machine only), $0.22 (activities only), $1.05 (both). Without the cost of hardware or the tech support contract, activities are clearly the cheapest intervention. It is important to remember, however, that these comparisons are based on mean scores alone and these calculations do not take into account the differences in the distribution of these effects.

Compared to other interventions that have been evaluated in similar contexts, these programs are particularly cheap (Kremer, Miguel, and Thornton (2007)). A computer assisted learning program achieved a 0.21 tenth standard deviation improvement in combined math and language scores for a cost of $7.60 per child per tenth standard deviation, a teaching assistant for remedial tutoring achieved an improvement in average combined math/language score of 0.24 at a per child per tenth standard deviation cost of $1 (Banerjee, Cole, Duflo, and Linden (2007)). A girls' scholarship program which reported average increases of 0.12 and 0.19 standard deviation in two different locations cost $1.77 to $3.53 per child per tenth standard deviation (Kremer, Miguel, and Thornton (2007)); cash incentives for teachers have been documented to achieve an average increase of only 0.07 standard deviation for a cost of $3.41 per student per tenth standard deviation (Glewwe (2003)); and textbook provision achieved an increase of just 0.04 standard deviation for a per pupil per tenth standard deviation cost of $4.01 (Glewwe (1997)). While the unadjusted per pupil cost is high relative to these programs, the gains in average score are commensurate with costs.

IV. Conclusion

The English educational techniques that we assess are generally effective with all of the treatments generating gains of 0.25-0.35 standard deviations in English scores. This is true across different combinations of implementing technologies and both rural and urban localities. Since all of these interventions (and in particular the teacher implemented ones) are inexpensive, these results suggest that changing the curriculum in developing countries may be much less difficult to do than previously expected. The results also suggest that it is important to match the appropriate technology and implementation method to particular students as some students show stronger responses to different educational methods.

Finally, contrary to much of the literature on the behaviors of teachers in developing countries, these results suggest that equipping teachers with the appropriate tools and learning methods can be more effective than delivering similar interventions through outside agencies. For example, the teacher implemented interventions generate the same gains in test scores as the externally implemented treatment. They did so, however, by using the machines less often, but managed to generate changes in students scores in ancillary subjects as well. This is something that the externally implemented program could not do – perhaps because the external implementation forced teachers to allocate an inefficiently large amount of classroom time towards English.

Further, when teachers are provided with the new methodologies, the lowest performing students show significant improvements in test scores. This is particularly surprising given the perception that teachers only care about teaching well-performing students. Combined with the evidence that different populations of students respond to different teaching methods, it may be that this perceived preference may, in part, reflect the fact that the tools available to teachers are not appropriate for lower performing students.

VI. Bibliography

Angrist, Joshua and Victor Lavy. (1997) “The Effect of a Change in Language of Instruction on the Returns to Schooling in Morocco,” *Journal of Labor Economics*. 15(1): S48-S76.

Angrist, Joshua and Victor Lavy (2001) “Does Teacher Training Effect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools,” *Journal of Labor Economics*. 19(2): 343-369.

Angrist, Joshua and Victor Lavy (2002) “New Evidence on Classroom Computers and Pupil Learning,” *The Economic Journal*. 112(482): 735-765.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden (2006) “Remedying Education: Evidence from Two Randomized Experiments in India,” *Quarterly Journal of Economics.* 122(3): 1235-1264.

Barrow, Lisa, Lisa Markman, Cecilia E. Rouse (2007) “Technology’s Edge: The Educational Benefits of Computer-Aided Instruction,” *Federal Reserve Board of Chicago working paper No. 2007-17*.

Birdsong, David (1999) “Second Language Acquisition and the Critical Period Hypothesis,” *Lawrence Erlbaum Associates*.

Boozer, M., Krueger, A. B., & Wolkon, S. (1992). “Race and school quality since Brown vs. Board of Education,” In M. Baily, & C. Winston (Eds.), *Brookings Papers on Economic Activity: Microeconomics*.

Duflo, Esther, Pascaline DuPas, and Michael Kremer. (2007) “Peer Effects, Pupil-Teacher Ratios, and Teacher Incentives,” *Manuscript*, Dartmouth College, Department of Economics.

Dynarski, Mark (2007). “Effectiveness of Reading and Mathematics Software Products: Findings from the 1st Student Cohort.” *Mathematica Report.* http://www.mathematica-mpr.com/publications/redirct_pubsdb.asp?strSite=PDFs /effectread.pdf. Accessed: May 19, 2008.

Glewwe, P., Kremer M., Moulin, S. (2003) "Textbooks and Test Scores: Evidence from a Prospective Evaluation in Kenya", *Manuscript*, Harvard University, Department of Economics.

Glewwe, P. Nauman, I., & Kremer M. (2003). "Teacher Incentives", *National Bureau of Economic Research Working Paper # 9671*.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. (2004) “Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya,” *Journal of Development Economics*. 74: 251-268.

Goolsbee, A., & Guryan, J. (2002). “The Impact of Internet Subsidies in Public Schools,” NBER Working Paper, No. 9090.

Hakuta, Kenji, Ellen Bialystok, and Edward Wiley (2003) “Critical Evidence: A Test of the Critical-Period Hypothesis for Second-language Acquisition,” *Pyschological Science*. 14(1): 31-38.

Jacob, Brian and Lars Lefgren (2004) “The Impact of Teacher Training on Student Achievement: Quasi-Experimental Evidence from School Reform Efforts in Chicago,” *Journal of Human Resources*. 39(1): 50-79.

Johnson, Jacqueline S. and Elissa L. Newport (1989) “Critical Period Effects in Second Language Learning: The Influence of Maturational State on the Acquisition of English as a Second Language,” *Cognitive Psychology*. 21(1): 60-99.

Kennedy, Mary (1998) “Form and Substance in Inservice Teacher Education,” *National Institute for Science Education, Research Monograph No 13.*

Kim, E. J. (1996) “The Sensitive Period for Second-Language Acquisition: An Experimental Study of a Lexical-decision Task with Semantic Priming and a Grammaticality Judgement Test,” *unpublished doctoral thesis University of Illinois, Urbana-Champaign*.

Kirkpatrick, H., & Cuban, L. (1998). “Computers Make Kids Smarter—Right?” *Technos Quarterly*. 70(2).

Kremer, Michael, Edward Miguel, and Rebecca Thornton. "Incentives to Learn". Jan 2007. *Harvard University Working Paper.*

Kroll, Judith and A. M. B. de Groot (2005) “Handbook of Bilingualism: Psycholinguistic Approaches,” *Oxford University Press*.

LeapFrog (2004a) “Phoenix Language First Study,” *Unpublished Research.* http://www.leapfrogschoolhouse.com/content/research/LF_Phoenix_full.pdf.

LeapFrog (2004b) “Oakland Unified School District LeapTrack Study” *Unpublished Research.* http://www.leapfrogschoolhouse.com/content/research/LT_full_OAK.pdf.

Lee, Jeong-Won and Diane Schallert (1997) “The Relative Contribution of L2 Language Proficiency and L1 Reading Ability to L2 Reading Performance: A Test of the Threshold Hypothesis in an EFL Context,” *TESOL Quarterly*. 31(4): 713-739.

Linden, Leigh (2008) “Enabling Out-of-School Learning: An Evaluation of an After-School Computer Learning Program,” *Manuscript*, Columbia University, Department of Economics.

Long, Mike (2005) “Problems with Supposed Counter-evidence to the Critical Period Hypothesis,” *International Review of Applied Linguistics in Language Teaching*.

43(4): 287-317.

Machin, Stephen, Sandra McNally, and Olmo Silva (2006) “New Technology in Schools: Is there a Payoff?" *Forthcoming Economic Journal*.

McDonald, Janet (2000) “Grammaticality Judgements in a Second Language: Influences of Age of Acquisition and Native Language,” *Applied Pyscholinguistics*. 21: 395-423.

Munshi, Kaivan and Mark Rosenzweig (2006) “Traditional Institutions Meet the Modern World: Caste, Gender, and Schooling Choice in a Globalizing Economy,” *American Economic Review* 96(4):1225-1252.

Muralidharan, Karthik and Venkatesh Sundararaman (2006) “Teacher Incentives in Developing Countries: Experimental Evidence from India,” *Manuscript,* Department of Economics, University of California at San Diego.

Oyama, S. (1976) “A Sensitive Period for the Acquisition of a Non-native Phonological System,” *Journal of Psycholinguistic Research*. 5: 261-285.

Oyama, S. (1978) “The Sensitive Period and Comprehension of Speech,” *Working Papers on Bilingualism*. 16: 1-17.

Patkowski, M.S. (1980) “The Sensitive Period for the Acquisition of Syntax in a Second Language,” *Language Learning*. 30: 449-472.

Perani, Daniela and Jubin Abutalebi (2005) “The Neural Basis of First and Second Language Processing,” *Current Opinion in Neurobiology*. 15: 202-206.

RMC Research (2003a) “Las Vegas Literacy Center Study” *Unpublished Study.* http://www.leapfrogschoolhouse.com/content/research/TLC_LVstudy.pdf.

RMC Research (2003b) “Newark Ready, Set, Leap! Study.” *Unpublished Research.* http://www.leapfrogschoolhouse.com/content/research/RMC_RSLreport.pdf.

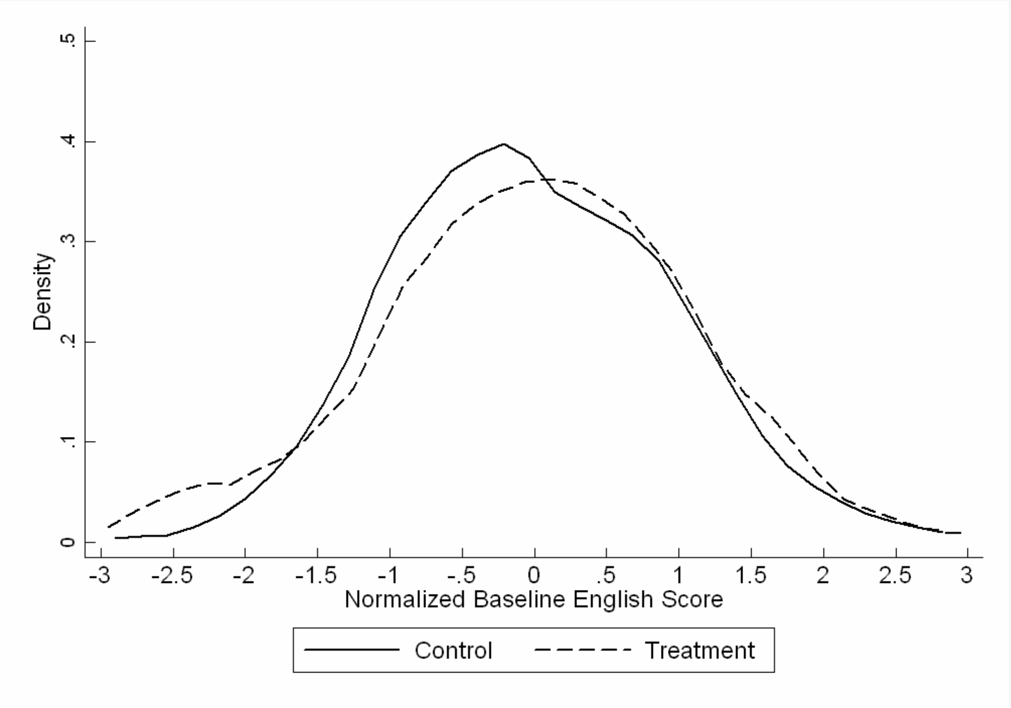
Rouse, Cecilia and Alan Krueger (2004) “Putting Computerized Instruction to the Test: A Randomized Evaluation of a “Scientifically Based” Reading Program,” *Economics of Education Review*. 23(4): 323-338.

Shim, Rosa Jinyoung (1993) “Sensitive Periods for Second Language Acquisition: A Reaction-Time Study of Korean-English Bilinguals,” *IDEAL*. 6: 43-64.

Ullman, Michael (2001) “The Neural Basis of Lexicon and Grammar in First and Second Language: the Declarative/Procedural Model,” *Bilingualism: Language and Cognition.* 4(1): 105-122.

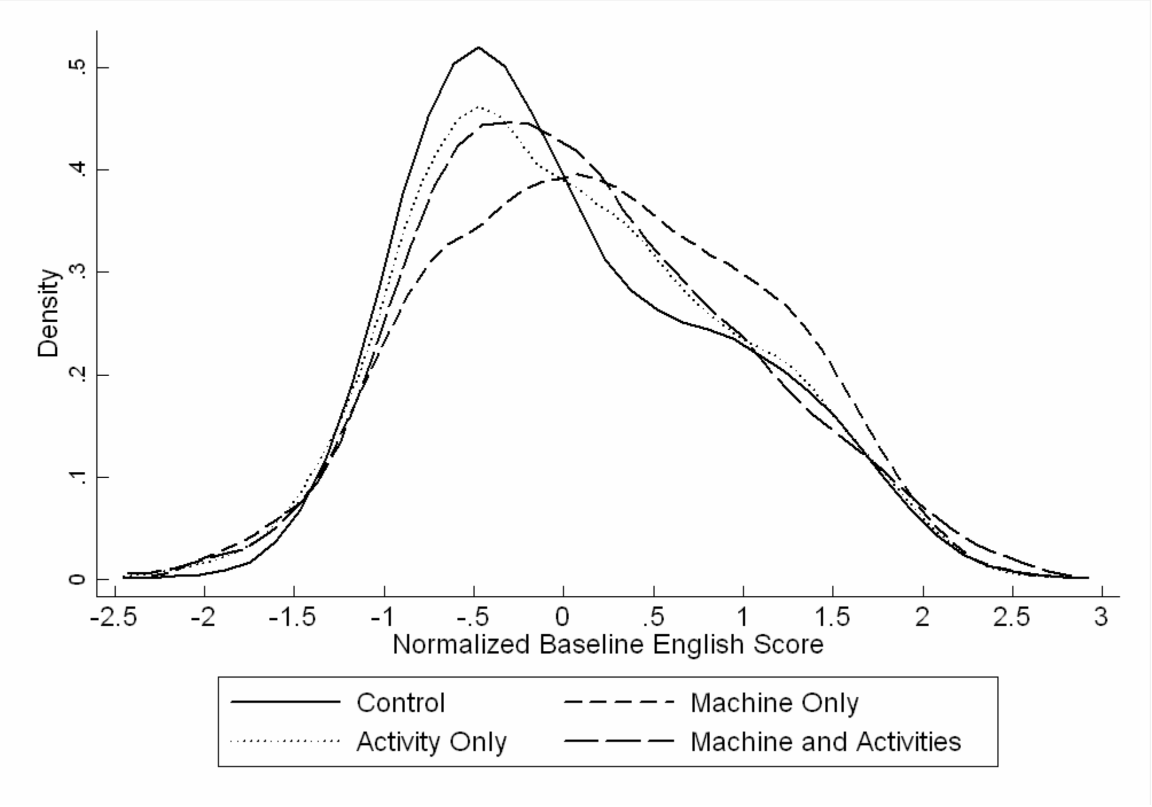
Wenglinsky, H. (1998). “Does it compute? The Relationship Between Educational Technology and Student Achievement in Mathematics.” Princeton, NJ: Educational Testing Service.

Figure 1: Baseline English Score Distribution, External Implementation (Year 1)



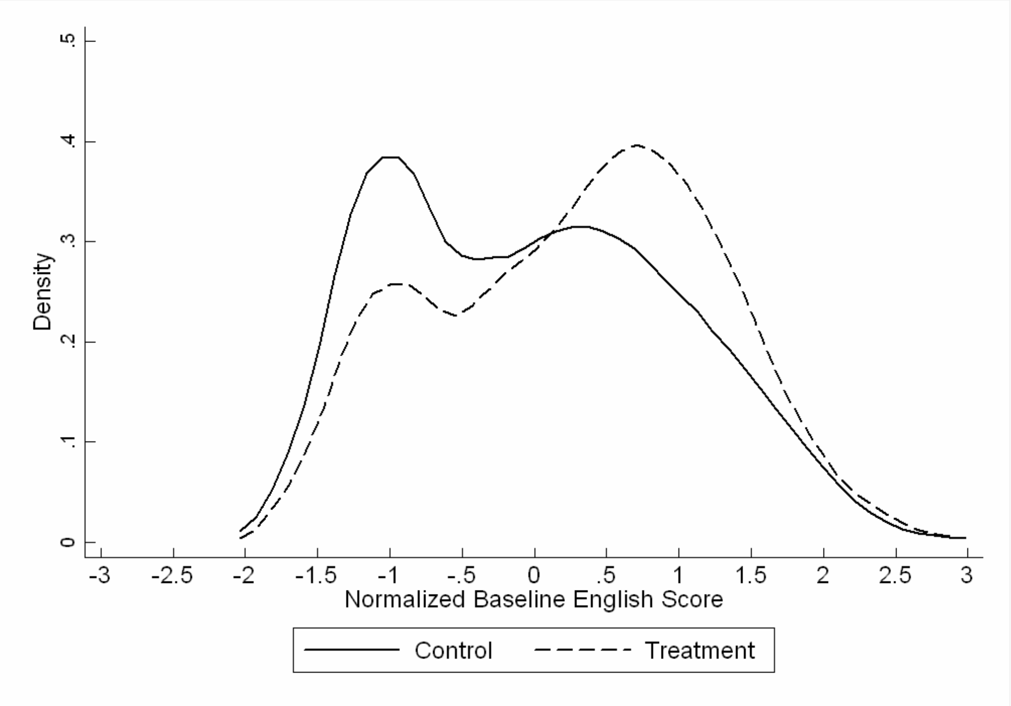
Note: Kernel density estimate with normal kernel and bandwidth of 0.4 standard deviations

Figure 2: Baseline English Score Distribution, Teacher Implementation (Year 2)



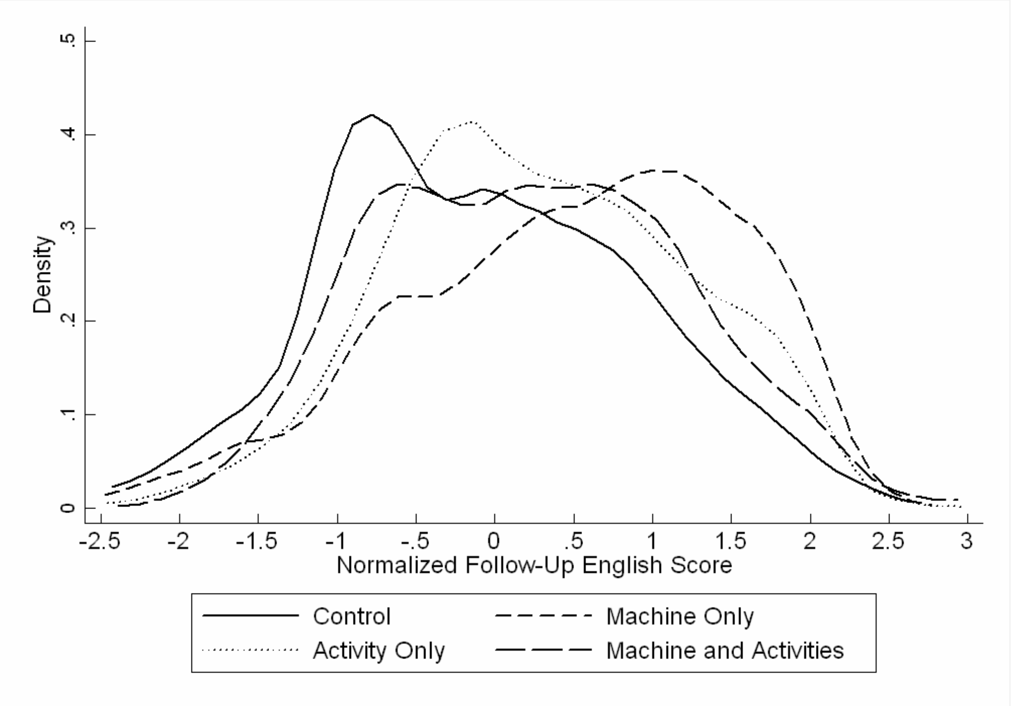
Note: Kernel density estimate with normal kernel and bandwidth of 0.4 standard deviations

Figure 3: Follow-Up English Score Distribution, External Implementation (Year 1)



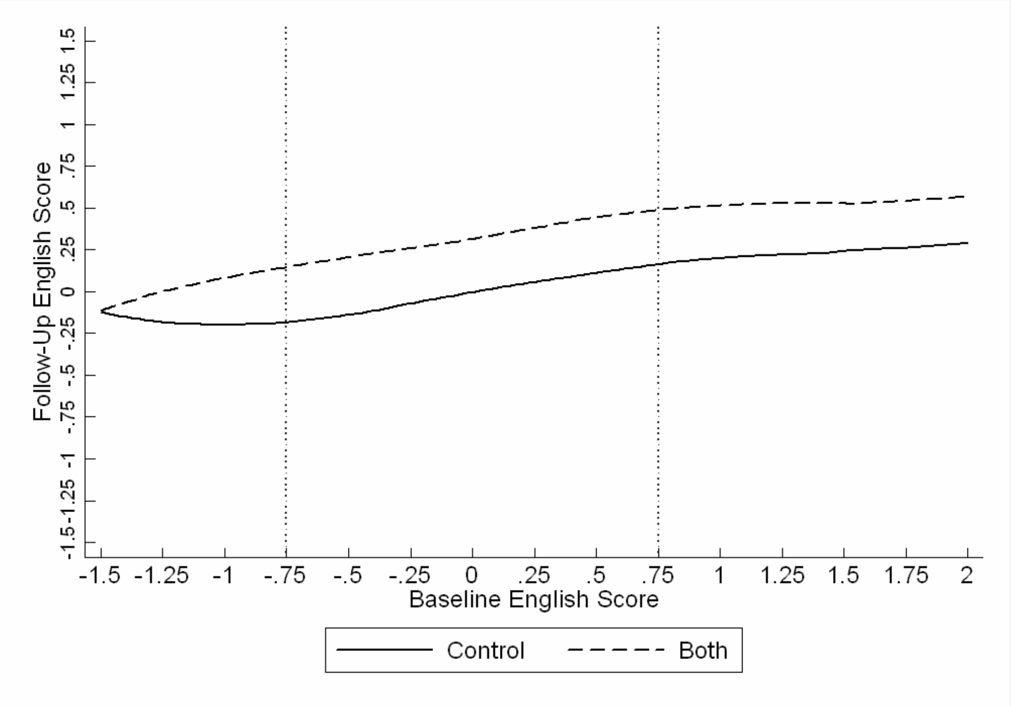
Note: Kernel density estimate with normal kernel and bandwidth of 0.4 standard deviations

Figure 4: Follow-Up English Score Distribution, Teacher Implementation (Year 2)



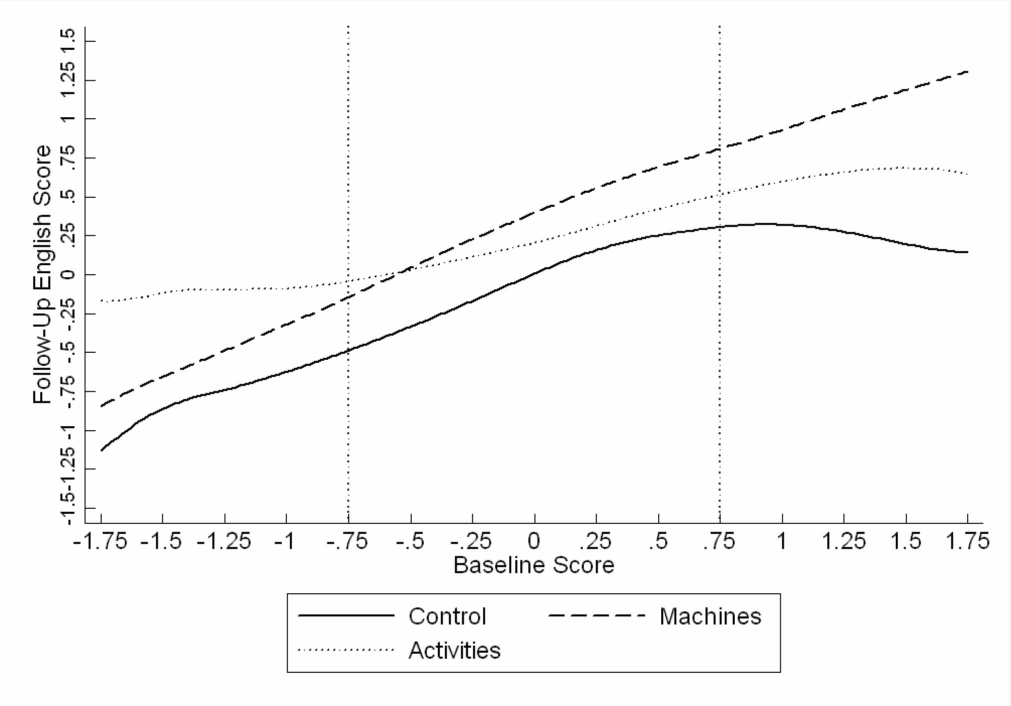
Note: Kernel density estimate with normal kernel and bandwidth of 0.4 standard deviations

Figure 5: Follow-Up English by Baseline Score, External Implementation (Year 1)



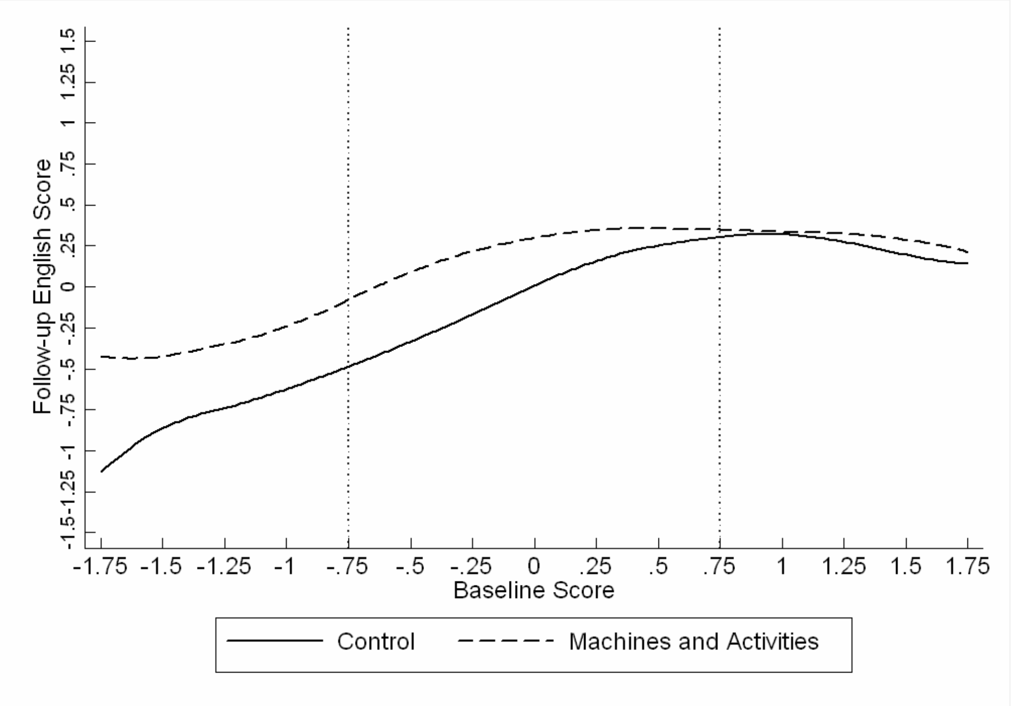
Note: Locally weighted polynomial estimate, bandwidth of 1.0 standard deviation

Figure 6: Follow-Up English by Baseline Score, Teacher Implementation (Year 2)



Note: Locally weighted polynomial estimate, bandwidth of 1.0 standard deviation

Figure 7: Follow-Up English by Baseline Score, Teacher Implementation (Year 2)



Note: Locally weighted polynomial estimate, bandwidth of 1.0 standard deviation

Table 1: Description of Sample

Sample Description	External (Year 1) Control Group	External (Year 1) Machines and Activities Group	Teacher Implementation (Year 2) Control Group	Teacher Implementation (Year 2) Machines and Activities Group	Teacher Implementation (Year 2) Machines Group	Teacher Implementation (Year 2) Activities Group
Number of Schools[†]			61	61	61	60
Number of Classes	97	97	253	254	254	246
Number of Students	2618	2699	2458	2514	2449	2324
Male Students	1329	1295	1276	1310	1194	1283
Female Students	1289	1404	1182	1204	1255	1041
Grade One			563	565	595	555
Grade Two	1281	1323	543	563	577	546
Grade Three	1337	1376	507	599	530	525
Grade Four			565	589	528	501
Grade Five			280	198	219	197

Note: This table describes the structure of the two randomized treatment-control experiments. [†]In the first year, the randomization was conducted within schools. Forty-eight schools received the treatment in grade 2 and 49 received the treatment in grade 3.

Table 2: Baseline Characteristics

	External (Year 1)			Teacher Implementation (Year 2)					
Characteristics	Control Average	Both - Control	Post-Test Marginal Correlation	Control Average	All Treat-Control	Both - Control	Machine - Control	Activities - Control	Post-Test Marginal Correlation
Panel A: School Characteristics									
Number of Students	55.354	-1.068	-0.006	40.938	-0.431	2.2	-2.55	-1.09	-0.005**
		(3.503)	(0.014)		(5.286)	(6.765)	(6.192)	(6.631)	(0.003)
Avg Pre-Test Score	-0.036	-0.005	0.007	0.022	0.073	0.091	0.161	-0.039	0.283
		(0.117)	(0.126)		(0.095)	(0.116)	(0.115)	(0.118)	(0.266)
Students per Class	28.015	0.091	0.016	9.825	-0.019	0.451	-0.513	-0.023	-0.007
		(1.177)	(0.024)		(1.160)	(1.456)	(1.370)	(1.454)	(0.015)
Panel B: Individual Characteristics									
Pre-Test English	0.045	-0.014	0.162***	0.025	0.079	0.049	0.159	0.032	0.202**
		(0.080)	(0.052)		(0.103)	(0.126)	(0.135)	(0.124)	(0.094)
Pre-Test Marathi	0.011	-0.038	0.044						
		(0.110)	(0.058)						
Pre-Test Math				0.053	0.029	0.075	0.035	-0.033	0.171**
					(0.082)	(0.120)	(0.107)	(0.108)	(0.078)
Panel C: Student and Family Characteristics									
Male	0.508	-0.033*	0.198***	0.519	-0.004	-0.036	-0.006	0.038	0.027
		(0.020)	(0.056)		(0.022)	(0.037)	(0.016)	(0.038)	(0.037)
Age	7.501	0.005	-0.015	7.712	-0.022	-0.168	0.067	0.061	-0.014
		(0.141)	(0.028)		(0.145)	(0.180)	(0.172)	(0.156)	(0.026)
Minority Religion	0.145	-0.016	0.041	0.108	-0.028	-0.004	-0.033	-0.052**	-0.359***
		(0.015)	(0.102)		(0.023)	(0.028)	(0.028)	(0.023)	(0.126)
Lower Castes	0.366	0.019	0.054	0.573	-0.043	-0.117	0.001	0	-0.165
		(0.026)	(0.089)		(0.073)	(0.086)	(0.088)	(0.095)	(0.140)
Mothers Education	3.762	-0.135	0.012						
		(0.273)	(0.009)						
Father's Education	4.965	-0.222	0.007						
		(0.361)	(0.009)						
Mother Has Occupation	0.465	-0.007	-0.051	0.21	0.045	-0.003	0.104	0.045	0.238
		(0.025)	(0.061)		(0.059)	(0.067)	(0.084)	(0.082)	(0.210)
Family Has Electricity				0.922	0.003	0.014	0.013	-0.02	0.26
					(0.020)	(0.020)	(0.024)	(0.028)	(0.177)
Family Has Running Water				0.161	0.145**	0.072	0.195*	0.184*	-0.127
					(0.074)	(0.083)	(0.107)	(0.102)	(0.188)
Number of Assets				0.803	-0.019	-0.009	-0.021	-0.027	0.094*
					(0.079)	(0.100)	(0.087)	(0.096)	(0.053)
Livestock				1.911	-0.163	-0.212	-0.249	-0.02	0.024
					(0.152)	(0.231)	(0.170)	(0.232)	(0.043)
Parent Living Elsewhere				0.292	0.005	0.034	-0.05	0.024	-0.139
					(0.072)	(0.095)	(0.081)	(0.095)	(0.191)

Note: Each cell contains the average difference between the indicated treatment and control group for the specified baseline characteristic. All estimates were made using equation 1. * indicates significance at the 10 percent significance level, ** the 5 percent level, and *** the 10 percent level. All estimated standard errors are clustered at the school level.

Table 3: Characteristics of Attriting Students

	External (Year 1)		Teacher Implementation (Year 2)				
	Control	Both Diff-	Control	All Diff-	Both Diff-	Mach Diff-	Act Diff -
Characteristics	Diff	Cont Diff	Diff	Cont Diff	Cont Diff	Cont Diff	Cont Diff
Panel A: Overall							
Overall attrition rate	0.084	-0.008	0.021	0.02	0.011	0.052	-0.001
		(0.028)		(0.018)	(0.013)	(0.051)	(0.009)
Panel B: Test Scores							
Pre-Test English	0.075	-0.171	-0.269**	0.251	0.187	0.241	0.22
	(0.067)	(0.160)	(0.135)	(0.217)	(0.266)	(0.223)	(0.299)
Pre-Test Marathi	0.199***	-0.283					
	(0.077)	(0.233)					
Pre-Test Math			-0.075	-0.348	-0.082	-0.726**	0.428
			(0.157)	(0.364)	(0.298)	(0.335)	(0.371)
Panel C: Student and Family Characteristics							
Male	-0.080**	0.130**	-0.071	0.043	0.013	0.057	0.073
	(0.040)	(0.064)	(0.080)	(0.066)	(0.077)	(0.068)	(0.098)
Age	0.037	-0.209	0.369	0.373	-0.542	0.717**	0.796
	(0.175)	(0.273)	(0.309)	(0.311)	(0.380)	(0.311)	(0.677)
Minority Religion	-0.035	-0.046	-0.086*	0.063	0.098	0.045	0.058
	(0.023)	(0.041)	(0.050)	(0.049)	(0.081)	(0.050)	(0.042)
Lower Castes	-0.157***	0.155	-0.081	-0.331*	-0.216	-0.435**	-0.259
	(0.031)	(0.101)	(0.079)	(0.188)	(0.186)	(0.196)	(0.199)
Mothers Education	-1.982***	1.283					
	(0.270)	(0.782)					
Father's Education	-2.923***	2.218**					
	(0.310)	(0.871)					
Mother Has Occupation	0.067	-0.278**	0.594***	-0.546***	-0.585**	-0.453**	-0.536**
	(0.057)	(0.120)	(0.105)	(0.197)	(0.231)	(0.225)	(0.214)
Family Has Electricity			-0.208***	0.094	0.163	-0.015	0.053
			(0.072)	(0.137)	(0.136)	(0.174)	(0.232)
Family Has Running Water			-0.009	0.322*	0.563***	0.131	0.059
			(0.103)	(0.177)	(0.185)	(0.193)	(0.229)
Number of Assets			-0.262	-0.06	-0.328	0.289	0.138
			(0.230)	(0.251)	(0.242)	(0.263)	(0.300)
Livestock			-0.451	-0.245	-0.482	0.119	0.035
			(0.337)	(0.419)	(0.466)	(0.403)	(0.699)
Parent Living Elsewhere			0.024	-0.15	-0.304*	0.02	-0.005
			(0.126)	(0.176)	(0.174)	(0.212)	(0.214)

Note: The table compares the average attrition patterns between the indicated treatment and control groups. The control column contains average differences between attriting and non-attriting students for the indicated characteristic. The treatment columns then contain the relative differences between attritors and non-attritors in the specified treatment and contrtol groups using equation 2. * indicates significance at the 10 percent significance level, ** the 5 percent level, and *** the 10 percent level. All estimated standard errors are clustered at the school level.

Table 4: Post-Test Scores

		Without Controls				With Controls			
	Control	All Treat-	Both -	Machines-	Activities-	All Treat-	Both -	Machines-	Activities-
Competency	**Post-Pre**	**Control**	**Control**	**Control**	**Control**	**Control**	**Control**	**Control**	**Control**
Panel A: External Implementation (Year 1)									
English									
Section 1 Total	0.218		0.298***				0.297***		
			(0.098)				(0.090)		
Section 2 Total			0.224**				0.229**		
			(0.096)				(0.092)		
English Total			**0.276*****				**0.278*****		
			(0.100)				**(0.093)**		
Math Total			0.038				0.052		
			(0.075)				(0.071)		
Post-Test Total			**0.253*****				**0.258*****		
			(0.098)				**(0.091)**		
Panel B: Teacher Implementation (Year 2)									
English									
Section 1 Total	0.703	0.310**	0.162	0.498***	0.301	0.316**	0.269*	0.323**	0.309**
		(0.149)	(0.184)	(0.178)	(0.183)	(0.125)	(0.142)	(0.139)	(0.147)
Section 2 Total	0.212	0.344*	0.305	0.459**	0.28	0.305**	0.304*	0.284*	0.25
		(0.182)	(0.211)	(0.231)	(0.199)	(0.150)	(0.168)	(0.148)	(0.154)
English Total	**0.474**	**0.380****	**0.279**	**0.546****	**0.338***	**0.355****	**0.328****	**0.345****	**0.320****
		(0.170)	**(0.203)**	**(0.216)**	**(0.189)**	**(0.149)**	**(0.165)**	**(0.152)**	**(0.160)**
Math Total	0.531	0.301***	0.274**	0.388***	0.250*	0.341***	0.391***	0.284**	0.319***
		(0.106)	(0.135)	(0.142)	(0.137)	(0.087)	(0.103)	(0.117)	(0.102)
Post-Test Total	**0.522**	**0.401****	**0.304**	**0.568*****	**0.352***	**0.384*****	**0.367****	**0.366****	**0.348****
		(0.167)	**(0.200)**	**(0.212)**	**(0.185)**	**(0.144)**	**(0.161)**	**(0.150)**	**(0.156)**

Note: This table presents the average treatment effects for each treatment. The control columns contain the average follow-up scores for studnets in the control group, and the treatment columns contain the average difference in follow-up scores between the treatment and control students. Estimates for columns 2 through 4 are simple differences without controling for baseline characteristics (equation 1) and estimates in columns 6 through 9 are simple differences controling for baseline characteristics (equation 3). * indicates significance at the 10 percent significance level, ** the 5 percent level, and *** the 10 percent level. All estimated standard errors are clustered at the school level.

Table 5: Utilization of Activities, Teacher Implemenation (Year 2)

English Activity	Control Group	Both Treatments	Machine Group	Activities Group
Panel A: Schools Using (Number of Schools)				
English Classes	61	61	61	60
Textbook	61	61	61	60
Chalkboard	61	61	61	60
Machine	2	59	61	3
Activities	9	52	19	50
Panel B: Average Utilization (Treatment Group Relative to Control)				
English Classes	0.945	-0.006	-0.007	0.002
		(0.019)	(0.018)	(0.018)
Textbook	0.978	0	-0.011	-0.006
		(0.010)	(0.013)	(0.013)
Chalkboard	0.969	0.006	-0.018	-0.021
		(0.012)	(0.017)	(0.017)
Machine	0.01	0.710***	0.635***	0.019
		(0.031)	(0.026)	(0.014)
Activities	0.165	0.564***	-0.005	0.546***
		(0.076)	(0.077)	(0.077)

Note: Table contains the results of questions posed to children about the activities experienced in class on the previous school day. Panel A contains the number of schools in which the specified activity is recorded a minimum of 10 percent of the time. Panel B estiamtes the average differences in utilization rates of each activity at the child level (using equation 1). * indicates significance at the 10 percent significance level, ** the 5 percent level, and *** the 10 percent level. All estimated standard errors are clustered at the school level.

Table 6: Heterogeneity in the Treatment Effects

Characteristics	Control Group	All Treat-Control	Both - Control	Machine-Control	Activities-Control
Panel A: External Implementation (Year 1)					
Entire Sample	0.071		0.287***		
			(0.092)		
Male	0.136		0.312***		
			(0.101)		
Female	0.002		0.271***		
			(0.098)		
Baseline < -0.75	-0.167		0.241*		
			(0.126)		
-0.75 > Baseline < 0.75	0.053		0.317***		
			(0.096)		
Baseline > 0.75	0.379		0.302**		
			(0.132)		
Panel B: Internal Implementation (Year 2)					
Entire Sample	-0.087	0.355**	0.328**	0.345**	0.320**
		(0.149)	(0.165)	(0.152)	(0.160)
Male	-0.065	0.338**	0.288*	0.354**	0.308*
		(0.150)	(0.170)	(0.158)	(0.163)
Female	-0.11	0.374**	0.382**	0.338**	0.327**
		(0.151)	(0.165)	(0.146)	(0.162)
Baseline < -0.75	-0.623	0.389**	0.529***	0.096	0.521***
		(0.158)	(0.152)	(0.174)	(0.185)
-0.75 > Baseline < 0.75	-0.064	0.300**	0.347**	0.247	0.211
		(0.141)	(0.162)	(0.153)	(0.146)
Baseline > 0.75	0.278	0.415*	0.072	0.646***	0.215
		(0.212)	(0.204)	(0.193)	(0.225)

Note: The cells in the first column provide the average English score on the follow-up exam for the specified subsample. In the remaining columns, each cell represents the average effect of the given treatment for the specified subsample. * indicates significance at the 10 percent significance level, ** the 5 percent level, and *** the 10 percent level. All estimated standard errors are clustered at the school level.

Table 7: Student Performance Relative to Machine Teacher Intervention (Year 2)

Dependent Variable	English All Post	English All Post	English All Post
Independent Variable	(1)	(2)	(3)
Panel A: Activities and Machine			
Both*(Baseline Eng < -0.75)	0.405*		
	(0.211)		
Both*(-0.75 < Baseline Eng < 0.75)	-0.026		-0.431**
	(0.172)		(0.203)
Both*(Baseline Eng > 0.75)	-0.590***		-0.995***
	(0.192)		(0.258)
Both		-0.04	0.405*
		(0.149)	(0.211)
Both*Baseline Eng		-0.430***	
		(0.104)	
Panel B: Activities			
Activities*(Baseline Eng < -0.75)	0.566***		
	(0.190)		
Activities*(-0.75 < Baseline Eng < 0.75)	-0.109		-0.675***
	(0.151)		(0.176)
Activities*(Baseline Eng > 0.75)	-0.401**		-0.967***
	(0.181)		(0.243)
Activities		-0.026	0.566***
		(0.137)	(0.190)
Activities*Baseline Eng		-0.358***	
		(0.110)	

Note: Each panel contains the results from regressions of the specified treatment against the machine only treatment group for the teacher implemented experiment (year 2). * indicates treatment at the 10 percent level, ** the 5 percent level, and *** the 1 percent level. All estimated standard errors are clustered at the school level.

Table 8: Treatment Effect by Grade

	Earlier Grades					Later Grades				
Competency	**Control Post-Pre**	**All Treat-Control**	**Both - Control**	**Machines-Control**	**Activities-Control**	**Control Post-Pre**	**All Treat-Control**	**Both - Control**	**Machines-Control**	**Activities-Control**
Panel A: External Implementation (Year 1)										
English										
Section 1 Total	0.185		0.327**			0.243		0.266**		
			(0.130)					(0.114)		
Section 2 Total			0.214					0.244*		
			(0.148)					(0.132)		
English Total			**0.286****					**0.271****		
			(0.141)					**(0.127)**		
Math Total			0.033					0.072		
			(0.128)					(0.125)		
Post-Test Total			**0.261***					**0.256****		
			(0.141)					**(0.122)**		
Panel B: Teacher Implementation (Year 2)										
English										
Section 1 Total	0.818	0.333***	0.323**	0.25	0.359**	0.606	0.309**	0.253	0.385***	0.286*
		(0.125)	(0.135)	(0.153)	(0.144)		(0.137)	(0.164)	(0.127)	(0.167)
Section 2 Total	0.293	0.200*	0.328**	0.146	0.046	0.143	0.391**	0.293	0.395**	0.429**
		(0.121)	(0.153)	(0.137)	(0.128)		(0.184)	(0.195)	(0.172)	(0.191)
English Total	**0.578**	**0.287****	**0.362****	**0.213**	**0.203**	**0.387**	**0.418****	**0.321**	**0.455*****	**0.436****
		(0.123)	**(0.142)**	**(0.146)**	**(0.133)**		**(0.179)**	**(0.197)**	**(0.165)**	**(0.197)**
Math Total	0.878	0.315**	0.366**	0.278*	0.286**	0.239	0.358***	0.408***	0.250*	0.359***
		(0.125)	(0.163)	(0.155)	(0.142)		(0.101)	(0.118)	(0.131)	(0.120)
Post-Test Total	**0.662**	**0.310****	**0.386*****	**0.238**	**0.230***	**0.405**	**0.451*****	**0.373****	**0.467*****	**0.466****
		(0.124)	**(0.146)**	**(0.146)**	**(0.136)**		**(0.173)**	**(0.188)**	**(0.163)**	**(0.188)**

Note: Table contains the average treatment effects disaggregated by grade. Earlier grades refer to grade 2 in panel A and grades 1-2 in panel B, later grades refer to grade 3 in panel A and grades 3-5 in panel B. All estimates are simple differences in follow-up scores controling for baseline characteristics (equation 3). * indicates significance at the 10 percent significance level, ** the 5 percent level, and *** the 10 percent level. All estimated standard errors are clustered at the school level.

Table 9: Effects on Attendance

Characteristics	Control Group	All Treat-Control	Both - Control	Machine-Control	Activities-Control
Panel A: External Implementation (Year 1)					
Entire Sample	0.878		0		
			(0.007)		
Male	0.879		0.004		
			(0.008)		
Female	0.877		-0.005		
			(0.009)		
Baseline < -0.75	0.877		-0.005		
			(0.016)		
-0.75 > Baseline < 0.75	0.88		0		
			(0.009)		
Baseline > 0.75	0.872		0.004		
			(0.016)		
Panel B: Internal Implementation (Year 2)					
Entire Sample	0.939	0.009	0.007	0.009	0.007
		(0.008)	(0.009)	(0.009)	(0.010)
Male	0.938	0.007	0.005	0.009	0.006
		(0.009)	(0.010)	(0.009)	(0.011)
Female	0.941	0.01	0.01	0.01	0.009
		(0.008)	(0.009)	(0.009)	(0.010)
Baseline < -0.75	0.92	0.02	0.037*	0.015	0.011
		(0.016)	(0.019)	(0.017)	(0.023)
-0.75 > Baseline < 0.75	0.943	0.008	0.004	0.009	0.009
		(0.007)	(0.008)	(0.008)	(0.009)
Baseline > 0.75	0.932	0.008	0.004	0.006	-0.001
		(0.014)	(0.014)	(0.013)	(0.014)

Note: Table contains estimated differences in attendance rates. Estimates are made by comparing the average attendance rates at the child level controling for baseline characteristics (equation 3). * indicates significance at the 10 percent significance level, ** the 5 percent level, and *** the 10 percent level. All estimated standard errors are clustered at the school level.

VII. Appendix

As mentioned in Section 3, the distribution of baseline scores by school for the second year of the study was extremely flat and thin at the tails. The small numbers of schools in the tails resulted in a slightly uneven distribution of schools between the research groups. To focus our analysis on a more homogenous group of schools, we trimmed the top and bottom 10 percent of the entire sample before conducting the analysis in the previous sections.

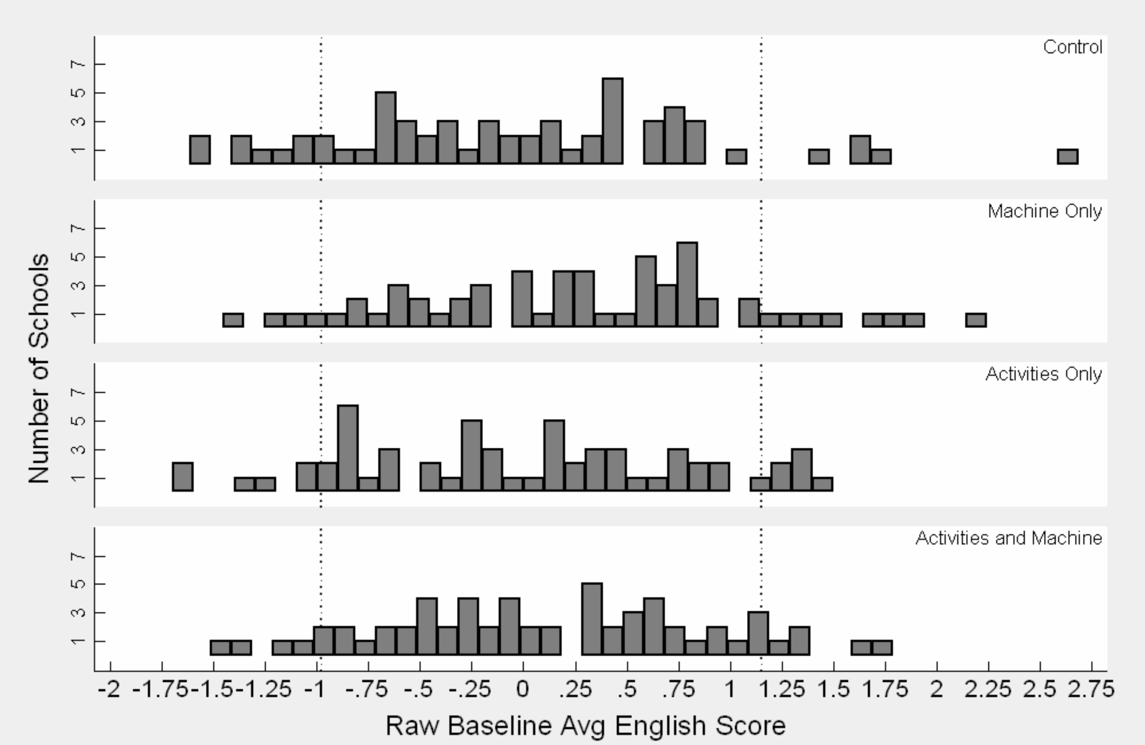
The rationale for trimming the sample is graphically depicted in Figure A1 which contains the entire distribution of normalized baseline average school test scores by research group. The dotted vertical line represented the 10 percent cut-off values (-0.9798 and 1.156 standard deviations). Comparing the distributions between the control and machine groups, the machine group has almost twice as many schools in the upper tail of the distribution (5 versus 8 schools) while the control group has twice as many schools in the bottom tail of the distribution (8 versus 4 schools). Outside of the tails, however, the distributions are much more similar.

The average effects of these changes are presented in Tables A1 and A2. The baseline characteristics of the full sample are presented in the first group of results in Table A1. The table is generally organized just like those in Table 2. Turning to the differences for the entire sample, the differences between the machine group and the control group are evident. On average, the machine group schools have an average baseline score of 0.29 standard deviations more than the control group. Since larger schools have higher pre-test scores, the differences at the student level are even larger – an average difference of 0.40 standard deviations in the baseline English test and then 0.29 standard deviations on the math section. There are also differences in the fraction of students whose mother has an occupation, whether or not the family has running water, and whether or not parents are living elsewhere, but as shown in Table 2 column 9, these variables are only weakly correlated with the students' test scores.

Trimming the sample makes the research groups much more similar. Comparing the results of the trimmed sample to those of the entire sample, the differences in test scores are significantly reduced. The difference in the student level average English score is reduced to 0.16 standard deviations and the difference in average math scores is reduced to 0.035 standard deviations. As explained in Section IV.A., these differences are small relative to the marginal correlation between the baseline and follow-up scores.

The treatment effects estimated for each sample is presented in Table A2. The layout of the table is similar to Table 4, but as in Table A1, we present the results for the entire sample and the trimmed sample. We also provide estimates for individual competencies as well as the section totals that are presented in Table 4. Comparing the trimmed sample effects to the full sample results, the pattern of effects is the same across both samples, but the point estimates are larger for the trimmed samples (ranging up to 0.15 standard deviations larger for the section totals). We are more confident in the trimmed estimates because they are created using a more homogenous sample.

Figure A1: Normalized Baseline Score Distribution, Teacher Implementation



Note: Dotted verticle lines represent cut-offs for top and bottom 10 percent of sample.

Table A1: Teacher Implementation (Year 2) Baseline Characteristics

	Untrimmed Sample					Trimmed Sample				
Characteristics	**Control Group**	**All Treat-Control**	**Both - Control**	**Machine - Control**	**Activities - Control**	**Control Group**	**All Treat-Control**	**Both - Control**	**Machine - Control**	**Activities - Control**
Panel A: School Characteristics										
Number of Students	40.295	-0.251	-0.131	0.918	-1.562	40.938	-0.431	2.2	-2.55	-1.09
		(4.856)	(6.001)	(6.402)	(5.677)		(5.286)	(6.765)	(6.192)	(6.631)
Avg Pre-Test Score	-0.006	0.136	0.12	0.289*	-0.002	0.022	0.073	0.091	0.161	-0.039
		(0.127)	(0.150)	(0.154)	(0.154)		(0.095)	(0.116)	(0.115)	(0.118)
Students per Class	9.711	-0.052	-0.085	0.186	-0.264	9.825	-0.026	0.432	-0.513	-0.023
		(1.056)	(1.296)	(1.378)	(1.241)		(1.161)	(1.458)	(1.370)	(1.454)
Panel B: Individual Characteristics										
Pre-Test English	0	0.204	0.11	0.397**	0.096	0.025	0.079	0.049	0.159	0.032
		(0.141)	(0.155)	(0.187)	(0.162)		(0.103)	(0.126)	(0.135)	(0.124)
Pre-Test Math	-0.001	0.191*	0.176	0.291*	0.099	0.053	0.029	0.075	0.035	-0.033
		(0.100)	(0.127)	(0.148)	(0.120)		(0.082)	(0.120)	(0.107)	(0.108)
Male	0.519	0.001	-0.032	0.002	0.033	0.519	-0.004	-0.036	-0.006	0.038
		(0.019)	(0.033)	(0.015)	(0.031)		(0.022)	(0.037)	(0.016)	(0.038)
Age	7.703	-0.024	-0.117	0.032	0.014	7.712	-0.022	-0.168	0.067	0.061
		(0.124)	(0.173)	(0.150)	(0.131)		(0.145)	(0.180)	(0.172)	(0.156)
Minority Religion	0.117	-0.040*	-0.02	-0.038	-0.064***	0.108	-0.028	-0.004	-0.033	-0.052**
		(0.023)	(0.028)	(0.026)	(0.024)		(0.023)	(0.028)	(0.028)	(0.023)
Lower Castes	0.566	-0.02	-0.117	0.021	0.037	0.573	-0.043	-0.117	0.001	0
		(0.068)	(0.081)	(0.078)	(0.086)		(0.073)	(0.086)	(0.088)	(0.095)
Mother Has Occupation	0.263	0.026	-0.043	0.093	0.029	0.21	0.045	-0.003	0.104	0.045
		(0.059)	(0.067)	(0.078)	(0.078)		(0.059)	(0.067)	(0.084)	(0.082)
Family Has Electricity	0.906	0.016	0.021	0.02	0.005	0.922	0.003	0.014	0.013	-0.02
		(0.023)	(0.024)	(0.027)	(0.028)		(0.020)	(0.020)	(0.024)	(0.028)
Family Has Running Water	0.159	0.128**	0.076	0.123	0.189**	0.161	0.145**	0.072	0.195*	0.184*
		(0.063)	(0.074)	(0.090)	(0.087)		(0.074)	(0.083)	(0.107)	(0.102)
Number of Assets	0.788	0.023	0.014	0.074	-0.022	0.803	-0.019	-0.009	-0.021	-0.027
		(0.069)	(0.089)	(0.085)	(0.082)		(0.079)	(0.100)	(0.087)	(0.096)
Livestock	1.936	-0.144	-0.191	-0.235	-0.002	1.911	-0.163	-0.212	-0.249	-0.02
		(0.149)	(0.213)	(0.234)	(0.195)		(0.152)	(0.231)	(0.170)	(0.232)
Parent Living Elsewhere	0.325	-0.041	0.004	-0.106	-0.019	0.292	0.005	0.034	-0.05	0.024
		(0.068)	(0.089)	(0.076)	(0.085)		(0.072)	(0.095)	(0.081)	(0.095)

* indicates significance at the 10 percent significance level, ** the 5 percent level, and *** the 10 percent level. All estimated standard errors are clustered at the school level.

Table A2: Teacher Implementation (Year 2), Post-Test Scores

	Untrimmed Sample					Trimmed Sample				
	Control	All Treat-	Both -	Machines-	Activities-	Control	All Treat-	Both -	Machines-	Activities-
Competency	Post-Pre	Control	Control	Control	Control	Post-Pre	Control	Control	Control	Control
Section 1										
Familiar Words	0.373	0.066	0.001	0.168	0.091	0.333	0.134	0.064	0.231*	0.106
		(0.112)	(0.144)	(0.125)	(0.126)		(0.118)	(0.152)	(0.134)	(0.127)
Follow Instructions	0.524	0.147	0.163	0.1	0.189*	0.597	0.15	0.123	0.149	0.167
		(0.099)	(0.111)	(0.133)	(0.110)		(0.104)	(0.118)	(0.130)	(0.112)
Identify Picture	0.387	0.142	0.158	0.079	0.152	0.304	0.262*	0.236	0.214	0.266*
		(0.125)	(0.141)	(0.136)	(0.136)		(0.135)	(0.155)	(0.133)	(0.149)
Identify Verb	0.558	0.225*	0.252*	0.193	0.241*	0.468	0.311**	0.304**	0.305**	0.280**
		(0.124)	(0.136)	(0.136)	(0.134)		(0.133)	(0.146)	(0.129)	(0.141)
Identify Noun	0.683	0.091	0.075	0.098	0.134	0.674	0.138	0.105	0.153	0.153
		(0.097)	(0.113)	(0.116)	(0.119)		(0.112)	(0.125)	(0.137)	(0.134)
Section 1 Total	0.741	0.216*	0.218	0.193	0.256*	0.703	0.316**	0.269*	0.323**	0.309**
		(0.120)	(0.139)	(0.144)	(0.141)		(0.125)	(0.142)	(0.139)	(0.147)
Section 2										
Identify, Focusing on End	0.244	0.14	0.205	0.093	0.06	0.141	0.269**	0.305**	0.217	0.201
		(0.131)	(0.140)	(0.147)	(0.129)		(0.137)	(0.153)	(0.136)	(0.139)
Identiy, Focusing on Constants	0.249	0.179	0.148	0.168	0.199	0.236	0.288**	0.208	0.324**	0.308**
		(0.128)	(0.141)	(0.139)	(0.136)		(0.145)	(0.162)	(0.145)	(0.155)
Identify, Focusing on Middle	0.262	0.138	0.191	0.095	0.077	0.156	0.269*	0.296*	0.231	0.211
		(0.133)	(0.144)	(0.150)	(0.133)		(0.139)	(0.157)	(0.142)	(0.145)
Identify Word	0.262	0.117	0.169	0.088	0.026	0.159	0.243*	0.267*	0.225	0.143
		(0.134)	(0.146)	(0.154)	(0.134)		(0.141)	(0.160)	(0.145)	(0.147)
Identify Sentence	0.373	0.108	0.139	0.078	0.082	0.286	0.212***	0.218**	0.196**	0.193**
		(0.073)	(0.087)	(0.087)	(0.085)		(0.076)	(0.094)	(0.091)	(0.088)
Section 2 Total	0.303	0.163	0.201	0.126	0.106	0.212	0.305**	0.304*	0.284*	0.25
		(0.141)	(0.152)	(0.157)	(0.141)		(0.150)	(0.168)	(0.148)	(0.154)
English Total	0.545	0.214	0.235	0.181	0.198	0.474	0.355**	0.328**	0.345**	0.320**
		(0.142)	(0.154)	(0.163)	(0.149)		(0.149)	(0.165)	(0.152)	(0.160)
Section 3										
Number Recognition	0.516	-0.019	0.122	-0.031	-0.072	0.482	0.048	0.197	0.02	-0.033
		(0.095)	(0.117)	(0.115)	(0.112)		(0.105)	(0.123)	(0.134)	(0.129)
Addition	0.473	0.11	0.141	0.13	0.07	0.442	0.201**	0.222**	0.222*	0.155
		(0.084)	(0.098)	(0.101)	(0.106)		(0.093)	(0.106)	(0.113)	(0.113)
Subtraction	0.502	0.262***	0.231**	0.251**	0.322***	0.406	0.381***	0.353***	0.349***	0.419***
		(0.090)	(0.104)	(0.102)	(0.104)		(0.094)	(0.107)	(0.105)	(0.112)
Multiplication	0.382	0.133*	0.14	0.088	0.109	0.355	0.199**	0.235**	0.111	0.187*
		(0.080)	(0.099)	(0.093)	(0.090)		(0.089)	(0.101)	(0.108)	(0.101)
Math Total	0.59	0.207**	0.246**	0.184*	0.196*	0.531	0.341***	0.391***	0.284**	0.319***
		(0.089)	(0.106)	(0.109)	(0.101)		(0.087)	(0.103)	(0.117)	(0.102)
Post-Test Total	0.596	0.232*	0.258*	0.2	0.216	0.522	0.384***	0.367**	0.366**	0.348**
		(0.140)	(0.151)	(0.161)	(0.146)		(0.144)	(0.161)	(0.150)	(0.156)

* indicates significance at the 10 percent significance level, ** the 5 percent level, and *** the 10 percent level. All estimated standard errors are clustered at the school level.