

Teacher Incentives

Paul Glewwe,^{*} Nauman Ilias,[♦] and Michael Kremer[•]

June, 2008

Abstract

Advocates of teacher incentive programs argue that such programs can strengthen the weak incentives faced by teachers, while opponents argue that they lead to “teaching to the test”, which increases short-run test scores but not necessarily long-run learning. In this paper we document that existing teacher incentives in Kenya are indeed weak, with teachers absent from the classroom about 20% of the time. We then report on a randomized evaluation of a program that provided primary school teachers in rural Kenya with incentives to reduce student drop out rates and increase student test scores. Students in program schools had higher test scores on exams linked to the incentives for the duration of the program. An examination of the channels through which this effect took place, however, shows that teacher attendance did not improve, homework assignment did not increase, and pedagogy did not change. Test scores did not increase significantly on exams that were not linked to incentives, and test score gains on exams linked to incentives did not persist after the program ended. Teachers in program schools did conduct more test preparation sessions and students in those schools were more likely to take exams. There is evidence that students in program schools acquired better test taking skills, scoring relatively well by focusing on answering multiple choice questions and, more generally, being more likely to answer all types of questions.

^{*} Department of Applied Economics, University of Minnesota. E-mail: pglewwe@umn.edu.

[♦] The Brattle Group, Washington, DC. E-mail: nauman.ilias@brattle.com.

[•] Department of Economics, Harvard University; The Brookings Institution; Center for Global Development; and NBER. E-mail: mkremer@fas.harvard.edu.

We would like to thank Rachel Glennerster, Ed Kaplan, Janina Matuszeski, and Courtney UMBERGER for very helpful comments and assistance. We are especially grateful to Sylvie Moulin and Robert Namunyu for outstanding work in the field and to Emily Oster for outstanding research assistance in the US. We thank the World Bank and the MacArthur Foundation for financial support.

1. Introduction

Teacher incentive programs are increasing in popularity. In the United States, several teacher incentive programs have been introduced, generally offering annual merit pay equivalent to 10% to 40% of teachers' monthly salaries (American Federation of Teachers, 2000).¹ Under the No Child Left Behind Act, poorly performing schools in the US face sanctions. Israel now offers incentives to teachers based on students' scores (Lavy, 2002b), and World Bank programs in Mexico provide performance incentives to primary school teachers.

In many developing countries, incentives for teachers are very weak. Chaudhury et al., (2006) show that primary school teachers were absent from school 27% of the time in Uganda, 25% in India, 14% in Ecuador, and 11% in Peru. Teachers in our sample have an absent rate of 20%. As in other countries, even when teachers are at the school they are often not in the classroom; our classroom observation data show a teacher absence rate of 27%.

Advocates of incentive pay argue that many teachers face weak incentives, with pay determined almost entirely by educational degrees, training, and experience rather than performance (Harbison and Hanushek, 1992; Hanushek, 1996; Lockheed and Verspoor, 1991). They argue that linking teacher pay to student performance would increase teacher effort.

Opponents of test score-based incentives argue that since educational goals are multi-dimensional and only some aspects are measured by test scores, linking pay to test scores may lead teachers to sacrifice curiosity and creative thinking in order to teach skills tested on standardized exams (Holmstrom and Milgrom, 1991; Hannaway, 1992). Yet teaching in many developing countries is usually by rote, so the risk of reducing efforts to stimulate creativity may be small. On the other hand, if incentive systems are very weak, schools may respond to test

¹ Examples include programs in Rhode Island in 1999; Denver in 1999-2000; and Douglas County, Colorado in 1994 (Olsen, 1999; Education Commission of the States, 2000). In 1999 California offered a one-time award of \$25,000 to teachers in under-performing schools if students showed large gains (Olsen, 1999). In 2007, New York City initiated a plan for teacher bonuses based on student test scores in schools with many low income children.

score-based incentives in ways that are more harmful than teaching to the test. For example, to raise exam scores they could deliberately force students to repeat grades or even drop out.

This paper examines the impact of teacher incentives in rural Kenyan primary schools. We report on a randomized evaluation of a program that rewarded teachers in 50 rural primary schools based on the average test scores of students. To discourage dropouts, students who did not take the test were assigned very low scores. The program awarded prizes valued at up to 43% of the typical monthly salary to teachers in grades 4 to 8 based on schools' average performance on the Kenyan government's district-wide exams. This prize to salary ratio is similar to that used in many US incentive programs.

We consider evidence on program impact through the lens of a Holmstrom-Milgrom style model in which teachers can devote effort to raising students' long-run human capital acquisition or to actions that influence students' short-run test scores but have less impact on human capital.

During the life of the program, students in incentive schools were more likely to take exams and scored higher on exams linked to teacher prizes. Further examination, however, provides little evidence of increased teacher effort that would reduce dropout rates or increase long-run learning. Dropout rates did not fall, teacher attendance and homework assignment did not increase, and pedagogy did not change. Test score gains on exams linked to incentives soon faded after the program ended. Test scores also did not increase significantly on concurrent exams that had a slightly different format and were not linked to the incentives.

There is evidence of changes in teacher behavior that seem oriented toward increasing scores on the formula used to award prizes. These scores were significantly higher in treatment schools. Conditional on enrollment, students in incentive schools were more likely to take tests. Teachers in those schools were more likely to conduct test preparation sessions, which not only review subject matter but also teach test taking strategies and review of old exams. The program appears to have reduced the number of blank answers on multiple choice exams and raised

students' propensity to complete exams. This is consistent with a model in which teachers focus effort on raising short-run scores. While we cannot rule out that the program also raised students' human capital, we find little evidence of it. The efforts we observe seem skewed towards activities with a high signalling component, which is consistent with the lack of any program impact on tests taken after the program ended and on tests with a different format that were not linked to the prizes. It also appears that teachers learned to adjust to the system over time; test preparation sessions increased from the first to the second year of the program, as did the gap between treatment and comparison schools in exam participation and test scores. While these test preparation sessions may contribute to longer-term skill formation, we found no persistent effects on any achievement tests after the program ended.

We can compare the impact of teacher incentives to those of other programs conducted in the same geographic area. Another teacher incentives program that provided bonuses to informal pre-school teachers conditional on low absence rates had no effect because headmasters would record absent teachers as present, even though the school could keep funds not paid to teachers. In contrast, a program providing merit scholarships to students seems to have raised human capital acquisition rather than signaling. This program increased test scores and teacher and student attendance, but not exam preparation sessions. (Kremer, Miguel, and Thornton, 2007)

Several earlier papers examine the impact of linking teacher pay to student test scores. Jacob (2002) explores a Chicago program in which low scoring students were not promoted to the next grade and schools and teachers were put on probation. He finds that the program raised students' scores, though the increases were largest for skills used on the high-stakes exam. Some schools manipulated scores by putting more students into special education. Koretz (2002) finds that a Kentucky teacher incentive program had significant positive impacts (0.5 to 0.6 standard deviations) on the test used to determine teacher rewards but smaller effects (0.1 to 0.2 standard deviations) on another test not tied to the rewards. Eberts, Hollenbeck, and Stone (2002) examine

a Michigan merit pay program designed to reduce the fraction of students who dropped classes after the quarter began. In general, the program did not attain its goals, and in some outcomes (student grades, daily attendance, and course passing rates) student performance declined.

Lavy (2002) finds that rewarding Israeli teachers based on average school (rather than individual teacher) performance raised test scores and matriculation exam participation, but not student performance on matriculation exams. Analyzing an Israeli program based on individual teacher performance, again based on student test scores, Lavy (2004) finds that pass rates on the high school matriculation exam rose by 7 to 18 percentage points for weak students (those below the median score on an earlier exam). At 70% to 600% of a teacher's monthly salary, these prizes were much larger than those in most US teacher incentive programs.

Some researchers find that high-stakes testing can lead teachers or administrators to manipulate test results. Figlio and Winicki (2002) show that Virginia school districts increase calories in school lunches on days when students take high-stakes tests, artificially inflating test scores. Jacob and Levitt (2003) estimate that 4-5% of Chicago elementary school teachers help their pupils cheat, and that cheating increased after the introduction of high-stakes testing.

Muralidharan and Sundararaman (2007) find that a teacher incentive program in Andhra Pradesh, India raised scores on tests that measure conceptual and mechanical skills during the years the incentives were in place. The program also increased prep sessions, but not teacher attendance. Unlike both that study and our study, Duflo, Dupas, and Kremer (2007) find that improved teacher incentives combined lower class sizes increased teacher attendance and test scores in Western Kenya.

This paper differs from earlier work in several ways. First, since advocates and opponents of teacher incentives agree that they can raise test scores, but disagree about whether this effect reflects increased learning or teaching to the test, we measure how teacher incentives affect both test scores and many types of teacher effort. In particular, we examine teacher absence, teachers'

classroom behavior, scores on both exams linked to incentives exams not linked to incentives (which have a somewhat different format), and scores on exams taken after the program ended. Second, since teacher incentive programs may be introduced in areas where teacher performance is relatively weak, it is difficult to identify econometrically these programs' impacts. We address this problem by randomly assigning schools to treatment and comparison groups. Third, while most research on teacher incentives is for the US, we examine a poor developing country.

The rest of the paper is organized as follows. Section 2 presents a Holmstrom-Milgrom style model in which linking bonuses to test scores could either increase teaching effort or divert effort towards teaching to the test. Section 3 discusses primary education in Kenya and argues that high teacher absenteeism suggests that current incentives are weak. Section 4 describes the teacher incentives program and the process that selected schools into it. Sections 5 and 6 report the impact of the program on teacher and student outcomes, respectively. Section 7 concludes.

2. A Model of Productive and Signaling Effort

Holmstrom and Milgrom (1991) consider a model where linking pay to observed signals of performance may lead employees to focus on tasks where output is easily measured and divert effort from tasks where output is hard to measure. They present two examples. First, linking teacher pay to test scores may cause teachers to teach to the test rather than encourage creativity. Second, employees who both produce output and maintain an asset, such as a piece of equipment or a firm's reputation, may neglect the asset if they face strong incentives to increase output.

Our model combines elements of both of these examples, and is a special case of their general model. Suppose teachers exert two types of effort: effort to promote long-run learning and "signaling effort," which raises scores in the short-run but does little for long-term learning. School officials observe only test scores, which depend on both underlying learning (produced by current and past teaching effort) and current signaling effort. Let $T_t = L(e_t, e_{t-1}, e_{t-2} \dots) + \gamma(s_t) +$

ε_t , where T_t is a test score at time t , L is student learning, e_t is teaching effort on long-run learning at time t , s is signaling effort, and ε is a random shock. Thus, teaching effort produces student learning, while signaling effort yields only short-run effects on test scores. Teaching effort is the unobservable effort to maintain assets in the Holmstrom-Milgrom's model.

Assume that teachers' utility is given by $U = M - C(e, s)$ where M is teacher pay and C is a utility cost that depends on both teaching and signaling effort. Note that e and s can be either substitutes or complements. For example, they are substitutes if there is a fixed amount of time per day that must be allocated between them. Yet they can be complements if there is a fixed cost to teachers of attending school at all. Teachers may also care directly about test scores, but adding this to the utility function does not alter our results.

For now, let the decision-maker as an individual teacher. In fact, the program incentives we examine below operate at the school level, yet one can think of the decision in the model as a collective one, made by the teachers and headmaster together. Teachers work closely and interact repeatedly, and although Kenyan headmasters cannot hire and fire teachers, they may be able to use their control of internal teaching assignments to help coordinate teachers on specific courses of action. The school thus may be able to sustain a cooperative solution to a repeated game.

Suppose teacher pay is $M = \alpha + BT$. If $B = 0$, pay is independent of performance. As noted by Holmstrom and Milgrom, $C_1(0,0)$ and $C_2(0,0)$ may be negative. Teachers may care about their students, or enjoy exerting some effort even without performance incentives. If so, they will exert effort in teaching and signaling even if $B = 0$.

If the government or an NGO announces that pay will be linked to test scores for a single year, teachers will change both teaching and signaling effort to satisfy the first order conditions implied by the above model. Specifically, they will exert teaching and signaling effort so that:

$\frac{\partial L}{\partial e} B = \frac{\partial C}{\partial e}(e, s)$ and $\frac{\partial \gamma}{\partial s} B = \frac{\partial C}{\partial s}(e, s)$. If e and s are complements in the utility function, or if

utility is additively separable, then both types of effort will increase. If they are substitutes in the utility function, incentives may increase one type of effort and decrease the other. Thus in this model, incentives can either increase or decrease teaching effort.²

If it were possible to cheaply and accurately monitor individual performance on both tasks, then a wage contract could induce teaching effort without inducing signaling effort. But measuring teaching and signaling effort may be expensive and inaccurate at the individual level, particularly if teachers know that their wages depend on these measures. Yet it may be possible to distinguish teaching and signaling effort empirically.

First, outside observers could directly observe teachers' activities. Some activities, such as teacher attendance or homework assignment, should increase students' human capital, while others probably have a higher signalling component. For example, some Kenyan schools conduct what are known as "preps": extra test preparation or coaching outside of normal class time, often during school vacations. Prep sessions may include some human capital acquisition, but relative to normal classes a higher proportion of effort is aimed at raising test scores, such as reviewing old exams or teaching students not to leave blanks on multiple choice exams.

Second, improved learning should have long-run effects on test scores, while signaling has only a short-run effect.³ Thus, a finding that test score gains disappear after the incentive program ends suggests that the program provided only skills specific to the test at hand, or served only as a short-run reminder to students about test taking techniques.

A third possible way to distinguish efforts to increase real learning from test preparation activities is to examine patterns in any test score increases. Indications of signaling include test

² Clearly, there is a continuum between exerting effort to promote long-term learning and trying to raise short-run test scores. An extreme example of the latter would be cheating at the time of the test; less extreme versions include going over previous years' exams and teaching test-taking strategies such as guessing on multiple choice questions. Within the category of promoting learning, teachers could focus only on the curriculum to be tested or promote learning more broadly. One could generalize this model to allow teachers to choose from a menu of activities, with varying combinations of long-term learning and signaling effort, but the main results would be similar.

scores that increase only on exams linked to incentives, increased scores mostly in subjects prone to memorization, or increased scores on exams with formats subject to coaching (e.g. multiple choice exams, for which students can be coached to guess rather than leave blanks).

Under this model, parents and local communities may support teachers' focus on short-run test preparation, since the distribution of certain rents (such as high-paying jobs or subsidized places in the next level of education) may depend on test scores as well as underlying learning. Thus raising test scores may benefit not just teachers, but also pupils. Yet such test preparation in the model is socially wasteful; it requires teacher effort but does not raise the overall stock of human capital that determines total output for the society.

3. Background

This section describes teacher incentives in Kenya, which are generally scant. In Kenya, decisions regarding hiring, firing, and transferring teachers have long been made centrally by the Ministry of Education. Hiring depends primarily on academic qualifications.⁴ Salary primarily depend on education and experience, and is set by collective bargaining between the government and the politically powerful Kenyan National Union of Teachers. During the project, primary school teachers earned, on average, Ksh 126,921 (US\$ 2162) including allowances (World Bank, 2005).⁵ At about five times the annual GDP per capita, this a very high salary high. There is little scope for performance-based promotion or pay increases. Teachers have strong civil service and union protection and so are difficult to fire. In rare cases, teachers with very poor performance are transferred to undesirable locations, and the government may look more favorably on requests for transfers to desirable postings or home areas from teachers who perform well.

³ Some types of signaling may have a long-run effects on test scores. For example, teaching students to guess on multiple choice exams or better allocate their time could raise scores on other tests and in the long run.

⁴ Most Kenyan primary school teachers have two years of teacher training beyond secondary school. Yet some were hired under an older system where primary teachers had only a grade 7 education and two years of teacher training.

Overall, teachers are offered few incentives by their employer. Parent committees provide some incentives to teachers, but usually these incentives are weak. All schools should have a parent committee, and these committees sometimes provide teachers gifts when schools do well on the national exams. Similarly, parents sometimes demonstrate against teachers who have behaved badly, which pressures the Ministry of Education to arrange a transfer for the teacher. However, only a minority of school committees provide supplemental bonuses, and they usually attempt to influence the national authorities only in extreme situations.

To the extent that teachers face incentives, they are based on the national testing system. Results on the national primary school leaving exam (the KCPE) are front-page news in Kenya; newspapers name the highest-scoring schools. Results from the KCPE and from district government exams administered in the upper primary grades are often posted in headmasters' offices.

The KCPE determines what secondary schools, if any, will accept graduating primary school students, so teachers devote considerable effort to prepare for it, administering "preps" during evenings, weekends, and vacation periods. Teachers in the rural schools covered by this program sometimes receive payments from parents for these sessions, but the amounts are small; a typical student pays only 10-20 Ksh (about 16 to 33 US cents) per school term to attend "preps" held at school, and students who cannot pay are not excluded. Preps are sometimes offered in lower grades, but not often. For example, in the year before the program began (1997), preps were offered in only about one third of grade 4 and grade 5 classes, and not every child in these classes attended. The amount of time spent in preps varies widely, but a typical amount of time is five hours per week during the term, and 25 hours per week in vacation periods (five hours per day for five days). Prep sessions are not limited to areas where students are already doing well; in the year before the program began students without

⁵ This is assuming an exchange rate of 58.7 shillings per dollar, the 1997 dollar-shilling exchange rate.

opportunities to take preps scored almost as high as students who did have the opportunity (means of -0.03 and 0.02 standard deviations, respectively, averaged across grades 4-8).

While the wide attention given to the KCPE results likely spurs effort by some teachers to raise test scores, some of this effort may be undesirable. For example, 7th graders with low scores may be required to repeat grade 7 instead of entering 8th grade and taking the KCPE.

On the whole, aside from those provided by the KCPE, teacher incentives in Kenya are weak. One indicator of this is high teacher absenteeism. Compared to the 6% staff absence rate at a nearby non-profit organization, random visits to comparison schools suggest that teachers were absent about 20% of the time.

High absence rates have two interpretations. First, they may reflect an optimal contract in which teachers supply less than full work for less than full pay. This may be optimal if teachers need to tend their farms at particular times of the year. Alternatively, high absence rates can arise because teachers have political power and so can prevent enforcement by school administrators. We believe that the facts support the second hypothesis. Kenyan teachers have a very strong union, and they are paid about five times the per capita GDP, a high ratio even for poor countries. There is considerable queuing for teaching posts, with many qualified applicants waiting for jobs due to a national hiring freeze.

Moreover, if high teacher absences indicate an implicit contract that allows teachers to pursue other activities, the absences should be scheduled in a predictable way so that students need not come to school on days when teachers expect to be absent, such as peak agricultural times. However, we find few patterns in teacher absences. Indeed, visits to schools often reveal many pupils milling around unsupervised because their teachers are absent. In addition, we find that teachers are often in school but not in their classrooms. Teachers were absent from school 20% of the time in our sample, yet trained observers who visited classes found teachers absent from class 27% of the time. Even when teachers come to class, they usually arrive late; only a

small percentage of teachers were in the classroom at the time the class officially started. Casual observation indicates that teachers often drink tea in the staff room with other teachers during class time. Similarly findings exist in India, where public school teachers are absent from school one-fourth of the time but absent from class about half the time (Kremer, et al., 2004).

Absences seem widely distributed among almost all teachers, rather than concentrated among a subset with very high rates. Figure 1 and Table 1 show the percentage of teachers who were absent zero times out of eight visits, one time out of eight visits, twice out of eight visits, etc. With only a few visits to each school, the dispersion of absence rates in Figure 1 exaggerates the underlying dispersion of attendance probabilities among teachers. For example, even if all teachers had the exact same absence probability, the empirical distribution of absence rates in a few visits will show some dispersion. To correct for this, and better assess the distribution of teacher absences, we calibrate two models of absences, a non-parametric model that assumes there are five types of teachers, with different probabilities of attendance, and a model in which each teacher's attendance probability is drawn from a beta distribution.⁶ As seen in Table 1 and Figure 1, both models suggest that while there are a few teachers who are rarely present, most absences are from teachers who attend between 50% and 80% of the time and a large share of absences are from those who attend *more* than 80% of the time. The absence rate of the median teacher is 19% in the five group model and 14% in the beta distribution model.

A regression of student test scores on teacher attendance and test preparation sessions in 1998 suggests that the marginal product of test preparation sessions, which can be regarded as having a high share of signaling effort, may be much higher than that of teacher attendance, a plausible index of teacher effort.⁷ Teachers who attend school 20 percentage points more have

⁶ For technical details, see Glewwe, Ilias, and Kremer (2003). We thank Emily Oster for calibrating these models.

⁷ Data on teacher attendance and test scores are described below.

students who score 0.012 standard deviations higher (standard error 0.012).⁸ We do not have data on exactly how long teachers spent in preps; we only know whether they taught during each of the three vacation periods or outside of normal school hours during each of the three terms. Still, teachers who report coaching in one additional vacation period have pupils who score 0.044 standard deviations higher (standard error 0.009). Based on discussions with teachers about the number of days spent coaching per vacation period, it seems likely that, if interpreted causally, these point estimates imply that the marginal test score effect of a day of coaching is an order of magnitude greater than that of a day of (teacher) school attendance. Of course, these results should be interpreted with caution since we do not have random variation in measures of teacher effort either in building students' learning or in signaling.

The claim that test preparation sessions raise test scores is supported by US evidence on the effect of commercial test preparation. Even though most US admission tests are designed to measure aptitude rather than achievement, and so should be hard to study for, studies often show gains of 0.15 to 0.40 standard deviations from test preparation (e.g. Bangert-Drowns et al., 1983).

In an environment where teacher pay is not linked to test scores, teachers may make only limited efforts at test preparation, leaving the marginal product of test preparation on test scores much higher than that of teacher attendance. Denote e_0 and s_0 as the levels of e and s when $B=0$. It is plausible that teachers' utility from exerting teaching effort may exceed their utility from signaling (although signaling may have a benefit if it enhances teachers' reputations). If this is true, then $\frac{\partial L}{\partial e}(e_0, s_0) < \frac{\partial \gamma}{\partial s}(e_0, s_0)$, so it is easier for teachers to raise student test scores by exerting signaling effort than by exerting effort that promotes long-run learning.

4. Program Description

⁸ These estimates may have attenuation bias as teacher absence is from random visits and thus measured with error.

International Child Support (ICS), a Dutch NGO, offered schools in Busia and Teso districts of Western Kenya the opportunity to participate in a program that provided gifts to teachers and headmasters whose students performed well. The program provided in-kind prizes, rather than cash, because this was seen as more politically acceptable in the Kenyan context. Anecdotal evidence suggests that teachers valued the prizes. They were given to teachers in grades 4 to 8 based on school performance on the annual government (district) exams. All teachers who taught these grades were eligible for prizes. Teachers of lower grades were ineligible, because no district-wide exams existed for those grades.⁹

To encourage cooperation among teachers within schools, and avoid creating incentives for teachers to sabotage each other's work, ICS prizes were based on the average performance of all grade 4 to 8 pupils in each school, with each subject weighted equally, rather than on a teacher-by-teacher basis. Education experts are more sympathetic to this type of incentive (Richards and Sheu, 1992; Hanushek, 1996).

To create incentives for headmasters, and encourage further cooperation among teachers at the school level, each winning school also received a briefcase for the headmaster, a wall clock, a time keeping clock, and a bell. A potential disadvantage of setting prizes at the school level is that doing so may allow free-riding by teachers. However, teachers are in a repeated game with each other and can observe each others' attendance, for example, at high frequency. The typical school in the sample had only 200 students and 12 teachers, of whom half taught in the upper grades, so coordination within the school seems feasible. Moreover, headmasters have ways to help teachers reach a cooperative solution; for example, they can choose a deputy from among the teachers.

ICS prizes were for two types of performance: top-scoring schools and most-improved schools. Schools could win in only one type, and prizes were the same for each type.

⁹ They received a lantern as a token prize, whether or not they belonged to a winning school.

Improvements were calculated relative to performance in a baseline year. Since results on government exams were not available for 1997, the 1996 scores were used as the base for measuring improvements. (Henceforth, we call the last pre-program year for which we have data year 0,¹⁰ the first (1998) and second (1999) years of the program are years 1 and 2 respectively, and the post-program year (2000) is year 3.) In each category, three first prizes (a suit, worth about \$51); three second prizes (plates, glasses, and cutlery, about \$43); three third prizes (a tea set, about \$34); and three fourth prizes (bed linens and a blanket, about \$26) were awarded. Overall, 24 of the 50 schools participating in the program received prizes, and teachers in most schools should have felt that they had a chance to win a prize.¹¹ The prizes ranged in value from 21-43% of typical teacher monthly salaries and are comparable to other merit pay programs.¹²

The ICS incentives were designed not only to raise test scores but also to reduce dropout rates. All students who were enrolled at the start of the program were included in the calculation of scores. Students who did not take the test were assigned very low scores: students who did not take the English essay exam were assigned a score of zero, and students who did not take the multiple choice tests in other subjects were given scores of 15, lower than the average guessing score of 25. To discourage schools from recruiting strong students to take the exams, only students enrolled in February 1998 were included in calculations of the school mean score.

In February 1998, 50 schools were given the opportunity to participate in the program, and all accepted. When ICS first announced the program, schools were told that ICS could

¹⁰ This is either 1996 or 1997, depending on the type of data.

¹¹ Since Busia and Teso districts had separate district exams, prizes were offered separately in each district in proportion to the number of schools in the district.

¹² For example, the 1993-94 Dallas merit pay program, which was also based on school-wide performance, awarded \$1000 annual bonuses, equivalent to 39% of the average monthly salary of Texas teachers, and presumably a lower percentage of salaries for teachers in Dallas (Clotfelter and Ladd, 1996; American Federation of Teachers, 2000). Similarly, a 1999 Rhode Island program awarded \$1,000 annual bonuses, worth about 25% of the average monthly salary (Olsen, 1999; American Federation of Teachers, 2000). Programs in Colorado awarded from 10-50% of a teacher's monthly salary in merit-based annual bonuses (Education Commission of the States, 2001). In Israel, the annual bonuses examined by Lavy (2002) ranged from 10% to 40% of the average teacher's monthly salary.

commit to only one year of funding, yet favorable field reports convinced the NGO to extend the program for a second year. Prizes were awarded at a ceremony for all schools in the program that was held at the end of the Kenyan school year in November.

Overall, the context seems very favorable for a teacher incentives program: the level of teacher absence suggests that teacher effort was an issue in these schools; standardized curricula and the prevalence of rote teaching implied little scope for diverting teacher effort from creativity and towards teaching to the test; and the program's short duration allowed for a design that was less amenable to manipulation of the student body or the set of teachers in the school.

The 50 schools invited to participate in the program were randomly selected from 100 schools designated by the Ministry of Education as particularly in need of assistance but which did not participate in an earlier World Bank program that had given textbooks to the most needy schools in the area. These 100 schools scored somewhat below the district average before ICS began to assist them. ICS had also provided textbooks or modest grants to 75 of these schools before or during the teacher incentive program as determined by random assignment into four groups of 25 schools. In each of these groups, half of the schools were randomly selected for the incentive program. By construction, the schools selected and not selected for the incentive program were divided in the same proportions across Busia and Teso districts, by geographic divisions within districts, and by whether they received textbooks or grants in earlier years. It is unlikely that receipt of other ICS assistance by 75 of the schools seriously compromises external validity, since previous ICS assistance was very small relative to overall school budgets, and the previous programs had only a modest impact on the schools (see, *inter alia*, Glewwe, Kremer, and Moulin, 2008). Moreover, many NGOs assist schools in Kenya, and while these 100 schools received more support than average, they were not in the upper tail of the distribution.

The 50 schools in the comparison group for this evaluation participated in another program that provided pre-school teachers with training, materials, and higher salaries (the last

conditional on teacher attendance). Unlike primary school teachers, pre-school teachers are semi-volunteers who typically have no formal training and are hired and paid by parents' committees, not by the Ministry of Education. Their salaries come from parent contributions, which are often irregular. The pre-school program had little effect on pre-school pupils' performance, so it seems unlikely that it affected outcomes in grades 4 to 8 in the time period we examine. Resources from that program went to pre-school teachers or supplies that pre-schools would not have bought in the absence of the program and thus could not have leaked in a substantial way to grades 4-8. Moreover, pre-schools are locally-run and administratively separately from the associated Ministry of Education-operated schools, with separate financial accounts.

Interviews with the headmaster and three teachers in each treatment school in year 2 suggest that they liked the program. All teachers interviewed supported the idea of motivating teachers by given them incentives; 83% of the teachers said that prizes were justly awarded; 75% reported increased homework assignment due to the program (yet our data contradict this, as noted below); 67% reported more cooperation among teachers; and 88% reported more preps.

To discourage teachers from trying to move into treatment schools to benefit from the program, eligibility was limited to teachers employed (in any grade) in program schools as of March 1998. Exit and entry of teachers were not significantly different between program and comparison schools (Table 2),¹³ and there is no evidence of differential reassignment of teachers within schools to different grades. In treatment schools 7.4% of teachers transferred from a non-incentive to an incentive grade during the program; in comparison schools, 7.3% did.

5. Impact of Incentives on Teachers

This section examines the impact of the program on teacher attendance, homework assignment, pedagogy, and teachers' propensity to offer prep sessions.

5.1 Teacher Attendance. Teacher attendance was not affected by the incentive program.

In the year prior to the start of the program, all 100 schools were subject to two random, unannounced visits where the present/absent status of each teacher in grades 4 to 8 was recorded. Similar visits were made five times in year 1 and three times in year 2.¹⁴ In each year for each teacher, an attendance rate was defined as the proportion of visits during which the teacher was present. Teachers were counted as present if they were at school, even if they were not teaching during the visit. Following standard intention-to-treat (ITT) methodology, the sample included only those teachers assigned to program or comparison schools in year 0; teachers who changed schools from year 0 and year 1 or from year 1 and year 2 were assigned to their initial schools.¹⁵

Before the program, the schools later selected to be program schools had slightly higher teacher attendance, but the difference was insignificant (Table 3, Column 1).¹⁶ In year 1 of the program, teacher attendance was slightly lower in the incentive schools, and in year 2 the attendance was slightly higher (Table 3, Columns 2 and 3), yet both coefficients are completely insignificant.¹⁷ Difference-in-difference estimates (columns 4 and 5) are also insignificant.

5.2 Pedagogy. There is no evidence that the program affected the presence or behavior of teachers in the classroom. Trained observers watched each teacher once per year, spending one class period recording several measures of teacher behavior, both objective information on teacher activities and subjective impressions of their energy level and caring for students.

¹³ All regressions in this paper allow for school-level random effects, as explained below.

¹⁴ Some visits did not happen, for example due to vehicle breakdowns; 1.44 visits were made to the average school in year 0, 4.78 in year 1, and 2.95 in year 2. We focus on teacher absence data from school visits, not from official school logs, because the latter are often blank. Yet school-log data also show no program effect on teacher absences.

¹⁵ This could be done only for those teachers who switched schools and remained in the sample of 100 schools. Since there are no data on the teachers who switched to other schools, they were dropped from the analysis.

¹⁶ The results here are robust to a specification where each visit is treated as a binary opportunity for attendance and month of visit is controlled for. The samples in Table 3 are smaller than those in Table 2 because teacher attendance data exist only for teachers in upper grades.

¹⁷ Results are similar when lower primary school teachers are used as a control, i.e. attendance of all the teachers is regressed on a program dummy, a dummy indicating a teacher is an upper primary teacher, and an interaction term.

Prior to the program, there was no significant difference in any of the pedagogy measures between the incentive and comparison schools (Table 4, Column 1). During the program, there was no difference between treatment and comparison schools in the rate teachers were present in class. This implies that the incentive program did not even lead the teachers who were at school but not in class to come to class, a low-cost type of effort; this suggests little impact of the program on human capital acquisition.

We also find no significant difference during the program period (years 1 and 2) between the two school groups in any pedagogical practices (Table 4, Columns 2 and 3). We examined many pedagogy measures but present results for only two objective measures (teaching aids and use of blackboards) and two subjective ones (teacher caring and energy). Point estimates are near zero for all measures. The difference-in-difference estimates in the last two columns are school-level regressions since individual teachers cannot be matched across years. These estimates are also close to zero and far from statistically significant for each pedagogical practice.

There is also no evidence that the program raised homework assignment. In grades 4 to 8, information was collected for each school from a random subset of students on whether they were assigned homework the previous day. Homework assignment was more common in higher grades. For example, in year 0, 14% of grade 4 students and 45% of grade 8 students report being assigned homework the previous day. Treatment schools assigned slightly more homework than control schools prior to the program, but the difference is completely insignificant (Table 4, Column 1). During the program, treatment schools assigned slightly less homework, yet the gap was never significant at the 5% level in either levels or differences (Columns 2, 3, 4, and 5).

5.3 Test Preparation Sessions. In contrast to no program impact on teacher attendance, pedagogy, and homework assignment, the program did lead program school teachers to offer more preps than did teachers in comparison schools. School headmasters provided information

on the existence of preps in grades 4 through 8 in six time periods: each of the three terms (outside of normal school hours) and the three vacation periods (April, August, and December).

Prior to the program, incentive schools were slightly less likely to offer preps (Table 5, Column 1), but after the program started, they started to conduct more preps (Columns 2 and 3). They were 4.2 percentage points more likely to conduct preps in year 1 and 7.4 percentage points more likely in year 2, and the latter estimate is significant at the 5% level. Differenced results (Columns 4 and 5) are similar but less precisely estimated. These results are driven primarily by preps over vacations, as seen by the stronger results in the lower panel of Table 5.

6. The Impact of Incentives on Student Outcomes

. The program increased scores on the formula used to award prizes, but this impact seems not to be due primarily to changes in student drop out rates or long-run learning. While students were more likely to take the government exams, and schools thus avoided the score reduction penalty for students who did not participate in these exams, this was due to higher proportions of enrolled students taking the government tests, not lower dropout rates. Moreover, while scores increased for the tests on which incentives were based (government tests), gains were largely absent on tests that were not tied to incentives, which used a somewhat different format (NGO tests). . Test score gains on the government exams did not persist beyond the end of the program, consistent with the hypothesis that the program mainly increased incentives for short-term signaling. Indeed there is evidence the program led to improved test-taking skills; performance improved more on multiple choice questions than on fill-in-the-blank questions, and students more likely to answer all questions on the exams.

Subsection 6.1 reviews the exams taken by students and the econometric approach used to analyze the data. Subsection 6.2 presents estimate of program impacts on dropout rates,

repetition, and exam participation. Section 6.3 examines test scores as well as scores on the formula used to reward teachers.

6.1 The Exams. Incentives for teachers were based on their students' performance on the government exams taken by students in grades 4 through 8 in October. Participation is incomplete since students had to pay a fee of 120 Ksh (US\$ 2) to participate. Since students who did not take the government exams were assigned lower scores on the formula used to determine teacher rewards than students could obtain by guessing, teachers in program schools had an incentive to encourage their students to take the government exams.

There are no government exam scores for the 32 schools in Teso district for year 1 because that district chose not to offer exams that year due to the cost to parents. (Program schools in Teso district in year 1 did administer exams for the purpose of awarding prizes, but no government exams were administered in the control schools of that district). Consequently, analysis of the government exam scores for year 1 is restricted to schools in Busia district, which comprised 68 of the 100 schools.

In order to provide an independent source of evidence on the impact of the program that is not directly tied to teacher incentives and to address the substantial attrition issues that would be created by analyzing only data from the government exams, the NGO administered its own test. Hence students were also given exams prepared by ICS that were not tied to the teacher incentive program and thus provide an independent assessment of the impact of the program. Note also that the scores on the government exams were quite low, while the NGO (ICS) exams were designed to detect performance differences among a wider range of students and thus were easier for the typical student. In years 0 and 1, separate NGO exams were given for each grade and the exams had a multiple-choice format. However, to facilitate comparisons across grades, given high levels of repetition the NGO exams in year 2 were "multilevel," with the same test

given to all students in grades 3 through 8. Easy questions at the beginning of the tests could be answered by all students, including those in grade 3, while subsequent questions became progressively harder. Moreover, in year 2, most of the NGO exam questions were “fill in the blank,” as opposed to multiple choice. We have information on the NGO tests for all 100 schools for years 0, 1, and 2. NGO exams were administered to students in grades 3 through 8.¹⁸

In Kenya as in many other countries, exams are supervised by monitors from outside the school, typically teachers from other schools. Because of this, cheating was not a large concern, although one case was discovered in year 1 (1988) where the headmaster of a program school colluded with the teachers assigned to supervise the school to allow cheating on the government exam. That school was disqualified from the competition in year 1 but was allowed to participate in year 2. The scores from that school were not included in the analysis in year 1, but its scores were included in year 2.

Finally, we also have data on the KCPE exam, which is given in grade 8. Technically, prizes were not awarded based on this test, so it could be considered as an indicator of learning not tied to the prizes. Yet the format is very similar to the government exam, so scores on the KCPE could be influenced by actions designed to increase the scores of students on the district mocks. Because of this ambiguity, we do not analyze the results in the rest of this paper; the results for the KCPE, which are often similar to those for the government tests, are available from the authors.

We examine differences in test scores between the incentive and comparison schools using a random effects regression framework that allows for the possibility that scores of students in the same grade and same school might be correlated due to unobserved characteristics of teachers and headmasters. In particular, we use an error components econometric model with

¹⁸ In years 0 and 1, ICS administered tests in English, Math, and Science. In year 2, ICS administered only English and Math tests. These are a subset of the subjects covered in the government exams. For a description of the NGO

school random effects, subject random effects within each school, and grade random effects within each subject within each school:

$$(1) \quad t_{ijks} = \alpha_{4k}D_{4i} + \alpha_{5k}D_{5i} + \dots + \alpha_{8k}D_{8i} + \beta_k p_s + u_{ks} + v_{jks} + e_{ijks}$$

where k = English, Math, Science, Swahili, Geography/History/Christian Religion (GHCR), Arts/Crafts/Music (ACM), and Home Science and Business Education (HSBE). Equation (1) combines data from several grades to measure the impact of the incentive program for a given subject. The test score of student i in grade j in subject k in school s is t_{ijks} . The dummy variables D_{ji} indicate whether child i is in grade j . The variable p_s is a dummy variable that equals 1 if school s is an incentive school (i.e. a school that was eligible for teacher incentives) and 0 otherwise. Thus, if the impact of the incentive program varies across grades, β_k will measure the (weighted) average impact of the program across all grades. The error term contains three components, the school-specific error term (for subject k), u_{ks} , a grade-specific term conditional on being in that school, v_{jks} , and a child specific term, e_{ijks} .

We estimate these equations using Generalized Least Squares (GLS) without imposing a specific distribution (e.g. the normal distribution) on the error terms. The regressions also include controls for sex and geographic division within Busia and Teso. Given the prospective design of the program, regressions without such controls are consistent, but adding controls to the regression may increase the precision of the estimates. As a check, we ran regressions without the controls for region and sex; they yield similar results.

Following standard intention-to-treat (ITT) methodology, we examine only those students who were enrolled as of February 1998 (year 1) and assign the few students who switched schools during the program to their original schools.

Because the units in which test scores are measured are arbitrary, for each year and each subject and grade combination we normalize all test scores by subtracting the mean test score in the comparison schools and then dividing by the corresponding standard deviation for those schools. Thus, a student with a normalized score of 0.1 was 0.1 standard deviations above the mean score in the comparison schools. For reference, note that for a normal distribution an increase of 0.1 standard deviations would move a student from the 50th percentile to the 54th.¹⁹

We estimate “difference-in-differences” estimates of the impact of the program by replacing the t_{ijks} in equation (1) with that test score minus a “pre-program” test score. The benefit of this specification is that it may reduce the standard error of the estimated impact of the program; since the program was randomized, it is not needed to reduce bias.

Government exams were given to students based on the grade they were in at the time, rather than the grade they would have been in if they had not repeated since the start of the program. In order to facilitate comparisons between repeaters and non-repeaters the same test was given to students in adjacent grades outside the sample schools. Students generally scored about a standard deviation lower on exams designed for the next grade. We therefore use a correction factor of one standard deviation to compute scores that repeaters would have received had they advanced to the next grade. Note also that there were no significant differences in dropout and repetition rates across the incentive and comparison schools.

6.2 Dropout, Repetition, and Exam Participation.

As discussed above, the formula used to reward teachers was designed to discourage dropping out and grade repetition by rewarding teachers in part based on the proportion of

¹⁹ Since the exams differed in Busia and Teso districts, these normalizations were done separately for each district. (The test score regressions in this paper were also estimated separately for Busia and Teso; in only one out of 14 regressions did the program’s impact differ across the two districts, and this difference was significant only at the 10% level.)

originally enrolled students who took the government exams for their grade. However, dropout rates were very similar in the treatment and comparison schools (Table 6, Columns 1 and 3), and the repetition rate was insignificantly lower in incentive schools (Table 6, Columns 2 and 4).

Exam participation rates were higher in program schools than in comparison schools for the government exams (on which the incentives were based), but this was not the case for the (non-incentive) NGO exams. Baseline participation (year 0) was around 80% on the government exam and around 85% on the NGO exam, with no statistically significant difference in the participation rates across treatment and comparison schools. In year 1, the point estimate indicates that participation in the government exams was 5.7 percentage points higher in the incentive schools, although this is not statistically significant (Table 7, Column 2). By year 2, participation was 11.2 percentage points higher in the incentive schools, a difference significant at the 5% level. (The main differences between incentive and treatment schools in exam participation were in grades 4 through 7; participation in grade 8 on the government exam was already close to 90% prior to the program.) In the post-program year, when there was no longer an incentive to encourage students to take the test, the participation rate was actually 1.9 percentage points lower in the incentive schools than in the comparison schools, though the difference was insignificant. In contrast, the participation rates in the NGO exams, which were not linked to teacher incentives, were similar between the two school groups in all years (Table 7, Panels B and C).

The sample size drops significantly from year to year, primarily because over time more students complete their primary education. The sample is limited to students who were enrolled in grades 4 through 8 in February 1998 (year 1) and did not graduate or drop out. To illustrate, 14,982 students were enrolled in grades 4-8 in year 1 (Table 7, Panel C, Column 2). Of these, approximately 10% finished grade 8, either going on to secondary school or finishing their education after completing the primary level; 12% dropped out before finishing grade 8; and 7%

transferred to schools outside the 100 school sample. This leaves 10,654 students for the year 2 regression.

. Theoretically, efforts by treatment school teachers to increase exam participation could bias scores either upwards or downwards, but looking at pre-test scores of subsequent government test takers in the treatment and comparison groups does not suggest substantial bias. If teachers in the treatment schools put equal effort into encouraging all students who would not otherwise have taken the exam to do so, then the addition of marginal students would likely have reduced average test scores, since academically weaker students are less likely to pay the fee to take the government exam. But if teachers selectively chose to concentrate on convincing potentially high-scoring students and their parents of the exam's importance, then average scores in the treatment schools could be biased upward. To get a sense of the potential bias, we compared pre-test (year 0) scores of the students in treatment and comparison schools who were eligible to take the government exams in years 1 and 2. Of the 11,122 students eligible to participate in the government exams in year 1, the mean of the normalized pre-test score of the treatment group students was 0.022 standard deviations higher than that of the comparison group students, but this difference is not statistically significant (t-statistic of 0.81). For the 7,259 students eligible for the year 2 government exams (and who were not repeaters), the mean of the year 0 score of the treatment group was 0.032 standard deviations higher than that of the comparison group, but again this was not significant (t-statistic of 0.93).

6.3 Test Scores. This subsection first demonstrates that by the second year of the program, students in program schools had higher scores on the exams that were tied to the incentives. Yet differences in test scores across the program and comparison schools were smaller and statistically insignificant on the exams that were not linked to incentives. The test scores gains did not persist after the program ended. The pattern of test score gains is consistent with the hypothesis of improved test-taking techniques. Finally, we show that students in

program schools had much higher scores on the formula used to reward teachers, which assigned very low scores to students who did not take the government tests.

There is no significant difference in pre-program scores on the government exam between incentive and comparison schools (Table 8, Panel A, Column 1), and the same is true for the NGO exams (Panel B, Column 1). Since grade 2 students were not given government exams in 1996, we used 1997 NGO tests, where available, as pre-tests for the government exams for students who were in grade 4 in 1998.

The difference in test scores between treatment and comparison schools, and the difference-in-difference estimator of the effect of the program are shown in Table 8 (Columns 2 - 7) and Table 9. The difference-in-difference estimates use a restricted sample, i.e. those students who took exams in at least one subject in the pre-program year and in at least one of the intervention years.²⁰ As discussed, we restrict attention to those students who were enrolled prior to the announcement of the program in February 1998 (year 1). In Table 8, we restrict the sample to those students who did not repeat or drop out in any year. In Table 9, repeaters are assigned to the grade in which they would have been enrolled had they not repeated and their scores are adjusted by subtracting one from their normalized test scores.

Excluding repeaters, the difference estimate for the government exam is insignificantly negative in the first year of the program (point estimate -0.04 , Table 8, Panel A), but this could potentially be due to the differential exam participation between treatment and comparison schools on the government exams. The difference-in-difference estimate, which should be less subject to bias from attrition, is positive, although not significantly so. Both the difference and

²⁰ Note that the sample sizes in Table 8 are related to those in Table 7. For example, there were 11,122 students enrolled in Busia District in year 1 (Table 7, Panel A, Column 2) of which 7,799 took at least one government exam. Since in the regressions for Table 8, Panel A there can be up to 7 observations per student (one for each subject), the actual sample was 50,842. Similarly, 14,982 students were enrolled in the two districts combined in year 1 (Table 7, Panel B, Column 2) of which 13,339 took at least one NGO test. With up to 3 observations per student in Table 8 (Panel B), the actual sample size was 39,510. Table 8 was re-estimated using a consistent sample (only children with test scores in years 0, 1 and 2), the results were very similar to those shown here.

difference-in-difference estimate of the treatment effect in year 2 are significantly positive at 0.136 and 0.139, respectively. Separate estimates by sex (not shown in Table 8) indicate that the impact in year 2 is stronger for boys than for girls. In particular, the difference estimate for girls is 0.107 (and statistically insignificant) while the effect for boys is 0.159 (and statistically significant), and the difference in difference estimate for girls is 0.061 (insignificant) while the effect for boys is 0.199 (significant).

Including repeaters weakens significance in year 2, consistent with the hypothesis that teachers did not find it worthwhile to exert effort on repeaters, since they were going to be assigned the same low score for these students, regardless of how well they performed on the exam.

For the NGO exams, differences are never significant. Point estimates are around 0.09 in levels (Tables 8 and 9, Panel B Columns 2 and 3) and close to zero in difference-in-difference estimation (Table 8 and 9, Panel B, Columns 5 and 6).

Summarizing the overall pattern of difference-in-difference results in Table 8, test scores improved the most on the government exam, while they did not increase as much, if at all, on the NGO test. This suggests that some of the test preparation activities may have been aimed specifically at the government exams, on which the teacher incentives were based, and did not carry over to the NGO exam, with its different format that included fill-in-the-blank questions

There is little or no evidence that gains in test scores persisted to year 3, after the program ended. Estimates of the program effect in year 3 on government exams are slightly negative (although not significantly so) or close to zero when repeaters are excluded from the analysis (see Table 8, columns 4 and 7), and when repeaters are included (see Table 9, column 4) the level point estimate suggests that post-program test scores are actually lower in the treatment schools, although the difference is significant only at the 10% level. There is thus no evidence of

an increase in students' underlying long-term learning. Although learning may depreciate over time (see Andrabi et al, 2008), the rate of depreciation would have to be close to 100 percent to drive down any post-program gains to zero. The test score data thus seem consistent with the hypothesis that teachers were exerting signaling effort that led to higher test scores on the incentive exam during the program years with little or no effect on longer-run learning.²¹

There is also some evidence that the program effects were greatest in subjects most prone to memorization. The average effect for the two program years was strongest for the subject test on GHCR (Appendix Table A.1 in the appendix). In year 1, the difference-in-difference estimate on GHCR was 0.205 for the government exam, which is significant at the 10% level. In year 2, the program impact on GHCR scores was even stronger, with the difference-in-difference estimates being 0.341, significant at the 5% level. The next largest effects were for science and math, with no significant effect for other subjects. Arguably, GHCR is the subject with most memorization and thus is particularly susceptible to extra-coaching and short-run teaching strategies. Primary school science also involves a fair amount of memorization, but math presumably requires less memorization.²²

There is also direct evidence that students' test taking techniques were improved by the program. As noted above, in prep sessions teachers taught students not to leave blanks on multiple choice questions and to reach the end of exams. For the 1999 NGO test, we have item-level data on whether students had correct answers to individual questions in their English and math tests. Although the English test has mostly a fill-in-the-blank format, there are some multiple choice questions (all math test questions were fill-in-the-blank questions). The test began with 20 relatively easy multiple choice questions, followed by 74 questions that become

²¹ Data for the NGO exams in year 3 are available for only 27 of the 100 schools—those that participated in a de-worming project in that year. Point estimates (not shown in Tables 8 or 9) are positive, but none of the t-statistics exceeds one.

progressively more difficult. Of the 74 more difficult questions, 70 are fill-in-the-blank and 4 are multiple choice. The item by item data on the 1999 NGO test do not indicate whether a student answered a question, but of course a student who does not answer a question will receive a score of zero.

To investigate whether the program improved students' test taking skills, we constructed two variables, one indicating the percentage of the four relatively hard multiple choice questions that were answered correctly, and the other indicating the percentage of the 70 relatively hard fill-in-the-blank questions answered correctly. A random effects probit regression of whether any of the multiple choice questions were answered correctly on the program dummy variable and the sex, grade, and division dummy variables produces a positive program impact that is significant at the 1% level (Table 10, col. 1). Moreover, regressing the ratio of the percentage of multiple choice questions answered correctly to the percentage of fill in the blank questions answered correctly on the program dummy and the same set of control variables yields a positive program impact of 0.042 points that is statistically significant at the 5% level (Table 10, col. 2). Students in incentive schools were also less likely to get none of the last 10 answers correct, and none of the last 20 questions correct; these results are significant at the 10% level (Table 10, cols. 3 and 4), which suggests that they learned to answer all questions, guessing on those for which they did not know the answer. Finally, analysis of school level data on leaving questions blank (based on revisiting schools and examining randomly drawn exams) suggests a reduction in leaving answers blank, although this was significant only at the 10% level (Table 10, col. 5). Together, these results suggest that the extra prep classes in the program schools increased students' test taking skills, such as how best to answer multiple choice questions, and that these skills had a small impact on the NGO exams (which were not the sort of test that prep classes

²² ICS staff members familiar with the curriculum suggested that G.H.CR and HS.BE. require the most memorization; science requires a medium amount of memorization; and English, Math, and Swahili require the least

were designed to teach to), and only on those questions that had a format similar to those on the government tests.²³

Evidence that teachers coached students on test taking skills and encouraged more children to take government exams suggests that teachers respond to incentives. Perhaps the strongest evidence of this is seen by analyzing the impact of the program on scores on the formula that determined rewards to teachers (described above in Section 4). Recall that this formula assigned low scores to students who dropped out or did not take the exam. The program increased scores on this formula by a very highly-significant 0.145 standard deviations (Table 9, column 8). This shows that teachers strongly responded to incentives, in particular by ensuring that more students took the test. However, the program impact on long-run student learning is still marginal at best. However, we do not find evidence that the program improved long run learning although we cannot fully exclude this possibility.

7. Conclusion

Schools randomly selected to participate in a teacher incentive program scored considerably higher on the formula used to determine teacher rewards. Students in these schools were more likely to take exams and had higher test scores in the short run on the exams linked to incentives. There is little evidence, however, that teachers responded to the program by taking steps to reduce dropouts or making efforts to increase long-run learning. Teachers in program schools had neither higher attendance rates nor higher homework assignment rates. Pedagogy and student dropout rates were similar in program and comparison schools. Instead, teachers in

²³ Cheating appears not to have been a concern on the tests administered in this study. Analysis of item responses to detect cheating using techniques developed by Jacob and Levitt (2003) provides little evidence of suspicious strings of questions for which all students in the class got the question right. There was one instance of cheating discovered at a program school. If cheating were widespread, however, we would expect to see much larger test score differentials on the government exam (on which incentives were based) than on other exams. The similar program impact on the government exams and on the heavily-monitored KCPE exams suggest that cheating was not the main source of the program effect.

program schools increased test preparation sessions and encouraged students enrolled in school to take the test. Test scores did not differ significantly on exams that were not linked to incentives. There is some evidence that the gain in test scores came from multiple choice questions and from the questions at the end of exams, consistent with the hypothesis that students were coached not to leave blanks on multiple choice exams and to finish exams. Following the end of the program, the test score difference, even on tests linked to incentives, between students in treatment and comparison schools disappeared, consistent with a model in which the program led teachers to focus effort on short-run scores rather than to concentrate on long-run training. The program would have rewarded teachers for reducing dropout rates, since pupils who did not take tests were assigned low scores for the purposes of calculating teacher incentives. However, dropout rates were no lower in program schools. Conditional on enrollment, students were more likely to take exams linked to incentives, consistent with the possibility that program school teachers encouraged students to take exams.

There is evidence that teachers learned over time how to take advantage of the program. Estimated differences in preparation sessions between treatment and comparison schools grew between the first and second year. Anecdotal evidence from the first year's prize award ceremonies suggests that prior to these ceremonies some teachers did not fully understand that having students drop out or not take the test would reduce their chances of receiving a prize. After this experience, differences in exam participation rates between program and comparison schools rose, presumably because teachers worked harder to persuade students to take the exam. Moreover, the test score gap between treatment and comparison schools was greater in the second year than in the first year.

The interpretation of our results does have several caveats. First, we cannot rule out the possibility that a larger incentive program or teacher-specific incentives would have increased efforts to improve underlying learning rather than simply increasing test preparation. However,

at up to 40% of monthly income, the incentives were comparable in magnitude to those in most US programs and in the Israeli program analyzed by Lavy (2002). Although the bonuses were a small percentage of yearly salary and thus the implied increase in daily wages was modest, if teachers chose attendance optimally prior to the program given their intrinsic motivation to teach, the other incentives implicit in the system, and their value of time in other activities, they should have been indifferent at the margin to small changes in attendance, and hence modest incentives could potentially have had a substantial effect. Indeed, the incentives in the program were sufficient to induce teachers to change their behavior—they simply did so in ways that may not have been consistent with the intentions of the program. Moreover, while larger incentives might induce more effort by teachers, they could also have induced more counter-productive signaling effort, for example through cheating on tests or forcing weak students to drop out. The Holmstrom-Milgrom multi-tasking model also suggests another disadvantage of stronger incentives: they would force teachers to bear more risk. Some also argue that individual-level incentives for teachers could potentially undermine cooperation within the school.

A second caveat is that incentives may work as much by encouraging people who will be good teachers to enter the profession as by eliciting higher effort from those who would become teachers in any case. However, given the queuing for teaching positions in Kenya, it is unlikely that people who have either teaching jobs or the academic qualifications to enter teacher training colleges (but not universities) are opting out of the profession in the current system. Any effect on this margin in Kenya, and other developing countries with queues for teaching jobs, is therefore likely to be small.

Third, the program was explicitly temporary. If teachers expected the program to continue indefinitely, and if they expected to remain at the schools for many years, they may

have had more incentives to make long-run investments in learning.²⁴ On the other hand, because the program was temporary it was possible to base incentives on improvements over baseline performance, to incorporate incentives to prevent students from dropping out, and to restrict the program to teachers already in school and thus to avoid strengthening incentives for teachers to seek transfers to schools with pupils from more advantaged backgrounds. A program without these features would be much less attractive since it would be difficult to provide incentives to teachers in weak schools, to prevent teachers from trying to influence the pool of pupils entering their school, or to avoid increasing incentives for good teachers to try to transfer to the best schools.

Fourth, teachers in program schools may have exerted little effort because they believed that learning has only a small impact on the scores of the tests currently in use. Alternative tests that better measure long-run learning might have provided better incentives. However, since the incentives set by ICS were based on the official government of Kenya exams, which in turn are based on the official curriculum, any incentive program based on these exams is likely to run into similar difficulties.

Experience with alternative programs conducted in the same areas suggests that an alternative incentive system for pre-school teachers based on headmaster monitoring failed because the headmasters did not in fact monitor the teachers accurately. This suggests that ultimately, an analysis of the problem must turn to the political economy of education.

Kenya's centralized education system is not producing adequate incentives and these results suggest that adding salary bonuses for teachers based on test scores under the existing system is not adequate to address the problem. It may be worthwhile to consider more fundamental reforms such as devolving control over teachers to local school committees or allowing parents to choose schools and tying school finance more tightly to their decisions, as in

²⁴ In practice, many teachers transfer between schools.

school voucher programs.²⁵ Recent evidence from a program in the same region suggests that hiring teachers locally on short term contracts leads to greater teacher attendance and persistently greater student achievement (Duflo, Dupas, and Kremer (2007)).

²⁵ Since students' placement in secondary school in Kenya depends on performance on the primary-school leaving exam, local communities and parents could share some of the same incentives to focus on test preparation as teachers (see Acemoglu et al., 2008). Nonetheless, since teachers and headmasters transfer between schools fairly frequently, parents are likely to take a somewhat longer run perspective, with at least a somewhat greater focus on human capital acquisition. Indeed, Kremer et al. (2007) find that incentives for students led to higher substantive effort and test score gains that persisted after the program ended rather than to increases in prep sessions.

References

- Acemoglu, Daron, Kremer, Michael, and Atif Mian (2008), "Incentives in Markets, Firms and Governments," *Journal of Law, Economics, and Organization*, forthcoming
- American Federation of Teachers, Teacher Salary Survey Archives at <http://www.aft.org/research/survey00/salariesurvey00.pdf>
- Andrabi, Tahir, Jishnu Das, Asim I. Khwaja, and Tristan Zajonc (2008), "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics," mimeo.
- Chapman, David W., Snyder, Conrad W., and Shirley A. Burchfield (1991), "Teacher Incentives in the Third World," Agency for International Development Report no. 143 (Washington, DC).
- Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan and Halsey Rogers (2006), "Missing in Action: Teacher and Health Worker Absence in Developing Countries". *Journal of Economic Perspectives*, vol. 20, 1, pp.91-116.
- Clotfelter, Charles T. and Helen F. Ladd (1996), "Recognizing and Rewarding Success in Public Schools," in Helen F. Ladd ed. *Holding Schools Accountable: Performance-Based Reform in Education*, Brookings Institution, Washington, D.C.
- Duflo, Esther, Pascaline Dupas, and Michael Kremer (2007), "Peer Effects, Pupil-teacher Ratios, and Teacher Incentives," mimeo.
- Eberts, Randall, Hollenbeck, Kevin, and Joe Stone (2003), "Teacher Performance Incentives and Student Outcomes," *Journal of Human Resources*, vol. 37, 4, pp. 913-927.
- Education Commission of the States (2000), "Pay-for-Performance: Key Questions and Lessons from Five Current Models," ECS Issue Paper, Education Commission of the States, at www.ecs.org/clearinghouse/28/30/2830.htm.
- Figlio, David N. and Joshua Winicki (2002), "Food for Thought: The Effects of School Accountability Plans on School Nutrition," National Bureau of Economics Working Paper 9319.
- Glewwe, Paul, Ilais, Nauman, and Michael Kremer (2003), "Teacher Incentives," National Bureau of Economics Working Paper 9671.
- Glewwe, Paul, Kremer, Michael, and Sylvie Moulin (2008), "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics* (forthcoming).
- Gramlich, Edward and Patricia Koshel (1975), "Educational Performance Contracting," Brookings Institution, Washington, D.C.

- Hannaway, Jane (1992), "Higher Order Thinking, Job Design, and Incentives: An Analysis and Proposal," *American Education Research Journal*, vol. 29, 1, pp. 3-21.
- Hanushek, Eric A. (1996), "Outcomes, Cost, and Incentives in Schools," in Hanushek, Eric A. and Dale W. Jorgenson, eds. *Improving America's schools: The Role of Incentives*, National Academy, Washington, D.C.
- Hanushek, Eric A., with Benson, Charles S., et al. (1994), *Making Schools Work: Improving Performance and Controlling Costs*, Brookings Institution, Washington, D.C.
- Hanushek, Eric A., Kain, John F., and Steven R. Rivkin (1999), "Do Higher Salaries Buy Better Teachers?", unpublished.
- Hanushek, Eric A., Kain, John F., and Steven R. Rivkin (1998), "Teachers, Schools, and Academic Achievement," National Bureau of Economic Research Working Paper 6691.
- Harbison, Ralph W. and Eric A. Hanushek (1992), *Educational Performance of the Poor: Lessons from Rural Northeast Brazil*, NY: Oxford University Press.
- Holmstrom, Bengt and Paul Milgrom (1991), "Multi-Task Principal-Agent Analysis: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics and Organization*, vol. 7, 0, pp. 24-52.
- Jacob, Brian (2002), "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools," National Bureau of Economics Working Paper 8968.
- Jacob, Brian and Stephen Levitt (2003), "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating," *Quarterly Journal of Economics*, vol 118, 3, pp.843-877.
- Kingdon, Geeta and Francis Teal (2002), "Does Performance Related Pay for Teachers Improve Student Performance? Some Evidence from India," mimeo.
- Koretz, Daniel (2002), "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity," *Journal of Human Resources*, vol. 37, 4, pp. 752-778.
- Kremer, Michael and Daniel Chen, (2004), "Notes on an Incentive Scheme for Pre-School Teachers."
- Kremer, Michael, Karthik Muralidharan, Nazmul Chaudhury, Jeffrey Hammer, F. Halsey Rogers, (2004), "Teacher Absence in India: A Snapshot," *Journal of the European Economic Association* (Forthcoming).
- Kremer, Michael, Edward Miguel and Rebecca Thornton (2007), "Incentives to Learn", *Review of Economics and Statistics*, forthcoming.

- Lavy, Victor (2002), "Evaluating the Effect of Teacher Group Performance Incentives on Students Achievements," *Journal of Political Economy*, vol. 110, 6, pp. 1286-1318.
- Lavy, Victor (2004), "Performance Pay and Teachers' Effort, Productivity, and Grading Ethics," National Bureau of Economics Working Paper 10622.
- Lockheed, Marlaine E. and Adriaan M. Verspoor (1991), *Improving Primary Education in Developing Countries*, NY: Oxford University Press.
- Muralidharan, Karthik and Venkatesh Sundararaman (2007), "Teaching Incentives in Developing Countries: Experimental Evidence from India," mimeo.
- Olsen, Lynn (1999), "Pay-Performance Link in Salaries Gains Momentum," *Education Week* October 13, 1999.
- PROBE (Public Report on Basic Education for India) (1999), Oxford University Press.
- Richards, Craig and Tian M. Sheu (1992), "The South Carolina School Incentive Reward Program: A Policy Analysis," *Economics of Education Review*, vol. 11, 1, pp. 71-86.
- World Bank, (2005), "Kenya - Public Expenditure Review 2004," Washington, D.C.

Table 1: Concentration of Teacher Absences

<i>Attendance Probability</i>	<i>Share of teachers</i>	<i>% of Total Absence From This Group</i> <i>Total Absence ≈ 20%</i>
Empirical Distribution		
0.000	0.67%	4.18%
0.125	0.22%	1.20%
0.250	0.22%	1.03%
0.375	2.00%	7.80%
0.500	4.67%	14.58%
0.625	9.56%	22.39%
0.750	14.44%	22.48%
0.875	33.78%	26.31%
1.000	34.44%	0%
Five Group, Non-Parametric Model		
0.011	3.8%	17.0%
0.079	2.8%	11.7%
0.581	12.2%	23.2%
0.815	42.5%	35.6%
0.929	38.7%	12.4%
MLE Estimate of Beta Distribution Model: $\alpha = 8.62$; $\beta = 1.57$		
0 < p < 0.5	0.73%	2.7%
0.51 < p < 0.6	2.4%	7.1%
0.61 < p < 0.7	7.4%	16.7%
0.71 < p < 0.8	17.9%	28.6%
0.81 < p < 0.9	34.0%	32.1%
0.91 < p < 1.0	37.5%	12.7%

Table 2: Program Impact on Teacher Entry and Exit

<i>Dependent Variable:</i>	(1)	(2)	(3)	(4)
	Recently Exited from Initial School		Recently Entered Current School	
	Exit between 1997 and 1998	Exit between 1998 and 1999	Enter in 1998	Enter in 1999
Incentive	0.041	0.007	0.026	-0.002
School	(0.033)	(0.026)	(0.030)	(0.034)
Male	0.047	0.020	0.043	-0.092
	(0.032)	(0.031)	(0.032)	(0.035)**
Constant	0.137	0.209	0.190	0.234
	(0.024)**	(0.019)**	(0.022)**	(0.025)**
Observations	1157	1227	1227	1228

Notes:

Standard errors in parentheses, based on regressions that include school-level random effects.

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

For exit regressions, incentive/non-incentive refers to the initial school; for entry regressions incentive/non-incentive refers to the school entered. The unit of observation in all regressions is the teacher.

Table 3: Program Impact on Teacher Attendance

<i>Dependent Variable:</i>	(1)	(2)	(3)	(4)	(5)
	Teacher Attendance Percentage			Attendance Differences (<i>Attendance</i> _{program year} – <i>Attendance</i> _{pre-program year})	
	Year 0	Year 1	Year 2	Year 1 - Year 0	Year 2 – Year 0
Incentive	0.012	-0.008	-0.011	-0.007	-0.063
School	(0.043)	(0.019)	(0.022)	(0.048)	(0.049)
Grade	-0.005	-0.010	0.000	-0.009	0.002
	(0.012)	(0.007)	(0.009)	(0.015)	(0.016)
Male (0/1)	0.015	0.007	-0.108	-0.028	-0.095
	(0.045)	(0.022)	(0.025)**	(0.053)	(0.055)*
Constant	0.828	0.882	0.904	0.049	0.064
	(0.073)**	(0.044)**	(0.049)**	(0.090)	(0.094)
Observations	466	397	320	396	319

Notes:

Standard errors in parentheses, based on regressions that include school-level random effects.

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

The dependent variable is the percentage of the visits for which the teacher was present, based on up to two visits in 1997, five visits in 1998 and three visits in 1999. The unit of observation is the teacher.

Table 4: Program Impacts on Pedagogy

	(1) Year 0	(2) Year 1	(3) Year 2	(4) Year 1 - Year 0	(5) Year 2 - Year 0
Panel A - Dependent Variable: Teacher Present in the Classroom					
Incentive	-0.066	-0.034	0.132	0.060	0.078
School	(0.157)	(0.139)	(0.167)	(0.153)	(0.130)
Grade	-0.061	0.038	-0.096	-0.092	-0.065
	(0.042)	(0.036)	(0.044)**	(0.052)*	(0.047)
Constant	0.797	0.313	0.211	0.589*	0.850
	(0.315)**	(0.276)	(0.338)	(0.356)	(0.330)**
Observations	631	826	481	380	400
Panel B - Dependent Variable: Use of Blackboard					
Incentive	0.018	-0.032	0.038	-0.051	0.078
School	(0.031)	(0.026)	(0.051)	(0.036)	(0.069)
Grade	0.010	-0.003	-0.018	-0.001	-0.029
	(0.009)	(0.007)	(0.013)	(0.014)	(0.022)
Constant	0.875	0.973	0.989	0.021	0.098
	(0.054)**	(0.044)**	(0.084)**	(0.085)	(0.133)
Observations	404	598	237	246	149
Panel C - Dependent Variable: Use Teaching Aid					
Incentive	-0.026	-0.006	0.012	0.025	0.052
School	(0.032)	(0.031)	(0.035)	(0.052)	(0.067)
Grade	-0.021	-0.006	-0.004	-0.016	0.002
	(0.012)*	(0.009)	(0.013)	(0.021)	(0.025)
Constant	0.235	0.143	0.094	0.093	-0.040
	(0.073)**	(0.059)**	(0.080)	(0.124)	(0.151)
Observations	399	567	235	241	147
Panel D - Dependent Variable: Teacher Caring (1 to 5: 1=very caring)					
Incentive	-0.080	-0.065	-0.051	0.052	-0.058
School	(0.104)	(0.062)	(0.125)	(0.133)	(0.178)
Grade	0.018	-0.010	0.125	-0.025	0.093
	(0.034)	(0.022)	(0.031)**	(0.048)	(0.062)
Constant	1.586	1.701	1.184	0.122	-0.280
	(0.204)**	(0.135)**	(0.205)**	(0.292)	(0.375)
Observations	382	571	234	238	146
Panel E - Dependent Variable: Teacher Energy (1 to 5: 1=energetic)					
Incentive	-0.030	-0.041	0.164	0.050	0.070
School	(0.096)	(0.080)	(0.120)	(0.167)	(0.195)
Grade	-0.023	-0.019	0.070	-0.017	0.092
	(0.035)	(0.023)	(0.027)**	(0.052)	(0.062)
Constant	1.926	1.870	1.126	0.073	-0.798
	(0.211)**	(0.146)**	(0.180)**	(0.324)	(0.377)**
Observations	383	570	233	239	146
Panel F - Dependent Variable: Homework Assignment					
Incentive	0.012	-0.052	-0.009	-0.092	-0.042
School	(0.042)	(0.045)	(0.047)	(0.055)*	(0.059)

Grade	0.079 (0.007)**	0.062 (0.007)**	0.149 (0.007)**	-0.017 (0.017)	0.036 (0.017)**
Constant	-0.176 (0.049)**	-0.060 (0.053)	-0.586 (0.055)**	0.137 (0.111)	-0.155 (0.111)
Observations	1914	1676	2371	431	427

Notes:

Standard errors in parentheses, based on regressions that include school-level random effects.

* significant at 10%; ** significant at 5%.

Each observation in columns 1 through 3 represents a classroom; differences in columns 4 and 5 are calculated at the school-grade level. Observations vary by year because the number of visits made to schools varied across years.

For homework regressions, in columns 1 through 3 each observation represents a student asked about homework assignment in the previous day; in columns 4 and 5 differences across years are calculated at the school-grade level.

Table 5: Program Impact on Test Preparation Sessions

<i>Dependent Variable:</i>	(1)	(2)	(3)	(4)	(5)
<i>Year 0</i>	<i>Year 1</i>	<i>Year 2</i>	<i>Year 1 - Year 0</i>	<i>Year 2 - Year 0</i>	
		Preparations			Preparation Differences (<i>Preparation</i> _{program year} - <i>Preparation</i> _{pre-program year})
<i>Preparations (Vacation and During School)</i>					
Incentive School	-0.007 (0.044)	0.042 (0.037)	0.074 (0.034)**	0.049 (0.042)	0.081 (0.047)*
Grade	0.155 (0.009)***	0.135 (0.007)***	0.103 (0.007)***	-0.021 (0.009)**	-0.052 (0.009)***
Constant	-0.502 (0.064)***	-0.372 (0.053)***	-0.121 (0.052)**	0.130 (0.064)**	0.381 (0.064)***
Observations	3000	3000	3000	3000	3000
<i>Vacation Preparations</i>					
Incentive School	0.035 (0.034)	0.089 (0.031)***	0.091 (0.035)**	0.055 (0.038)	0.056 (0.046)
Grade	0.156 (0.008)***	0.139 (0.006)***	0.118 (0.007)***	-0.017 (0.008)*	-0.038 (0.009)***
Constant	-0.527 (0.049)***	-0.425 (0.049)***	-0.219 (0.054)***	0.103 (0.057)*	0.308 (0.059)***
Observations	1500	1500	1500	1500	1500

Notes:

Standard errors in parentheses, based on regressions that include school-level random effects.

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

Preparations are reported at 6 times during the year for each grade: 3 vacation terms and three periods during the year; each observation represents a school grade at a given time during the year. Dummy variables for these six time periods are included in all regressions, but the associated parameter estimates are not shown here. The mean number of times that preps were held in year zero, by grade, were: 1.0 for grade 4, 1.1 for grade 5, 2.4 for grade 6, 3.5 for grade 7, and 4.5 for grade 8.

Table 6: Program Impact on Dropout and Repetition Rates

	(1)	(2)	(3)	(4)
	Year 1		Year 2	
	Dropout	Repetition	Dropout	Repetition
Incentive School	0.008 (0.021)	-0.023 (0.017)	-0.002 (0.010)	-0.013 (0.026)
Male	0.030 (0.005)**	0.009 (0.008)	0.026 (0.006)**	-0.007 (0.007)
Constant	0.109 (0.015)**	0.291 (0.013)**	0.110 (0.007)**	0.247 (0.019)**
Observations	14,014	11,892	13,622	12,718

Notes:

Standard errors in parentheses, based on regressions that include school-level random effects.

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

Each observation represents an upper primary school student.

Table 7: Program Impacts on Participation in Exams

	(1) Year 0 (Pre-Program)	(2) Year 1	(3) Year 2	(4) Year 3 (Post-Program)
Panel A				
<i>Dependent Variable: Take Government exam (0/1)</i>				
Incentive School	-0.002 (0.025)	0.057 (0.037)	0.118 (0.027)**	-0.007 (0.030)
Male (0/1)	-0.0007 (0.006)	0.015 (0.008)	-0.007 (0.008)	-0.004 (0.011)
Grade	0.018 (0.002)**	0.059 (0.003)**	0.041 (0.003)**	0.007 (0.005)
Constant	0.709 (0.022)**	0.356 (0.031)**	0.477 (0.028)**	0.563 (0.040)**
Observations	15,224	11,122	10,654	8,055
Panel B				
<i>Dependent Variable: Take NGO exam (0/1)</i>				
Incentive School	0.005 (0.013)	0.023 (0.023)	0.045* (0.024)	0.007 (0.042)
Male (0/1)	-0.004 (0.006)	0.003 (0.005)	0.002 (0.006)	0.021 (0.019)**
Grade	0.012 (0.002)**	0.018 (0.002)**	0.010 (0.003)**	0.028 (0.009)**
Constant	0.767 (0.016)**	0.757 (0.020)**	0.793 (0.023)**	0.540 (0.068)**
Observations	15,718	14,982	10,654	2,226

Note: Standard errors in parentheses, based on regressions that include school-level random effects.

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

Government test data were not available for Teso District in Year 1 (1998), and the ICS data for Year 3 are limited to 27 schools.

ITT methodology employed.

Each observation represents an upper primary school pupil in year 0; columns 2 and 3 are limited to pupils who did not drop out or transfer to another school in those years.

Table 8: Program Impact on Test Scores, by Type of Test (Excluding Repeaters)

<i>Dependent Variable</i>	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Test Scores (Standardized relative to comparison schools)				Test Score Differences $\text{Test Score}_{\text{program year}} - \text{Test Score}_{\text{pre-program year}}$		
	Year 0	Year 1	Year 2	Year 3	Year 1 – Year 0	Year 2 – Year 0	Year 3 – Year 0
Panel A -- Incentive Tests							
<i>Dependent Variable: Government exams Test Scores</i>							
All Subjects & Grades	0.011 (0.091)	-0.040 (0.079)	0.136 (0.077)*	-0.087 (0.083)	0.054 (0.054)	0.139 (0.065)**	-0.008 (0.084)
Observations	24,716	50,842	37,620	15,893	24,677	15,641	5,330
Panel B -- Non-Incentive Tests							
<i>Dependent Variable: NGO test Scores</i>							
All Subjects & Grades	0.071 (0.086)	0.089 (0.079)	0.083 (0.090)	--	0.017 (0.033)	0.016 (0.063)	--
Observations	33,162	39,510	12,996	--	32,993	10,512	--

Note: Standard errors in parentheses, regressions include school-level random effects.

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

Year 1 government test results are available only for Busia.

Each row represents a random effects regression of test scores on a dummy variable for teacher incentive schools and on region and sex dummy variables, based on data on the 100 schools in Teso and Busia Districts. For each grade/subject combination, test scores were standardized by subtracting the mean score and dividing by the standard deviation of the test score from the comparison schools.

Each observation represents a test score in a particular subject for an upper primary school pupil; columns 2, 3, and 4 are limited to pupils who were enrolled in year 1 and did not repeat or drop out. Columns 5, 6 and 7 impose the additional restriction that a pre-test score is available.

7,846 students (grades 4 to 8) took at least one government exam in year 1. Of these, 5,751 had pre-test scores from a pre-program year, in this case 1996. In year 2, when exam results are also available for Teso, 10,927 students (grade 4 to 8) took at least one exam and 6,365 of these students also had pre-test scores from the same pre-program year. In the post-program year, 9,613 students (grade 4 to 8) took at least one exam and 4,016 of these had pre-test scores. In later years more students have no pre-test scores because students who enter the sample (by reaching 4th grade) after the first year do not have pre-test scores. So, for example, in the post-program year students in 4th and 5th grade do not have pre-test scores.

Year 3 NGO tests were given only in 27 schools so scores are not reported.

13,339 students (grades 4 to 8) took at least one subject of the NGO exams in year 1. Of these, 11,298 had pre-test scores from year 0, in the form of normalized government exam scores from year 0. 15,647 students took at least one NGO exam in year 2, of which 8,638 had pre-test scores from year 0.

Table 9: Program Impact on Test Scores, by Type of Test (Including Repeaters)

<i>Dependent Variable</i>	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Test Scores (Standardized relative to comparison schools)				Test Score Differences $\text{Test Score}_{\text{program year}} - \text{Test Score}_{\text{pre-program year}}$			Teacher Reward Formula Score
	Year 0	Year 1	Year 2	Year 3	Year 1 – Year 0	Year 2 – Year 0	Year 3 – Year 0	Year 2
Panel A -- Incentive Tests								
<i>Dependent Variable: Government Exams Test Scores</i>								
All Subjects & Grades	0.011	-0.040	0.090	-0.141	0.054	0.148	0.006	0.145***
	(0.091)	(0.079)	(0.076)	(0.076)*	(0.054)	(0.059)**	(0.067)	(0.023)
Observations	24,716	50,842	54,293	33,036	24,677	23,679	12,519	72,224
Panel B -- Non-Incentive Tests								
<i>Dependent Variable: NGO test Scores</i>								
All Subjects & Grades	0.071	0.089	0.105	--	0.017	0.043	--	--
	(0.086)	(0.079)	(0.098)		(0.033)	(0.053)		
Observations	33,162	39,510	18,558	--	32,993	15,225	--	--

Note: Standard errors in parentheses, regressions include school-level random effects.

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

Year 1 government test results are available only for Busia. Year 2 and year 3 government test results are available for both Busia and Teso. This leads to the increase in sample size from year 1 to year 2 on the government exam, when repeaters are included.

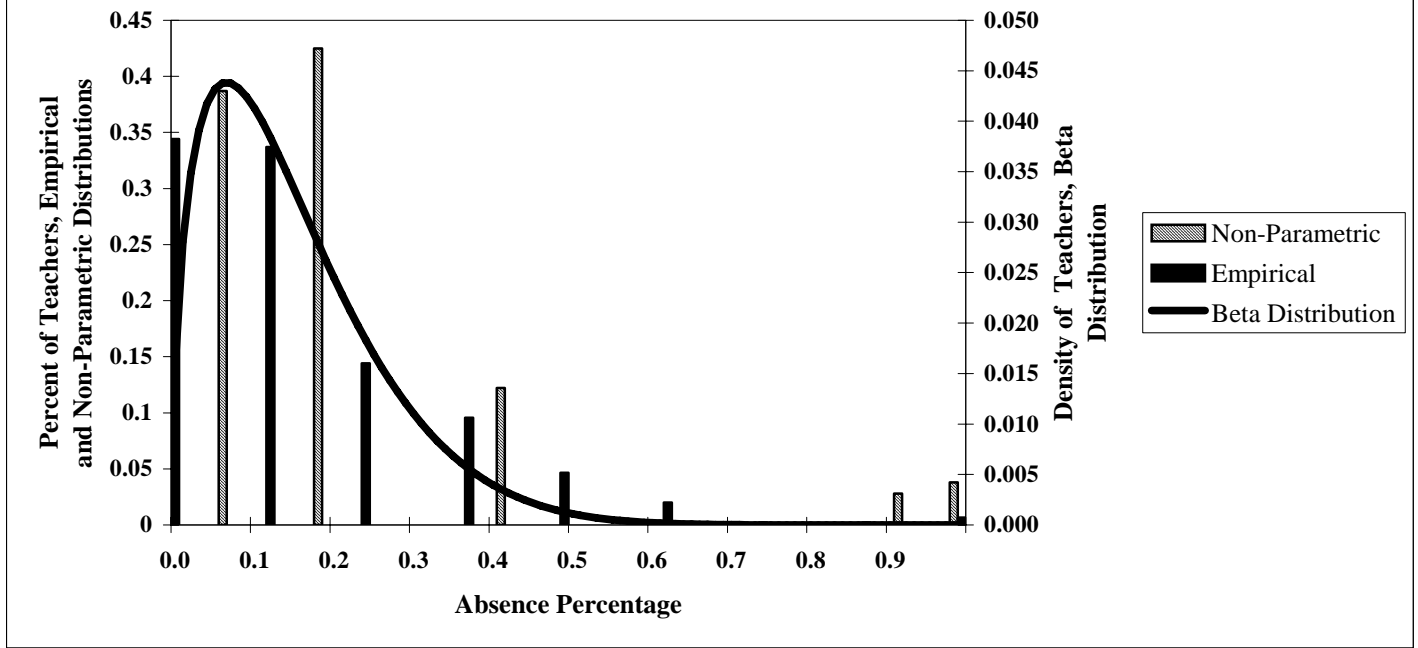
The scores of the repeaters were adjusted by subtracting 1 from their normalized test scores, due to the fact that by repeating the same year, they were essentially taking easier exams and their scores were about one standard deviation above the average test scores in comparison schools.

Table 10: Impact of Program on Correctly Answering Different Types of Questions
(Year 2 tests, grades 4-8)

	<i>Any of last 4 mult. choice questions correct? (random effects probit)</i>	<i>Ratio of % of mult. choice questions correct over % of fill in blank questions correct</i>	<i>None of last 10 questions correctly answered</i>	<i>None of last 20 questions correctly answered</i>	<i>Number of answers left blank (negative binomial regression)</i>
Grades	5-8	5-8	4-8	4-8	4-8
Subject	English	English	English, math	English, math	English, math, science
Incentive school	0.264*** (0.92)	0.041** (0.019)	-0.036* (0.022)	-0.025* (0.015)	-0.616* (0.374)
Observations	8449	10,157	27,756	27,756	603

Notes: The first four regressions included sex, grade and division dummy variables as control variables. The first two columns omit grade 4 because grade 4 did not have any multiple choice questions after the first easy 20 multiple choice questions. The first four columns use the student as the unit of observation. The last column has grade and subject combinations as the unit of observation, and the only control variables are interactions between grade and subject dummy variables.

Figure 1
Three Models of Teacher Absence Distribution



Appendix Tables: Subject-Specific Test Score Regressions

Excluding Repeaters

<i>Dependent Variable</i>	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Test Scores (Standardized relative to comparison schools)				Test Score Differences Test Score _{program year} – Test Score _{pre-program year}		
	Year 0	Year 1	Year 2	Year 3	Year 1 – Year 0	Year 2 – Year 0	Year 3 – Year 0
<i>Dependent Variable: Government exams Test Scores</i>							
English	0.041 (0.121)	-0.059 (0.107)	0.094 (0.094)	0.017 (0.112)	-0.024 (0.071)	-0.003 (0.086)	-0.091 (0.122)
Math	-0.032 (0.099)	0.058 (0.089)	0.099 (0.084)	-0.077 (0.089)	0.076 (0.054)	0.150 (0.064)**	-0.106 (0.089)
Science	-0.045 (0.098)	0.015 (0.091)	0.155 (0.102)	0.121 (0.115)	0.050 (0.076)	0.206 (0.094)*	0.194 (0.128)
Swahili	-0.020 (0.101)	-0.052 (0.093)	-0.105 (0.072)	0.091 (0.084)	0.023 (0.083)	0.019 (0.094)	-0.134 (0.221)
G.H.CR.	-0.139 (0.115)	-0.039 (0.089)	0.202 (0.097)**	0.055 (0.105)	0.205 (0.107)*	0.341 (0.129)**	-0.021 (0.262)
A.C.M.	-0.128 (0.114)	-0.007 (0.096)	0.010 (0.092)	-0.049 (0.102)	0.116 (0.121)	0.108 (0.154)	-0.218 (0.249)
HS. BE.	-0.008 (0.152)	0.049 (0.092)	0.073 (0.107)	-0.079 (0.113)	0.073 (0.161)	0.167 (0.196)	-1.232 (0.525)**
All Subjects & Grades	0.011 (0.091)	-0.040 (0.079)	0.136 (0.077)*	-0.087 (0.083)	0.054 (0.054)	0.139 (0.065)**	-0.008 (0.084)
Observations	24,716	50,842	37,620	15,893	24,677	15,641	5,330

Note: Standard errors in parentheses, regressions include school-level random effects.

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

Year 1 government test results are available only for Busia.

KCPE tests are taken by grade 8 students only.

Each row represents a random effects regression of test scores on a dummy variable for teacher incentive schools and on region and sex dummy variables, based on data on the 100 schools in Teso and Busia Districts. For each grade/subject combination, test scores were standardized by subtracting the mean score and dividing by the standard deviation of the test score from the comparison schools.

Normalized government test scores from year 0 (1996) were used as the KCPE pre-program scores.

Each observation represents a test score in a particular subject for an upper primary school pupil; columns 2, 3, and 4 are limited to pupils who were enrolled in year 1 and did not repeat or drop out. Columns 5, 6 and 7 impose the additional restriction that a pre-test score is available.

7,846 students (grades 4 to 8) took at least one government exam in year 1. Of these, 5,751 had pre-test scores from a pre-program year, in this case 1996. In year 2, when exam results are also available for Teso, 10,927 students (grade 4 to 8) took at least one exam and 6,365 of these students also had pre-test scores from the same pre-program year. In the post-program year, 9,613 students (grade 4 to 8) took at least one exam and 4,016 of these had pre-test scores. In later years more students have no pre-test scores because students who enter the sample (by reaching 4th grade) after the first year do not have pre-test scores. So, for example, in the post-program year students in 4th and 5th grade do not have pre-test scores.

1,490 eighth graders took at least one KCPE exam in year 1, of which 1,026 had pre-test scores from year 0. 1,584 students took at least once KCPE exam in year 2, of which 944 had pre-test scores. 1,537 students took at least one KCPE exam in year 3, of which 839 had pre-test scores.

**Table A.2: Program Effect on Test Scores by Subject (ICS Subject Tests)
Excluding Repeaters**

<i>Dependent Variable</i>	(1)	(2)	(3)	(4)	(5)
	Test Scores (Standardized relative to comparison schools)			Test Score Differences	
	Year 0	Year 1	Year 2	Year 1 – Year 0	Year 2 – Year 0
English	0.070 (0.097)	0.077 (0.090)	0.077 (0.138)	0.001 (0.040)	0.031 (0.099)
Math	0.084 (0.088)	0.053 (0.074)	0.069 (0.074)	-0.042 (0.041)	-0.009 (0.058)
Science	0.036 (0.088)	0.129 (0.082)		0.091 (0.043)**	
All Subjects & Grades	0.071 (0.086)	0.089 (0.079)	0.083 (0.090)	0.017 (0.033)	0.016 (0.063)
Observations	33,162	39,510	12,996	32,993	10,512

Note: Standard errors in parenthesis; regressions include school-level random effects.

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

Year 3 NGO tests were given only in 27 schools so scores are not reported. Each row represents a random effects regression of test scores on a dummy variable for teacher incentive schools and on region and sex dummy variables, based on data on the 100 schools in Teso and Busia Districts. For each grade/subject combination, test scores were standardized by subtracting the mean score and dividing by the standard deviation of the test score from the comparison schools.

Each observation represents a test score in a particular subject for an upper primary school pupil; columns 2 and 3 are limited to pupils who were enrolled in year 1 and did not repeat or drop out. Columns 4 and 5 impose the additional restriction that a pre-test score is available.

13,339 students (grades 4 to 8) took at least one subject of the NGO exams in year 1. Of these, 11,298 had pre-test scores from year 0, in the form of normalized government exam scores from year 0. 15,647 students took at least one NGO exam in year 2, of which 8,638 had pre-test scores from year 0.

Table A.3: Program Effect on Test Scores by Subject (Government exam), Including Repeaters

<i>Dependent Variable</i>	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Test Scores (Standardized relative to comparison schools)				Test Score Differences			Teacher Reward
	Year 0	Year 1	Year 2	Year 3	Year 1 – Year 0	Year 2 – Year 0	Year 3 – Year 0	Formula Score Year 2
<i>Dependent Variable: Government exams Test Scores</i>								
English	0.041 (0.121)	-0.059 (0.107)	0.094 (0.094)	-0.026 (0.104)	-0.024 (0.071)	0.030 (0.080)	-0.031 (0.098)	0.134* (0.059)
Math	-0.032 (0.099)	0.058 (0.089)	0.127 (0.083)	-0.032 (0.083)	0.076 (0.054)	0.157 (0.055)**	-0.063 (0.072)	0.142* (0.063)
Science	-0.045 (0.098)	0.015 (0.091)	0.139 (0.100)	0.115 (0.099)	0.050 (0.076)	0.193 (0.089)**	0.143 (0.100)	0.153** (0.065)
Swahili	-0.020 (0.101)	-0.052 (0.093)	0.105 (0.077)	0.045 (0.080)	0.023 (0.083)	0.034 (0.085)	-0.006 (0.117)	0.154** (0.055)
G.H.CR.	-0.139 (0.115)	-0.039 (0.089)	0.214 (0.100)**	0.054 (0.097)	0.205 (0.107)*	0.385 (0.112)**	0.141 (0.134)	0.193** (0.063)
A.C.M.	-0.128 (0.114)	-0.007 (0.096)	0.018 (0.093)	0.002 (0.093)	0.116 (0.121)	0.143 (0.146)	-0.054 (0.177)	0.106* (0.059)
HS. BE.	-0.008 (0.152)	0.049 (0.092)	0.107 (0.111)	-0.025 (0.111)	0.073 (0.161)	0.109 (0.167)	-0.093 (0.333)	0.133* (0.069)
All Subjects & Grades	0.011 (0.091)	-0.040 (0.079)	0.090 (0.076)	-0.141 (0.076)*	0.054 (0.054)	0.148 (0.059)**	0.006 (0.067)	0.145*** (0.023)
Observations	24,716	50,842	54,293	33,036	24,677	23,679	12,519	72,224

Note: Standard errors in parentheses, regressions include school-level random effects.

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

Year 1 government test results are available only for Busia. Year 2 and year 3 government test results are available for both Busia and Teso. This leads to the increase in sample size from year 1 to year 2 on the government exam, when repeaters are included.

The scores of the repeaters were adjusted by subtracting 1 from their normalized test scores, due to the fact that by repeating the same year, they were essentially taking easier exams and their scores were about one standard deviation above the average test scores in comparison schools.

Table A.4: Program Effect on Test Scores by Subject (ICS Subject Tests) Including Repeaters

<i>Dependent Variable</i>	(1)	(2)	(3)	(4)	(5)
	Test Scores (Standardized relative to comparison schools)			Test Score Differences $\text{Test Score}_{\text{program year}} - \text{Test Score}_{\text{pre-program year}}$	
	Year 0	Year 1	Year 2	Year 1 – Year 0	Year 2 – Year 0
English	0.070 (0.097)	0.077 (0.090)	0.100 (0.129)	0.001 (0.040)	0.041 (0.083)
Math	0.084 (0.088)	0.053 (0.074)	0.098 (0.079)	-0.042 (0.041)	0.036 (0.050)
Science	0.036 (0.088)	0.129 (0.082)		0.091 (0.043)**	
All Subjects & Grades	0.071 (0.086)	0.089 (0.079)	0.105 (0.098)	0.017 (0.033)	0.043 (0.053)
Observations	33,162	39,510	18,558	32,993	15,225

Note: Standard errors in parenthesis; regressions include school-level random effects.

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

The scores of the repeaters were adjusted by subtracting 1 from their normalized test scores, due to the fact that by repeating the same year, they were essentially taking easier exams and their scores were about one standard deviation above the average test scores in comparison schools.