Teacher Incentives

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

April 2003

Abstract

Advocates of teacher incentive programs argue that they can strengthen weak incentives, while opponents argue they lead to "teaching to the test." We find evidence that existing teacher incentives in Kenya are indeed weak, with teachers absent 20% of the time. We then report on a randomized evaluation of a program that provided primary school teachers in rural Kenya with incentives based on students' test scores. Students in program schools had higher test scores, significantly so on at least some exams, during the time the program was in place. An examination of the channels through which this effect took place, however, provides little evidence of more teacher effort aimed at increasing long-run learning. Teacher attendance did not improve, homework assignment did not increase, and pedagogy did not change. There is, however, evidence that teachers increased effort to raise short-run test scores by conducting more test preparation sessions. While students in treatment schools scored higher than their counterparts in comparison schools during the life of the program, they did not retain these gains after the end of the program, consistent with the hypothesis that teachers focused on manipulating short-run scores. In order to discourage dropouts, students who did not test were assigned low scores. Program schools had the same dropout rate as comparison schools, but a higher percentage of students in program schools took the test.

[*] Department of Applied Economics, University of Minnesota. E-mail: pglewwe@apec.umn.edu

[♦] Competition Economics, Inc., Washington, DC. E-mail: nilias@competitioneconomics.com

[•] Department of Economics, Harvard University; The Brookings Institution; Center for Global Development 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 U.S. We thank the World Bank and the MacArthur Foundation for financial support.

1. Introduction

Teacher incentive programs have enjoyed growing popularity. In the United States, a number of teacher incentive programs have been introduced in the past decade, generally offering annual merit pay on the order of 10% to 40% of an average teacher's monthly salary (American Federation of Teachers, 2000).[1] Under the No Child Left Behind (NCLB) act, passed in 2001, poorly performing schools face sanctions across the United States. Israel has provided incentives to teachers based on students' scores (Lavy 2002a, b) and a World Bank program in Mexico will provide performance incentives to primary school teachers.

Advocates of incentive pay for teachers note that teachers currently face weak incentives, with pay determined almost entirely by educational attainment, training, and experience, rather than performance (Harbison and Hanushek, 1992; Hanushek, Kain, and Rivkin, 1998; Hanushek, 1996; Lockheed and Verspoor, 1991), and argue that linking teachers' pay to students' performance would increase teacher effort.

Opponents of test score-based incentives argue that since teachers' tasks are multi-dimensional and only some aspects are measured by test scores, linking compensation to test scores could cause teachers to sacrifice promoting curiosity and creative thinking in order to teach skills tested on standardized exams (Holmstrom and Milgrom, 1991; Hannaway, 1992).

In many developing countries, incentives for teachers are even weaker than in developed countries. Thus, for example, in our data teachers are absent from school 20% of the time and absent from their classrooms even more frequently. Work in progress suggests that absence rates among primary-school teachers are 26% in Uganda, 23% in India, 16% in Ecuador and 13% in Peru. (Chaudhury, Hammer, Kremer, Muraldhiran, and Rogers, 2003).

[1] Examples include programs in Rhode Island in 1999, Denver in 1999-2000, Douglas County, Colorado beginning in 1994 and Iowa beginning in 2001 (Olsen, 1999; Education Commission of the States, 2000). A 1999 program in California offered a one-time award of $25,000 to teachers in under-performing schools whose students showed substantial gains (Olsen, 1999).

In environments with very weak incentives, it could be argued that the key problem is to get teachers to show up at all. Given that most teaching in many developing countries is by rote, the risk of reducing efforts to stimulate creativity may seem remote. On the other hand, if incentive systems are very weak, schools could potentially respond to test score-based incentives in more pernicious ways than teaching to the test. For example, they could deliberately force students to repeat grades or even drop out in order to raise average scores on the exam.

This paper examines the issue of teacher incentives in Kenya, where some local school committees strengthen teacher incentives by providing bonuses to teachers whose students perform well on exams. We report on a randomized evaluation of a program along these lines that provided incentives to teachers in 50 rural schools based on the average test score of students already enrolled at the start of the program. Students who did not take the test were assigned very low scores so as to discourage dropouts. Each year the program provided prizes valued at up to 43% of typical monthly salary to teachers in grades 4 to 8 based on the performance of the school as a whole on the Kenyan government's district-wide exams. This ratio of prize to salary was similar to that used in typical U.S. incentive programs.

During the life of the program, students in treatment schools were more likely to take exams, and scored higher, at least on some exams. An examination of the channels through which this effect took place, however, provides little evidence of more teacher effort aimed at preventing dropouts or increasing long-run learning. Dropout rates did not fall, teacher attendance did not improve, homework assignment did not increase, and pedagogy did not change. There is, however, evidence that teachers increased efforts to increase the number of pupils taking tests in the short run and to raise short-run test scores. Conditional on being enrolled, students in treatment schools were more likely to take tests, and teachers in treatment schools were more likely to conduct test preparation sessions. While students in treatment schools scored higher than their counterparts in comparison schools during the two years that the program operated, they did not retain these gains after the end of the program, consistent with a model in which teachers focused on manipulating short-run scores.

There is evidence that teachers learned how 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 rates and overall test scores.

1.1 Related Literature

A number of earlier papers examine the impact of linking teacher pay to students' test scores. Lavy (2002a) finds that an Israeli program providing teachers individual cash prizes for increases in student test scores on a high-school matriculation exam increased high school matriculation exam rates from 42% to 45.3%. At 60% to 300% of the average monthly salary, the prizes given in this case were much larger than those in most teacher incentive programs in the U.S. Lavy (2002b) finds that rewarding Israeli teachers based on school average performance (rather than individual performance) increased test scores and participation in matriculation exams, but not the percentage of students receiving matriculation certificates.

Jacob (2002) explores a Chicago program in which students with low test scores were not promoted to the next grade and schools and teachers were put on probation. He finds that the program increased student achievement, although the improvement was larger in skill sets used on the high-stakes exam. Some schools manipulated scores by putting more students in special education classes. Figlio and Winicki (2002) show that school districts in Virginia increase the number of calories in school lunches on days when high-stakes tests are administered, thus artificially inflating test scores.

This paper differs from earlier work in several ways. First, since both advocates and opponents of teacher incentive programs agree that incentives can increase test scores, but disagree about whether these higher test scores would be due to increased overall teacher effort or more teaching to the test, we measure not only how teacher incentives affect test scores, but also how they affect different types of teacher effort. In particular, we examine teacher behavior in the classroom and scores not only on exams to which incentives were linked, but also on other exams given both contemporaneously with the program and after its conclusion. Second, since teacher incentive programs are likely to be introduced in areas where teacher performance is worse than expected, and since the introduction of teacher incentives may be correlated with other factors affecting teacher performance, it is difficult to econometrically identify the effect of such teacher incentive programs. We address this problem by examining a context with random assignment of schools to treatment and comparison groups. Third, we examine teacher incentives in a developing country context.[2] Finally, by collecting panel data on teacher absence, we are able to show that teacher absence is widespread, suggesting existing incentives are weak in the context we examine.

The remainder of the paper is organized as follows. Section 2 sketches a simple Holmstrom-Milgrom style model in which linking bonuses to test scores could potentially either increase teaching effort or divert effort towards teaching to the test. Section 3 discusses primary education in Kenya and argues that the high rate of teacher absence suggests current incentives are inadequate. Section 4 describes the teacher incentives program that we examined, and the process by which schools were selected for the program. Section 5 reports the impact of the program on teacher outcomes, while Section 6 reports the impact on student outcomes. Section 7 discusses how teacher behavior changed in response to incentives over time, and Section 8 concludes.

[2] Kingdon and Teal (2002) show that in private schools in India teacher pay and student achievement are linked, but they do not demonstrate specifically that linking pay to performance has any effect.

2. A Model of Productive and Signaling Effort

Holmstrom and Milgrom (1991) consider a model in which linking pay to observable signals of performance can potentially lead employees to focus on tasks for which output is easily measured and divert effort away from tasks for which output is difficult to measure. They motivate their analysis using two main examples. In the first, linking teacher pay to test scores may cause teachers to teach to the test rather than encourage creativity. In the second, employees who are responsible both for producing output and for maintaining the value of an underlying asset, such as a piece of equipment or a firm's reputation, may neglect the long run value of the asset if they are provided with strong incentives to focus on current output.

We consider a model that combines elements of both Holmstrom and Milgrom's motivating examples, and can be considered a special case of their general model. Teachers can exert two types of effort: efforts to promote long-run learning and "signaling effort," which improves scores in the short-run but has little effect on long-term learning. Employers observe only test scores. In particular, we assume test scores depend both on underlying learning (produced by teaching effort over time) and contemporaneous signaling effort. Suppose that $T_t = L(e_t, e_{t-1}, e_{t-2}...) + \gamma(s_t) + \varepsilon_t$, where T_t denotes test scores during period *t*, *L* denotes student learning, e_t denotes teaching effort on long-run learning during period *t*, *s* denotes signaling effort, and ε is a random shock. Thus, teaching effort produces long-run improvements in students' understanding, while signaling effort produces only short-run effects on test scores. (Teaching effort can thus be seen as unobservable effort to maintain asset value in Holmstrom and Milgrom's framework.)

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. In this specification, *e* and *s* can be either substitutes or complements. For example, they could be substitutes if there is a fixed amount of time in the day that must be allocated between them. On the other hand, they could be complements if there is a fixed cost to teachers of attending school at all.

Suppose teacher pay is $M = \alpha + BT$. If $B = 0$, so pay is independent of performance, teachers will choose effort in teaching and signaling such that the marginal product of each is equal to zero. As noted by Holmstrom and Milgrom (1991), C_1 (0,0) and C_2 (0,0) may be negative, so some effort may be exerted even if B = 0. Teachers may care about their students, or enjoy exerting some effort even in the absence of performance incentives.

If the government or an NGO makes a surprise announcement 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 equations. Specifically, teachers will exert teaching and signaling effort such 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 then incentives may increase one type of effort at the expense of the other. Thus in this model, incentives could potentially either increase or decrease teaching effort.[3]

In the model, if it were possible to cheaply and accurately monitor individual performance on both tasks as part of an incentive program, then a wage contract could induce teaching effort without inducing signaling effort. However, while distinguishing teaching and signaling effort would be expensive and inaccurate at the individual level, particularly if tied to an incentive program, there are potential ways to distinguish them empirically at the aggregate

[3] Clearly, there is a continuum between exerting effort on promoting long-term learning and trying to manipulate short-run test scores. The extreme of manipulating short-run scores would be actually cheating at the time of the test; less extreme versions would include going over questions from previous years' exams, and teaching test-taking strategies such as guessing on multiple choice questions. Within the category of promoting learning, teachers could focus narrowly on the curriculum to be tested or could promote learning more broadly. One could imagine generalizing this model to allow teachers to choose from a menu of activities, with varying components of true and signaling effort, but results would presumably be similar.

level, at least if results are not tied to compensation. First, outside observers may be able to observe teachers' activities directly. For example, in Kenya, some schools conduct what are known locally as "preps"—extra test preparation or coaching outside of normal class time, often during school vacations. One could potentially interpret preps as including a higher rate of signaling to teaching effort than ordinary classroom attendance. Second, improved learning should have a long-run effect on test scores, whereas under the model signaling has only a short-run effect.[4] Thus a finding that test score gains do not persist is consistent with the hypothesis that the program led only to extra coaching specific to the test at hand. It is more difficult to reconcile this result with the hypothesis of increased long-run learning. A third potential way to distinguish efforts to increase long-run learning from test preparation activities is to check if test scores improved primarily in subjects prone to memorization.

Note that under the model, parents and local communities may not object to teachers' investing in short-run test preparation, since students' prospects for further education and labor market success depend on test scores as well as underlying learning. Test preparation, however, is assumed to be socially wasteful, in that it requires teacher effort but does not improve the underlying learning that affects total output in society.

3. Background

This section provides some background on primary school teaching in Kenya and provides evidence that teacher absence is widespread in the area of the study.

---

[4] In practice some types of signaling may have a long-run effect on test scores. For example, helping students cheat will only increase scores in the short run, but teaching students to guess on multiple choice exams or better allocate their time could raise scores on other tests and in the long run.

Teacher hiring, firing, and transfer decisions in Kenya are made centrally by the Ministry of Education. Hiring is based primarily on academic qualifications.[5] Salaries are set through collective bargaining between the government and the politically powerful Kenyan National Union of Teachers (KNUT). In 1997, the starting salary for teachers was Ksh 5,175 ($88) per month, and a typical teacher in our sample earned approximately Ksh 7,000 ($119) per month.[6] Taking into account generous benefits, total teacher compensation was approximately $2,000 a year, or more than five times annual GDP per capita.[7]

Teachers' salaries depend primarily on education and experience. There is little opportunity for performance-based promotion or increases to salary. Teachers have strong civil service and union protection and are difficult to fire. In some cases teachers who have performed very badly are transferred to less desirable locations, while the government may look more favorably on requests for transfers to more desirable postings or to home areas from teachers who perform well.

Although incentives provided to teachers by their employer are weak, every school is supposed to have a parent committee, and these committees sometimes provide gifts for teachers when schools perform well on the national exams. Similarly, communities sometimes refuse to allow exceptionally bad teachers to enter the school, thus putting pressure on the Ministry of Education to arrange a transfer for the teacher. However, only a minority of school committees provide supplemental bonuses, and school committees typically only attempt to influence the national authorities in extreme situations.

---

[5] Primary school teachers in Kenya typically have completed two years of teacher training beyond secondary school. A small number of teachers were hired under an older system in which primary teachers had only a 7th grade education and two years of teacher training.

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

[7] Authors' calculations based on value of housing allowance and other benefits. This is calculated from salary scales and represents a salary for a teacher with average education and experience in a sample of schools in the area.

To the extent that incentives do exist, they are typically based on the system of national testing. Results on the national primary school leaving exams (the KCPE) are front-page news in Kenya, and newspapers publish the relative positions of different districts in the country on the national exam and lists of the highest-scoring schools. Results from this exam and from district exams administered in the upper grades of primary school are often posted in headmasters' offices. Since the primary school leaving exam determines what secondary schools, if any, will accept graduating primary school students, teachers devote considerable effort to preparing for these exams. An important method of preparation is to review books of questions from old exams. Some schools hold extra preparation sessions for the exams outside of normal class hours. These sessions – locally referred to as "preps" – are held during evenings, weekends or vacation periods. These sessions are made up of a variety of activities, ranging from class-work similar to normal classroom sessions to direct test preparation activities like going over old exams. In general, these sessions will be more heavily weighted to specific test-preparation activities than normal classroom sessions.

While the considerable attention given to results on national exams clearly spurs effort by some teachers to raise their scores, not all of this effort is necessarily desirable. For example, seventh graders who do not perform well are often required to repeat a grade rather than being allowed to go on to 8th grade and take the KCPE exam.

One indication that teacher incentives are weak lies in high absence rates. Random visits conducted to check pupil attendance and observe pedagogy suggest that teachers in comparison schools were absent about 20% of the time. While the 20% absence rate may reflect a variety of factors, including the prevalence of infectious diseases such as malaria and AIDS, these unavoidable absences are unlikely to account for all absences. For comparison, absence rates among staff at a non-profit organization working in the same area are around 6.3%.

Absences seem fairly broadly distributed among the population of teachers rather than primarily accounted for by a subset of teachers with very high absence rates. To see this, it is helpful to consider the empirical distribution of absences, as well as two ways of calibrating the underlying distribution of absences across teachers that correct for the additional dispersion in absence rates across teachers that is created by sampling error.

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, and so on. However, it is important to note that with only a few visits to each school, the dispersion of absence rates in the empirical distribution will exaggerate the underlying dispersion of probabilities of attendance among teachers. For example, suppose there were only two visits, and that one quarter of teachers were absent on both visits, half were absent once, and a quarter were present during both visits. Note that this hypothetical data would be consistent with the hypothesis that all teachers attend half the time but is inconsistent with the hypothesis that a quarter of the teachers have a zero probability of absence, half attend half of the time, and one quarter are always present. In general, due to sampling error, the variance of empirically observed absence rates across teachers will be greater than the underlying variance of probabilities of absence. As the number of visits to each school increases, the empirical distribution of observed teacher absence rates converges to the underlying distribution.

To correct for this problem and assess the extent to which teacher absences are concentrated, we calibrate two more structural models of absences.[8] The first is a non-parametric model in which we assume the population of teachers consists of 5 groups, with proportions α_1 through α_5 and probabilities of attendance p_1 through p_5. There are nine independent unknowns in this model: $\alpha_1 \ldots \alpha_4$ and $p_1 \ldots p_5$. The model is therefore identified

[8] We are grateful to Emily Oster for outstanding work calibrating these two models.

with eight visits, since this gives nine possible outcomes (ranging from never attending to attending at all visits). We observe a proportion x_0 of teachers who are never present, x_1 who are present once, and so on. We solve a system of equations in which the expected proportion of teachers present for J<8 visits (based on $\alpha_1 \ldots \alpha_4$ and $p_1 \ldots p_5$) is set equal to the actual proportion x_J. The resulting group size and attendance probabilities are graphed in Figure 1 and detailed in Table 1. The median teacher in this model is absent about 19% of the time.

The second model imposes more structure, but allows for a continuous distribution of the probability of attendance. For this model we assume that each teacher's probability of attending p that is drawn from a beta distribution, $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$. With eight visits, the probability of observing a teacher with probability of attendance p at school J times is binomial: $\Pr(v = J \mid p) = \binom{8}{J} p^J (1-p)^{8-J}$.

Maximum likelihood analysis yields the parameters of the beta distribution most consistent with the data observed. The likelihood function for J attendances is:

$$\int_0^1 \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right]\left[\binom{8}{J} p^J (1-p)^{8-J}\right] dp\,,$$

or equivalently,

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{8!}{J!(8-J)!} \int_0^1 p^{J+\alpha-1}(1-p)^{\beta+8-J-1}\, dp = \frac{\Gamma(\alpha+\beta)\Gamma(N+1)\Gamma(J+\alpha)\Gamma(\beta+N-J)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(J+1)\Gamma(N-J+1)\Gamma(\alpha+\beta+N)}.$$

The beta distribution of attendance probabilities implied by the ML estimates of α and β is graphed in Figure 1 and detailed in Table 1. The median teacher is absent about 14% of the time in this calibration.

Both models yield similar results. Although a few teachers are rarely present, the majority of absences appear to be due to those who attend between 50% and 80% of the time. In addition, in both models (although more so in the beta distribution model) a large minority of absence is actually due to those who attend *more* than 80% of the time, as can be seen in Table 1.

The widespread nature of absences suggests that teachers who are absent frequently may not be violating a social norm. This is not, however, an implicit contract in which the government pays teachers only for part-time work. As noted above, including benefits, Kenyan teachers are paid up to 5 times average per capita income. There is substantial unemployment among people who would be qualified to become teachers and queuing for teaching jobs. The government imposed a hiring freeze in 1998 so several cohorts of graduates of teacher training colleges are unemployed. Some are working for free in the hope of eventually obtaining a paying job.

Further evidence of weak incentives is that while teachers are absent from school about 20% of the time, they are absent from their classroom much more often. In 1998, the time of teacher arrival was not recorded during classroom visits to observe pedagogy, but in 1999, when it was, 45% of the time teachers never arrived in the classroom.

Recalling the model in Section 2, the fact that we have information on both teacher attendance and preps provides us the opportunity to run a simple regression of test scores on teacher attendance and test preparations sessions. To the extent that one is willing to interpret this causally, this regression suggests that the marginal product of test preparation sessions, which can plausibly be interpreted as signaling effort, may be much higher than that of teacher attendance, a plausible measure of teacher effort.[9] The evidence, based on visits conducted in 1998, suggests that teachers who attend school 20 percentage points more of the time have

[9] Data on teacher attendance and test scores are described below.

students who score 0.0115 standard deviations higher (standard error 0.0123).[10] We do not have information on exactly how long teachers spent on coaching, but only know whether they taught during each of the three vacation periods over the year or outside of normal school hours during each of the three terms. However, those who report coaching in one additional time period have pupils who score 0.044 standard deviations higher (standard error 0.0087). It is somewhat difficult to draw conclusions about the relative marginal productivity of coaching and teaching given that we do not have data on how many days of coaching took place. Based on discussions with teachers about the frequency of coaching, however, it seems likely that, if interpreted causally, our point estimates would imply that the marginal test score effect of a day of coaching is an order of magnitude or more greater than that of a day of school attendance.

Of course, it is not clear that these correlations should be interpreted causally. Nonetheless, the hypothesis that test preparation activities can raise test scores is consistent with evidence from the U.S. on the effects of commercial test preparation. These studies often show gains of 0.15 to 0.4 standard deviations on admissions tests, even though most U.S. admission tests are supposed to measure aptitude, rather than achievement, and thus to be difficult to study for. Extra coaching raised scores on an achievement test by 0.25 standard deviations (Bangert-Drowns *et al.,*1983). In an environment in which teacher pay is not linked to test scores, it seems plausible that teachers might make only limited efforts at test preparation, leaving the marginal test score product of test preparation substantially greater than that of teacher attendance.

The hypothesis that the marginal test score impact of a day of preps is larger than a day of teaching is reasonable if one assumes that incentives are initially weak. Denote e_0 and s_0 as the levels of e and s when B=0. It seems reasonable that teachers' utility from putting in teaching effort may be greater than their utility from signaling (although there may be some benefit from

[10] Note, however, that these estimates will be subject to attenuation bias since teacher attendance is estimated based on random visits and thus measured with substantial error.

signaling if it enhances the teacher's reputation with students and parents). If this is true, then

$$\frac{\partial L}{\partial e}(e_0, s_0) < \frac{\partial \gamma}{\partial s}(e_0, s_0).$$

4. Program Description

As noted above, some school committees in Kenya provide bonuses to teachers whose students perform well in exams. This paper evaluates a program conducted by International Christelijk Steunfonds (ICS), a Dutch NGO, in Busia and Teso Districts of Western Kenya. The program offered schools the opportunity to provide gifts to teachers if students performed well. It provided prizes to teachers in grades 4 to 8 based on the performance of the school as a whole on the district exams in each year. All teachers who taught these grades were eligible for the prize.[11]

ICS awarded prizes in two categories: "Top-scoring schools" and "Most-improved schools." Schools could not win in more than one category. Improvements were calculated relative to performance in the baseline year. Since the results of the district exams were not available for 1997, the scores for 1996 were used as the base to measure improvements. (Henceforth, we will refer to the last pre-program year for which we have data as year 0,[12] the first (1998) and second (1999) years of the program as years 1 and 2 respectively, and the post-program year (2000) as year 3.) In each category, three first prizes were awarded, three second prizes were awarded, three third prizes were awarded, and three fourth prizes were awarded. Thus, overall, 24 out of the 50 schools participating in the program received prizes of some type, and teachers in most schools should have felt that they had a chance of winning a prize. 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.

[11] Teachers of lower grades were not a part of the competition, because there were no district exams for those classes. They received a lantern as a token prize, whether or not they belonged to a winning school

[12] This may be either 1996 or 1997 depending on the type of data.

Education experts generally are more sympathetic to school-based incentives than to individual-based incentives since they feel these are more conducive to cooperation among teachers (Richards and Sheu, 1992; Hanushek, 1996). In order to encourage cooperation among teachers within schools and to avoid creating incentives for teachers to sabotage each other's work, ICS prizes were based on the performance of all of the grade 4 to 8 pupils in the school, with each subject weighted equally, rather than on a teacher-by-teacher basis. Thus, every teacher in grades 4 to 8 in the winning schools received the same prize. Setting prizes at the school level could potentially allow free-riding within the teaching staff. However, teachers are in a relatively good position to monitor each other's performance, particularly on easily observed aspects such as attendance. Moreover, since teachers can observe each other's work at high frequency, they were in a repeated game with each other. Since the typical school in the sample had only 200 students and 12 teachers, about half of whom taught in the upper grades, teachers should have been able to enforce cooperation within the school.

In order to discourage schools from forcing weaker students to repeat, drop out, or not take the exam, students who did not take the exam were assigned low scores. Multiple choice exams were used in all subjects other than English. Students who did not take these exams were assigned a score of 15, whereas students who simply guessed would have obtained a score of 25 on average. On the English essay exam component, students who did not take the test were assigned a score of zero. In order to discourage schools from recruiting strong students to take the exams, only students enrolled in school as of February 1998 were included in the computation of the school mean score.

Prizes ranged in value from 21 to 43% of typical teacher monthly salaries. [13] This is comparable to merit pay programs in the United States. For example, the 1993-94 Dallas merit

[13] Each winning school also received a briefcase for the headmaster, a wall clock, a time keeping clock, and a bell.

pay program, which was also based on school-wide performance, awarded $1000 annual bonuses, which were 39% of an average monthly salary of Texas teachers that year, and presumably a somewhat 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 monthly salary (Olsen, 1999; American Federation of Teachers, 2000). Other programs in Colorado award between 10% and 50% of monthly salary in merit-based annual bonuses (Education Commission of the States, 2001).

Schools were offered the opportunity to participate in this program in February 1998 and all accepted. (The academic year in Kenya runs from January to November.) The prizes were awarded during a ceremony held in November of each year, and all the schools in the program were invited to attend. In order to discourage teachers from arranging transfers into treatment schools in order to be eligible for the program, eligibility was restricted to teachers who were employed in the school as of March 1998. As discussed below, teacher entry and exit rates did not differ significantly between treatment and comparison schools.

When the program was originally announced in February 1998, it was scheduled to run for a single year and teachers were informed of this. Later the program was extended for an additional year. Because the NGO had conducted other programs in the area, we think most teachers found the commitment to provide the prizes credible. However, it is possible that teachers did not react fully to the program until after they had actually seen the prizes awarded during the first year. The awards ceremony presumably increased the salience of the program. It is not clear whether teachers expected the program to continue longer than the NGO had promised, since the NGO tries to be conservative in announcing benefits it will provide to avoid creating overly high expectations.

The short duration of the program and its surprise introduction allowed several elements to be included that might not otherwise have been possible in a permanent program. Half the prizes were based on improvements in performance. Teachers in many lower performing schools may have felt these were the only prizes for which they could realistically compete. The incentives created by these prizes for improvement were presumably larger than they would have been in a permanent program in which teachers who increased their scores in one year would find it harder to win the prize in subsequent years. Moreover, the short duration of the program made it possible to base prizes on the test scores of all students originally enrolled in the school, which allowed the program to discourage repetition and dropouts. Under a permanent program, schools might have incentives to restrict admissions to school to students who they believe would score well on exams. Similarly, the program was restricted to the original teachers at this school. It thus reduced the opportunity for teachers to seek assignment to schools with high-scoring pupils due to non-teacher factors such as parental characteristics. All these factors imply that the incentives provided by the program were stronger than could have been provided with the same expenditure under a permanent program. On the other hand, teachers might have had stronger incentives to promote long-run learning if they expected the program to continue indefinitely.

Overall, the context seems particularly favorable for a teacher incentives program: the level of teacher absence suggests that teacher effort was an issue in schools; there was relatively little scope for diverting teacher attention away from creativity and towards teaching to the test; and the short duration of the program made possible a design that did not encourage manipulation of the student body or the set of teachers in the school.

4.1 School Selection

The 50 schools in the program were selected from a group of 100 schools that had originally been selected by the Ministry of Education because they were considered to be particularly in need of assistance, but had not participated in an earlier World Bank program which provided textbooks to the schools judged to be in greatest need. These schools scored somewhat worse than average for the area before ICS began working with them. ICS had provided textbooks or modest grants to these schools before the inception of the teacher incentive program. Schools were numbered according to the year they received textbooks or grants from ICS, their geographic division, and their alphabetical order. Within each group, the odd numbered schools were chosen to participate in the teacher incentive program. By construction, the odd and even numbered schools were split in comparable proportions across Busia and Teso districts, geographic divisions within these districts, and across schools which received textbooks and grants in different years.

The 50 even-numbered schools that serve as the comparison group for this evaluation participated in another program that was designed to improve pre-schools by providing training, materials, and salary supplements conditional on pre-school teacher attendance. Pre-school teachers, unlike the primary school teachers, are semi-volunteers who are not hired or paid by the central Ministry of Education, but instead are hired locally by parents' committees and receive only contributions from parents, which are often irregular. Unlike primary school teachers, pre-school teachers often have no formal training and do not belong to the Kenya teachers union. Since pre-school and primary school teachers are quite distinct, and since in any case the pre-school program had little effect on performance in the pre-school classes, it seems unlikely that this program affected outcomes in grades 4 to 8 during the time period we examine.

5. Impact of Incentives on Teachers

This section examines the impact of the program on teacher behavior by comparing outcomes between treatment and comparison schools. It is worth noting, however, that interviews of teachers in treatment schools about their satisfaction with the program suggest that teachers in participating schools liked the program. In particular, in the middle of the second year of the incentives program, the headmaster and three other teachers were interviewed in each program school. All teachers interviewed supported the idea of motivating teachers by providing them with incentives. Most reported a change in school activities and teacher attitudes because of the program. 83% reported that prizes were justly awarded in 1998. 75% of teachers in program schools reported an increase in homework assignment due to the program, 67% reported an increase in cooperation among teachers, and 88% reported an increase in preps.

5.1 Teacher Assignment

The incentives program was designed so as not to provide incentives for teachers to join program schools. Only teachers who were already assigned to an incentive school as of March of the first year were eligible for a prize. However, the program could potentially have reduced the exit rate of teachers from the incentive schools.

In fact, the exit rate was not significantly different between program and comparison schools (Table 2, Columns 1 and 2).[14] The entry rate was higher in the incentive schools for the first year of the program, and lower for the second, although in neither case was the difference statistically significant (Table 2, Columns 3 and 4).[15]

We also considered the possibility that teachers in treatment schools in lower grades would attempt to transfer into teaching higher grades, even though this would not actually make them eligible for the program. There is no evidence of differential transfers across grades. In treatment schools 7.4% of teachers transfer from a non-incentive grade to an incentive grade during the program; in comparison schools 7.3% do.

---

[14] All regressions in this paper allow for school-level random effects to take account of the possibility that there may be correlation between error terms for students or teachers in the same school. Note that in the random effect regression framework the coefficient on the constant term is not exactly equal to the mean of the omitted category.

[15] The transfers include voluntary transfers due to family reasons (such as marriage) or involuntary transfers such as disciplinary actions against teachers or staff balancing needs (to replace teachers who retire, die, or move.)

5.2 Teacher Attendance

Teacher attendance was not affected by the incentive program. In the year prior to the establishment of the program, each of the 100 schools was subject to two random, unannounced visits at which 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.[16] For each teacher in each year, an attendance rate was computed as the proportion of visits during which the teacher was present. Note that teachers were recorded as present if they were at the school, even if they were not teaching when the visit took place. Following standard Intention-to-Treat (ITT) methodology, the sample included only those teachers who were assigned to program or comparison schools in year 0. Any teachers who changed schools between year 0 and year 1 or between year 1 and year 2 were classified with their initial schools.[17]

Prior to the program, schools that would later be selected to be program schools have slightly higher teacher attendance, although the difference was insignificant (Table 3, Column 1).[18] In year 1 of the program, teacher attendance was actually slightly lower in the incentive schools, and in year 2 the attendance was slightly higher in incentive schools (Table 3, columns 2 and 3), although both coefficients are insignificant and quite small.[19]

[16] Some visits did not take place, 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 based on visits to schools, rather than on official school logs, because school logs are often not filled out. However, school-log data also suggest no effect of the program on absence.

[17] This could only be done 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.

[18] The results here are robust to a specification in which each visit is treated as a binary opportunity for attendance and month of visit is controlled for.

[19] Results are also similar when lower primary school teachers are used as a control in a regression in which attendance of all the teachers in the schools is regressed on a dummy for the program, on whether they are a lower primary teacher or an upper primary teacher, and on an interaction term.

5.3 Homework Assignment

For a random subset of students in grades 4 to 8 for each school, information was collected from the students on whether they were assigned any homework on the previous day. In general, homework assignment was much more common in the higher grades. In 1997, 14% of grade 4 students report having homework assigned the previous day, versus 45% of grade 8 students. Treatment schools assigned slightly more homework than comparison schools prior to the program, although the difference is far from significant (Table 4, Column 1). After the launch of the program, treatment schools assigned slightly less homework, although the gap was insignificant both in levels and in differences (Columns 2 and 3).

5.4 Pedagogy

Teacher behavior was not significantly different between the incentive and comparison schools. Trained observers observed each teacher annually, spending one period in a class and recording teacher behavior on a number of measures, including both objective information about what the teacher was doing and subjective impressions about their energy level and caring for the students. We examined a wide variety of pedagogy measures, and the results for two objective measures (blackboard use and teaching aid use) and two subjective ones (teacher caring and energy) are presented here.

There was no significant difference in pedagogy between the incentive and comparison schools for any of the classroom observations prior to the program (Table 5, Column 1). We also find no significant difference during the intervention period between the two school groups in any of the pedagogical practices (Columns 2 and 3). The point estimates are close to zero for each observation type. The difference-in-difference estimates shown in the last two columns were computed at the school-grade level since it was not possible to match individual teachers across observations years. These estimates are also close to zero for every observation type.

5.5 Test Preparation Sessions

Incentive schools conducted more preps than comparison schools. Headmasters in each school provided information on whether there were any preps for grades 4 through 8 in six time periods during the year – the three school vacations (April, August and December) and out of school hours during each of the three terms.

Prior to the program, incentive schools were slightly less likely to offer preps (Table 6, Column 1), but after the introduction of the program, treatment schools started to conduct more preps (Columns 2 and 3). They were 4.2 percentage points more likely to conduct preps in the first year and 7.4 percentage points more likely in the second, with the latter estimate being significant at the 5% level. The results shown above are driven primarily by preps over vacations, as can be seen by the stronger results in the lower panel of Table 6.

6. The Impact of Incentives on Students

Consistent with the hypothesis that teachers responded to the program primarily by seeking to manipulate the variables determining prize allocation rather than by increasing efforts to promote long-run learning, the program had little impact on dropout and repetition rates, but increased student participation in exams. During the period the program was in place, student scores increased, significantly so on some test measures. There is some suggestive evidence that the effect was larger in the subjects more vulnerable to coaching. After the end of the program the effect on test scores did not persist. Students who had been in program schools during the program scored no higher than their counterparts who had been in comparison schools. Below we discuss the program impact on dropout and repetition (subsection 6.1), review the structure of

Kenyan exams (subsection 6.2), and discuss program impact on exam participation (subsection 6.3) and test scores (subsection 6.4).

6.1 Dropout and Repetition

As mentioned above, schools were penalized for those students who did not take the test. However, dropout and repetition rates in incentive schools were not significantly different from those in comparison schools. Dropout dummy variables were set equal to one if a student enrolled in the previous year did not continue schooling in the current year. Repetition dummy variables were set equal to one if the student repeated the same class in the following year. Dropout rates were insignificantly higher in treatment schools (Table 7, Columns 1 and 2), while the repetition rate was insignificantly lower in incentive schools (Table 7, Columns 3 and 4).

6.2 The Kenyan Exams

Incentives for teachers were based on their students' performance on the district exams, which are administered in seven subjects: English, Math, Science, Swahili, Geography-History-Christian Religion (G.H.CR.), Arts-Crafts-Music (A.C.M.), and Home Science-Business Education (HS.BE.) Students in grades 4 through 8 take these exams in October of each year. Participation is incomplete since students have to pay a fee of 120 Shillings (US $2) to participate in the exams. Since the rules of the incentive program stipulated that any student who did not take the district exams would be assigned lower scores than students could obtain by guessing, teachers in program schools had incentives to encourage their students to take the district exams.

In addition to the district exams, students also took KCPE and ICS exams, which were not tied to the teacher incentive program. Thus, these tests provide us with an independent way to assess the impact of the program.

The Kenyan primary school leaving exam (KCPE) is administered by the Ministry of Education each year to pupils completing grade 8. It determines what secondary school, if any, students attend. ICS also administered exams to students in grades 3 through 8.[20] In year 1, the ICS tests were administered in English, Math, and Science. In years 2 and 3, they were administered only in English and Math.[21] We have data on the district exam scores and the KCPE test scores from both intervention years (1 and 2) as well as the post-program year (3). Finally, we have information on the ICS tests for all participating schools for both of the intervention years.[22] We were unable to obtain the data for comparison schools in Teso District for year 1. Consequently, analysis of the district exam scores for year 1 is restricted to schools in Busia District.

Security is generally tight in Kenyan exams to prevent cheating. District exams were supervised by three to four teachers from a neighboring school. Because the KCPE exam determines the future scholastic paths of the eighth grade students who take it, these exams are even more strictly monitored and supervised.

In 1998, one case was identified in which the headmaster of a program school colluded with the teachers assigned to supervise the schools to allow cheating on the district exam. That

[20] For a complete description of the ICS tests and their administration, see Glewwe, Kremer, and Moulin (2001). These tests were also administered in 1996 and 1997.

[21] In year 0 and year 1, all three exams were fairly similar in format and content. Separate exams were given for each grade and the exams had a multiple-choice format. However, the ICS exams in year 2 and 3 were "multilevel," with the same test given to all students in grades 3 through 8. Easy questions in the beginning of the test could be answered by all students, including those in grade 3, while questions became progressively harder. The final questions were based on material seen only in the eighth grade. These exams also had a "fill in the blank," as opposed to a multiple choice format.

school was disqualified from the competition in 1998 but was allowed to participate in 1999. The scores from that school were not included in the analysis in 1998, but their scores were included in 1999.

6.3 Exam Participation

Exam participation is important both as an outcome in its own right and because differential exam participation could complicate the interpretation of test score differences between program and comparison schools. Exam participation rates were higher in program schools than in comparison schools for the district exams (on which the incentives were based), but were similar between treatment and comparison schools on the ICS and KCPE exams.

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

Baseline participation on the district exams was around 70%; on the ICS and KCPE exams it was around 85%. In year 1, participation in the district exams was higher by 6.0 percentage points in the incentive schools, a difference which is statistically significant at the 10% level (Table 8, Column 1). By year 2, participation was higher by 10.8 percentage points in the treatment schools, a significant difference 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 district exam was 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 2.2% lower in the incentive schools than in the comparison schools, though the difference was not significant. In contrast, the participation rates in the ICS and KCPE exams,

[22] In the post-program year, 27 of the 100 schools were involved in a de-worming project. This enabled us to collect

which were not linked to teacher incentives, were similar between the two school groups in either year (Columns 2 and 3).

Theoretically, efforts by teachers in treatment schools to increase exam participation could bias scores either upwards or downwards, but available evidence suggests the bias is likely downward. 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 dragged down average test scores, since poorer students are less likely to pay the fee to take the district 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 potentially biased upward. To get a sense of the direction of bias, we compared pre-scores of test takers in treatment and comparison schools. The additional students who took the tests in the incentive schools had lower pre-test scores, but not significantly so.

6.4 Test Scores

Test scores on the district exam are higher in the incentive schools during the years of the program (significantly higher in the second year). Scores on other exams are also higher in program schools during the duration of the program, but not significantly higher. In the post-program year, however, we see no persistence of the test score gains. This provides some support for the hypothesis that teachers are focusing primarily on extra coaching specifically for the test in question. On the other hand, there is little evidence of outright cheating, and we did not see any evidence that teaching effort actually decreased during the program, such as a decline in long-run test scores for students exposed to the program.

ICS test score data from that subset of schools for that program year.

We examine differences in test scores between the incentive and comparison schools using a random effect 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 school, grade, and subject random effects: random effects at the school level and at the level of individual subjects and grades within the school.

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

where k = English, Math, Science, Swahili, G.H.CR., A.C.M., and HS.BE.

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 which was eligible for teacher incentives) and 0 if not. 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. 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.

Because the units in which test scores are measured are arbitrary, for each subject and grade combination we normalize all test scores (including district, ICS, and KCPE tests) 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. Note that for a normal distribution, an increase of 0.1 standard deviation would move a student from the 50th percentile to the 54th. Since the district test exams were different for Busia and Teso Districts, the normalization of these tests was done separately for each district.

There is no significant difference in pre-program scores on the district exam between incentive and comparison schools for any subject or grade. Overall, prior to the program, treatment schools scored almost the same as comparison schools. Since grade 2 students were not given district exams in 1996, we used 1997 ICS tests as pre-tests where available 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 Tables 9 (for the district exam) and 10 (for the ICS and KCPE exams).[23] The difference estimates were calculated using the full sample, i.e. all the students in grades 4 through 8 who took the exams in either intervention year, while 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 at least one of the intervention years. As discussed, we restrict attention to only those students who were enrolled prior to the announcement of the program, as of February 1998 (year 1). We also restrict the sample to those students who did not repeat or dropout in any year, since students who repeated would be taking a different exam.

Averaging across all subjects and grades, the difference estimate for the district exam is insignificantly negative in the first year of the program (point estimate –0.04), but this could

[23] Normalized district test scores from year 0 (1996) were used as the KCPE pre-program scores, since the KCPE exam is taken by grade 8 students only.

potentially be due to the differential exam participation between treatment and comparison schools on the district exams. The difference-in-difference estimate is positive, although not significant. Both the difference and difference-in-difference estimate of the treatment effect in year 2 are significantly positive at 0.136 and 0.139 respectively. For the ICS exams, differences are not significant in either year, and the point estimates across years are similar, at around 0.085 (Table 10, Panel A). For the KCPE, the overall difference estimate in year 1 is 0.138 and the difference-in-difference is 0.104, with the former being significant at the 5% level (Table 10, Panel B). In year 2, the effect is stronger, with the difference estimate equal to 0.165, significant at the 5% level.

By year 3, students in the incentive schools had been exposed to the program for all of year 1 and year 2. If the increases in test scores in year 1 and year 2 of the program were due to increases in students' underlying long-term learning, then students in incentive schools should also have scored higher in year 3, after the program ended. However, estimates of the program effect in year 3 on the district and KCPE exams were close to zero.[24]

Since the teacher incentives were based on performance on the district tests only, data on ICS and KCPE tests can be used to check whether the difference in student outcomes between the incentive and comparison schools was due to factors specific to the district exams, such as outright cheating or altering the pool of students taking the test. Program schools scored insignificantly higher on the ICS test (Table 10, Panel A), and significantly higher on the KCPE (Table 10, Panel B) in both program years.

Breaking down the results by subject, the average effect for the two program years was strongest for the subject test on Geography, History, and Christian Religion, which is arguably the subject that involves the most memorization (Tables 9 and 10). In year 1, the difference-in-

[24] Data for the ICS exams in year 3 was only available for 27 of the schools – those that participated in a de-worming project that year. Point estimates are positive, but t-statistics are less than or equal to one.

difference estimate on G.H.CR was 0.205 for the district exam and 0.149 for the KCPE exam, with the former being significant at the 10% level. In year 2, the program impact on G.H.CR. scores is even stronger, with the difference-in-difference estimates being 0.341 for the District Exams and 0.336 for the KCPE exams. Both estimates are significant at the 5% level. The next biggest effects were for science and math, with no significant effect for other subjects. Arguably, GHCR is the subject with most memorization and 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. [25]

For the 1999 ICS test, we have item-level data on whether students had correct answers to individual questions in their English and math tests. Treatment students scored higher than comparison students on the later part of the test but not on the early part of the test. This may be because teachers in treatment schools taught students to better allocate time across sections of the test. (However, given that both tests were designed so as to increase in difficulty as the test progressed, another possibility is that the program induced increased teaching effort, but that this was most effective at raising scores on more difficult questions.)

Analysis of the item responses to detect cheating using techniques developed by Jacob and Levitt (2002) provide little evidence of suspicious strings of questions for which all students in the class got the question right, suggesting cheating was not widespread, although there was one instance in which cheating was discovered at a program school. Were cheating widespread we likely would have seen much larger test score differentials on the district exam (on which incentives were based) than on other exams. Although the estimated program impact on the

[25] 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 most creative thinking.

district exam is somewhat larger than on the ICS exams, it is lower than on the heavily-monitored KCPE exams, suggesting that cheating was not the main source of the program effect.

7. Time-path of Changes in Teacher Behavior

There is evidence that teachers learned over time how better 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. The test score gap between treatment and comparison schools was greater in the second year than in the first year.

8. Conclusion

Students in schools with a teacher incentive program in were more likely to take exams and had higher test scores in the short run. There is little evidence, however, that teachers responded to the program by taking steps to reduce dropouts or increasing effort on stimulating long-run learning. Teachers in program schools had no higher attendance rates or homework assignment rates. Pedagogy and student dropout rates were similar across schools. Instead, teachers in program schools increased test preparation activities and encouraged students enrolled in school to take the test. Following the end of the program, the test score difference between students who had attended treatment and comparison schools disappeared, consistent with a model in which teachers increased signaling effort but did not significantly increase effort to promote long-run learning. Similarly, dropout rates were no lower in program schools, but conditional on being enrolled in school, students were more likely to take exams linked to incentives.

It is worth noting several caveats (as well as caveats to these caveats). First, we cannot rule out the possibility that a larger incentive program or teacher-specific incentives would have induced not only increased test preparation, but also increased effort to improve underlying learning. However, at up to 40% of monthly income, the incentives were comparable in magnitude to those in most U.S. programs. Although the bonuses were a small percentage of yearly salary so the implied increase in daily wages was modest, if teachers chose attendance optimally prior to the program given their intrinsic motivation to teach, 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. Moreover, while larger incentives might induce more effort by teachers, they could also have induced effort at counter-productive signaling, for example through cheating on tests or forcing weak students to drop out. They would also force teachers to bear more risk. Some 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 either already have teaching jobs or who have the academic qualifications to enter teacher training colleges (but not universities) will select out of the profession. 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 would have had more incentive to make long-run investments in learning.[26] 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 the test was such that learning has only a small impact on test scores. Alternative tests that better measure long-run learning might potentially create better incentives. However, since the incentives provided by ICS were based on the official government of Kenya exams, which in turn are based on the official curriculum, any incentive program based around these exams is likely to run into similar difficulties.

However, whatever the problems with teacher incentives, the status quo - with its 20% teacher absence rate - is inadequate, so it seems worth exploring other alternatives.

One strategy for improving incentives would be to attach incentives to measurable inputs, such as teacher attendance. In many countries, teachers’ pay is not linked to pupils' test scores, but a teacher who is absent 20% of the time would typically face some sort of disciplinary sanction. One problem with this approach is that attaching incentives to inputs rather than

[26] In practice, many teachers transfer between schools.

outputs can lead to undersupply of badly measured inputs. For example, since random audits can verify whether a teacher is in school, but not whether a teacher is in class, teachers might come to school, but not come to class, since outside inspectors cannot easily monitor presence in class. While it would be prohibitively expensive for outside inspectors to visit schools regularly enough to keep track of teachers' presence for incentive purposes, headmasters could keep these records. However, headmasters do not currently have incentives to risk getting into a confrontation with teachers over absences. One indication of this comes from an evaluation of the program the NGO implemented at pre-schools. (Preschool teachers are not formal Ministry of Education employees, but instead are locally hired and paid by parents.) In this program, school committees were given funds with which to provide bonuses to pre-school teachers, conditional on their not missing more than a specified number of days of class. If the funds were not spent on the teachers, school committees could use them for other pre-school purposes, so headmasters and school committees arguably had a strong incentive to monitor pre-school teachers. Preliminary work suggests that the program yielded little if any improvement in absence rates, and it is clear that headmasters did not strictly enforce the rules requiring teacher attendance as a condition of the bonus being provided. More broadly, headmasters are already required to keep log books of teacher attendance, and inspectors are supposed to monitor them but log books are often not even filled out.

Ultimately, an analysis of the problem must turn to the political economy of education. Given that Kenya's centralized education system is not producing adequate incentives, it may be worth considering decentralizing control over teachers to local school committees or allowing parents to choose schools and tying school finance more tightly to their decisions.[27]

[27] Since students' placement in secondary school depends on performance on the primary-school leaving exam, local communities and parents would share some of the same incentives to focus on test preparation as teachers (see Acemoglu, Kremer, and Mian 2002). Nonetheless, since teachers transfer schools fairly frequently, they likely have greater incentives than parents to focus on the short run.

References

Acemoglu, Daron, Kremer, Michael R., and Atif Mian (2002), "Markets, Firms and Governments." Unpublished.

American Federation of Teachers, Teacher Salary Survey Archives at http://www.aft.org/research/survey00/salarysurvey00.pdf

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, Kremer, Muraldhiran, and Rogers (2003), Absence Among Service Providers: A Preliminary Note. work in progress, World Bank.

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.

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 W, Kremer, Michael R. and Sylvie Moulin (2001), "Textbooks and Test Scores: Evidence from a Prospective Evaluation in Kenya," mimeo.

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.

Harbison, Ralph W. and Eric A. Hanushek (1992), *Educational Performance of the Poor: Lessons from Rural Northeast Brazil*, NY: Oxford University Press.

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 .

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 Charles S. Benson et al (1994), *Making Schools Work: Improving performance and controlling costs*. Brookings Institution. Washington, D.C.

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 (2002), "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." Unpublished manuscript.

Kingdon, Geeta and Francis Teal (2002), "Does performance related pay for teachers improve student performance? Some evidence from India," mimeo.

Lavy, Victor (2002a), "Paying for Performance: The Effect of Teachers' Financial Outcomes on Students' Scholastic Outcomes." mimeo.

Lavy, Victor (2002b), "Evaluating the Effect of Teacher Group Performance Incentives on Students Achievements.", *Journal of Political Economy,* vol. 110, no. 6, pp. 1286-1318.

Lockheed, Marlaine E. and Adriaan M. Verspoor (1991), *Improving Primary Education in Developing Countries*, NY: Oxford University Press.

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.

Table 1: Concentration of Teacher Absences

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

Table 2: Program Effect on Teacher Entry and Exit

	(1)	(2)	(3)	(4)
Dependent Variable:	Exit Current School (0/1)		Enter New School (0/1)	
	Exit in 1997	Exit in 1998	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; regressions include school-level random effects.
* significant at 10%; ** significant at 5%;
For exit regressions, incentive/non-incentive refers to the originating school; for entry regressions incentive/non-incentive refers to the destination school. The unit of observation in all regressions is the teacher.

Table 3: Program Effect on Teacher Attendance

	(1)	(2)	(3)	(4)	(5)
Dependent Variable:	Teacher Attendance Percentage			Attendance Differences *(Attendance $_{program\ year}$ – Attendance $_{pre\text{-}program\ year}$)*	
	Year 0	Year 1	Year 2	Year 1 - Year 0	Year 2 – Year 0
Incentive School	0.012	-0.008	-0.011	-0.007	-0.063
	(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; regressions 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 Effect on Homework Assignment

	(1)	(2)	(3)	(4)	(5)
Dependent Variable:	Homework Assignment (0/1)			Homework Differences *(Homework $_{program\ year}$ – Homework $_{pre\text{-}program\ year}$)*	
	Year 0	Year 1	Year 2	Year 1 - Year 0	Year 2 – Year 0
Incentive School	0.012 (0.042)	-0.052 (0.045)	-0.009 (0.047)	-0.092 (0.055)*	-0.042 (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; regressions include school-level random effects.
* significant at 10%; ** significant at 5%;
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 Effects on Pedagogy

	(1)	(2)	(3)	(4)	(5)
	Year 0	**Year 1**	**Year 2**	**Year 1 - Year 0**	**Year 2 – Year 0**
Panel A *Dependent Variable:* **Use of Blackboard (0/1)**					
Incentive School	0.018	-0.032	0.038	-0.051	0.078
	(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 B *Dependent Variable:* **Use Teaching Aid (0/1)**					
Incentive School	-0.026	-0.006	0.012	0.025	0.052
	(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 C *Dependent Variable:* **Teacher Caring (1 to 5: 1=very caring)**					
Incentive School	-0.080	-0.065	-0.051	0.052	-0.058
	(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 D *Dependent Variable:* **Teacher Energy (1 to 5: 1=energetic)**					
Incentive School	-0.030	-0.041	0.164	0.050	0.070
	(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

Notes:
Standard errors in parentheses; regressions 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. There are fewer observations in 1999 because only one class per school/grade was observed that year.

Table 6: Program Effect on Preparations

	(1)	(2)	(3)	(4)	(5)
Dependent Variable:		Preparations (0/1)		Preparation Differences *(Preparation $_{program\ year}$ – Preparation $_{pre\text{-}program\ year}$)*	
	Year 0	**Year 1**	**Year 2**	**Year 1 - Year 0**	**Year 2 – Year 0**
Preparations (Vacation and During School)					
Incentive School	-0.007	0.042	0.074	0.049	0.081
	(0.044)	(0.037)	(0.034)**	(0.042)	(0.047)*
Grade	0.155	0.135	0.103	-0.021	-0.052
	(0.009)***	(0.007)***	(0.007)***	(0.009)**	(0.009)***
August Holiday (0/1)	-0.020	0.058	-0.122	0.078	-0.102
	(0.016)	(0.030)*	(0.034)***	(0.037)**	(0.036)***
December Holiday (0/1)	-0.370	-0.452	-0.534	-0.082	-0.164
	(0.025)***	(0.029)***	(0.030)***	(0.031)**	(0.036)***
Term Visit 1 (0/1)	0.094	0.126	0.130	0.032	0.036
	(0.035)***	(0.040)***	(0.038)***	(0.035)	(0.052)
Term Visit 2 (0/1)	0.094	0.168	0.282	0.074	0.188
	(0.035)***	(0.041)***	(0.040)***	(0.052)	(0.052)***
Term Visit 3 (0/1)	0.096	0.158	0.242	0.062	0.146
	(0.035)***	(0.040)***	(0.043)***	(0.051)	(0.055)***
Constant	-0.502	-0.372	-0.121	0.130	0.381
	(0.064)***	(0.053)***	(0.052)**	(0.064)**	(0.064)***
Observations	3000	3000	3000	3000	3000
Vacation Preparations					
Incentive School	0.035	0.089	0.091	0.055	0.056
	(0.034)	(0.031)***	(0.035)**	(0.038)	(0.046)
Grade	0.156	0.139	0.118	-0.017	-0.038
	(0.008)***	(0.006)***	(0.007)***	(0.008)*	(0.009)***
August Holiday (0/1)	-0.020	0.058	-0.122	0.078	-0.102
	(0.016)	(0.030)*	(0.034)***	(0.037)**	(0.036)***
December Holiday (0/1)	-0.370	-0.452	-0.534	-0.082	-0.164
	(0.025)***	(0.029)***	(0.030)***	(0.031)**	(0.036)***
Constant	-0.527	-0.425	-0.219	0.103	0.308
	(0.049)***	(0.049)***	(0.054)***	(0.057)*	(0.059)***
Observations	1500	1500	1500	1500	1500

Notes:
Standard errors in parentheses; regressions 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. Rates for given time periods are reported compared to the omitted time period, the April holiday.

Table 7: Program Effect on Dropout and Repetition Rates

	(1)	(2)	(3)	(4)
	Year 1		**Year 2**	
	Dropout	**Repetition**	**Dropout**	**Repetition**
Incentive School	0.004	-0.012	0.002	-0.010
	(0.010)	(0.019)	(0.015)	(0.026)
Male (0/1)	0.021	0.010	0.018	-0.009
	(0.005)**	(0.008)	(0.005)**	(0.008)
Constant	0.073	0.295	0.116	0.256
	(0.007)**	(0.014)**	(0.011)**	(0.018)**
Observations	14153	12686	14545	12671

Notes:
Standard errors in parentheses; regressions include school-level random effects.
* significant at 10%; ** significant at 5%;
Each observation represents an upper primary school student.

Table 8: Program Effects on Participation in Exams

	(1) **Year 1**	(2) **Year 2**	(3) **Year 3 (Post-Program)**
	Panel A *Dependent Variable:* **Take District Exam (0/1)**		
Incentive School	0.060	0.108	-0.022
	(0.033)*	(0.028)**	(0.034)
Male (0/1)	0.015	-0.011	0.007
	(0.008)*	(0.009)	(0.016)
Grade	0.058	0.036	0.010
	(0.003)**	(0.004)**	(0.010)
Constant	0.389	0.525	0.560
	(0.029)**	(0.034)**	(0.069)**
Observations	10690	7158	3642
	Panel B *Dependent Variable:* **Take ICS Exam (0/1)**		
Incentive School	0.005	0.031	
	(0.007)	(0.024)	
Male (0/1)	0.004	0.000	
	(0.004)	(0.007)	
Grade	0.010	0.009	
	(0.001)**	(0.003)**	
Constant	0.880	0.822	
	(0.010)**	(0.027)**	
Observations	14397	7158	
	Panel C *Dependent Variable:* **Take KCPE Exam (0/1)**		
Incentive School	-0.014	0.026	0.021
	(0.047)	(0.017)	(0.039)
Male (0/1)	-0.000	0.001	0.013
	(0.019)	(0.015)	(0.029)
Constant	0.779	0.912	0.813
	(0.034)**	(0.014)**	(0.029)**
Observations	1624	1265	681

Note: Standard errors in parentheses; school-level random effects included.
* significant at 10%; ** significant at 5%;
District test data was not available for Teso District in 1998.
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 repeat or drop out in any year.

Table 9: Program Effect on Test Scores by Subject (District Exam)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	**Test Scores (Standardized relative to comparison schools)**			**Test Score Differences** ***Test Score$_{program\ year}$ – Test Score$_{pre\text{-}program\ year}$***		
	Year 1	**Year 2**	**Year 3**	**Year 1 – Year 0**	**Year 2 – Year 0**	**Year 3 – Year 0**
English	-0.059	0.094	0.017	-0.024	-0.003	-0.091
	(0.107)	(0.094)	(0.112)	(0.071)	(0.086)	(0.122)
Math	0.058	0.099	-0.077	0.076	0.150**	-0.106
	(0.089)	(0.084)	(0.089)	(0.054)	(0.064)	(0.089)
Science	0.015	0.155	0.121	0.050	0.206*	0.194
	(0.091)	(0.102)	(0.115)	(0.076)	(0.094)	(0.128)
Swahili	-0.052	0.105	0.091	0.023	0.019	-0.134
	(0.093)	(0.072)	(0.084)	(0.083)	(0.094)	(0.221)
G.H.CR.	-0.039	0.202**	0.055	0.205*	0.341**	-0.021
	(0.089)	(0.097)	(0.105)	(0.107)	(0.129)	(0.262)
A.C.M.	-0.007	0.010	-0.049	0.116	0.108	-0.218
	(0.096)	(0.092)	(0.102)	(0.121)	(0.154)	(0.249)
HS. BE.	0.049	0.073	-0.079	0.073	0.167	-1.232**
	(0.092)	(0.107)	(0.113)	(0.161)	(0.196)	(0.525)
All Subjects &	-0.040	0.136*	-0.087	0.054	0.139**	-0.008
Grades	(0.079)	(0.077)	(0.083)	(0.054)	(0.065)	(0.084)
Observations	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 district 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 and 3 are limited to pupils who were enrolled in year 1 and did not repeat or drop out. Columns 4, 5, 6 impose the additional restriction that a pre-test score is available.
7,848 students (grades 4 to 8) took at least one district 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 will not have pre-test scores. So, for example, in the post-program year students in 4th and 5th grade will not have pre-test scores.

Table 10: Program Effect on Test Scores by Subject (Non-Incentive Tests)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	**Test Scores (Standardized relative to comparison schools)**			**Test Score Differences** ***Test Score$_{program\ year}$ – Test Score$_{pre\text{-}program\ year}$***		
	Year 1	**Year 2**	**Year 3**	**Year 1 – Year 0**	**Year 2 – Year 0**	**Year 3 – Year 0**
Panel A						
***Dependent Variable:* ICS Subject Test Scores**						
English	0.077	0.077		0.001	0.031	
	(0.090)	(0.138)		(0.040)	(0.099)	
Math	0.053	0.069		-0.042	-0.009	
	(0.074)	(0.074)		(0.041)	(0.058)	
Science	0.129			0.091**		
	(0.082)			(0.043)		
All Subjects & Grades	0.089	0.083		0.017	0.016	
	(0.079)	(0.090)		(0.033)	(0.063)	
Observations	39,510	12,996		32,993	10,512	
Panel B						
***Dependent Variable:* KCPE Test Scores**						
English	0.116	0.103	0.002	-0.045	-0.120	-0.130
	(0.094)	(0.126)	(0.125)	(0.105)	(0.137)	(0.192)
Math	0.166	0.120	0.044	0.123	0.145	0.071
	(0.102)	(0.099)	(0.124)	(0.103)	(0.124)	(0.185)
Science	0.132	0.113	0.040	0.200*	0.189	0.156
	(0.098)	(0.114)	(0.142)	(0.114)	(0.142)	(0.194)
Swahili	0.212*	0.226**	-0.126	0.107	0.081	-0.473*
	(0.121)	(0.112)	(0.131)	(0.126)	(0.123)	(0.270)
G.H. CR	0.167*	0.257**	0.004	0.149	0.336**	-0.053
	(0.088)	(0.115)	(0.120)	(0.108)	(0.136)	(0.253)
A.C.M.	0.054	0.169	-0.027	0.139	0.169	-0.058
	(0.098)	(0.117)	(0.133)	(0.121)	(0.168)	(0.267)
HS. BE.	0.125	0.154	0.052	-0.008	0.268	-1.285*
	(0.091)	(0.111)	(0.128)	(0.146)	(0.175)	(0.658)
All Subjects & Grades	0.138*	0.165*	-0.009	0.104	0.152	-0.006
	(0.074)	(0.090)	(0.101)	(0.080)	(0.097)	(0.138)
Observations	10,430	8,427	4,053	7,152	5,247	1,505

Note: Standard errors in parenthesis; regressions include school-level random effects.
* significant at 10%; ** significant at 5%;
Year 3 ICS tests were given only in 27 schools so scores are not reported; 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.
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, 5, and 6 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 ICS exams in year 1. Of these, 11,298 had pre-test scores from year 0, in the form of normalized district exam scores from year 0. 15,647 students took at least one ICS exam in year 2, of which 8,638 had pre-test scores from year 0. 1,490 eight 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.

Figure 1
Three Models of Teacher Absence Distribution