American Economic Journal: Applied Economics 2 (July 2010): 205–227
http://www.aeaweb.org/articles.php?doi=10.1257/app.2.3.205

Teacher Incentives†

By Paul Glewwe, Nauman Ilias, and Michael Kremer*

We analyze a randomized trial of a program that rewarded Kenyan primary school teachers based on student test scores, with penalties for students not taking the exams. Scores increased on the formula used to reward teachers, and program school students scored higher on the exams linked to teacher incentives. Yet most of the gains were focused on the teacher reward formula. The dropout rate was unchanged. Instead, exam participation increased among enrolled students. Test scores increased on exams linked to the incentives, but not on other, unrelated exams. Teacher attendance and homework assignment were unaffected, but test preparation sessions increased. (*JEL* I21, I28, J13, O15)

Incentives for public school teachers are weak in many countries. Teacher absence is one symptom. A study of five developing countries found a 19 percent teacher absence rate (Nazmul Chaudhury et al. 2006). Many policies have been proposed to address weak incentives, including rewards for teacher attendance, adjusting teacher salaries based on students' exam scores, voucher programs, and increased community oversight. This paper examines a Kenyan program that rewarded teachers based on students' exam scores, with penalties for students missing the exam.

The program changed teacher behavior, particularly in the second year of the program, after teachers had had time to learn how it worked. Scores on the formula used to reward teachers were substantially higher in program schools.

Yet while there is evidence of narrow gains, that is gains on outcomes that were the focus of the incentives, there is less evidence for gains in other, broader measures of the stock of student human capital. In some cases, point estimates of effects are very close to zero, while in other cases they are positive, but fairly small and statistically insignificant. Students in program schools were more likely to take exams conditional on enrollment, but the dropout rate did not fall. A decomposition analysis suggests that two-thirds of the increase on the formula used to reward teachers is due to higher exam participation. Test scores increased on the exams linked to the incentives, but point estimates of increased scores on exams that were not linked to the incentives

* Glewwe: University of Minnesota, Department of Applied Economics, 1994 Buford Ave., St. Paul, MN 55108 (e-mail: pglewwe@umn.edu); Ilias: Compass Lexecon, 1101 K Street NW, Washington, DC 20005 (e-mail: nilias@compasslexecon.com); Kremer: Harvard University, Department of Economics, Littauer Center M20, Cambridge, MA 02138, Brookings Institution, Center for Global Development, and National Bureau of Economic Research (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 and Bijie Ren for outstanding research assistance in the United States. We thank the World Bank and the MacArthur Foundation for financial support.

† To comment on this article in the online discussion forum, or to view additional materials, visit the articles page at http://www.aeaweb.org/articles.php?doi=10.1257/app.2.3.205.

are small and statistically insignificant. Test score gains on multiple-choice questions are significantly greater than on fill-in-the-blank questions, suggesting that program school students learned strategies for answering multiple-choice questions. Gains on the formula used to reward teachers did not persist beyond the life of the program, and although point estimates for test score effects remain positive after the program ended, they are smaller and statistically insignificant.

Consistent with the hypothesis that teachers responded to the program fairly narrowly, teachers in program schools were no more likely to be at school or assign homework than teachers in the comparison schools. Instead, headmasters reported more test preparation sessions in program schools, especially in the second year of the program.

The paper is organized as follows. Section I describes primary education in Kenya and the teacher incentive program. Section II discusses the data and our analytical framework. Sections III and IV report program impacts on students and teachers, respectively. Section V compares our results to the existing literature and discusses policy implications.

I. Background and Program Description

A. *Teacher Incentives in Kenya*

Decisions regarding hiring, firing, and transferring teachers in Kenya have long been made centrally by Kenya's Ministry of Education. Hiring, promotion, and salary increments depend mostly on education and experience, rather than job performance. Teachers have strong civil service and union protection, and are difficult to fire. In rare cases, poorly performing teachers are transferred to undesirable locations.

Random visits to comparison schools suggest that the teacher absence rate in our sample is about 20 percent. We estimate that the median teacher is absent 14 to 19 percent of the time, suggesting that teacher absences are widely distributed, rather than concentrated among a few "ghost" teachers. (See Glewwe, Ilias, and Kremer 2003 for details of the estimation procedure.)

While high absence rates could reflect an optimal contract, so that teachers can tend their farms at certain times of the year, we think it more likely that they reflect reluctance of school headmasters to bear the costs of enforcing absence rules. Personal gains from such enforcement are minimal, as headmasters do not bear the cost of absence, but do bear the cost of dealing with the aggrieved teachers and their union, and of performing the bureaucratic steps required to sanction chronically absent teachers. Those costs could be large since Kenyan teachers have a powerful union, as seen by the fact that they are paid five times the per capita gross domestic product (GDP). Moreover, if high teacher absences reflect an implicit contract that lets teachers pursue other activities, they should be scheduled in a predictable way, so that students need not come to school on days when teachers are absent, such as during peak agricultural times. Yet we find few patterns in teacher absences, and school visits often reveal pupils left unsupervised because their teachers are absent. In addition, teachers are often in school, but not in their classrooms. Teachers were absent from school 20 percent of the time in our data, yet trained classroom observers found teachers absent from class 27 percent of the time. Even when they are in class, they usually arrive late. Only a small percentage of teachers were in the classroom when class officially started. Casual observation reveals many teachers sipping tea in the staff room during class time.

Kenyan schools have parent committees that provide some incentives to teachers, but usually these are weak. They sometimes provide teachers gifts if schools do well on national exams. Parents sometimes protest against badly behaved teachers, pressuring the Ministry of Education to transfer them. Yet most parent committees do not provide bonuses to teachers, and they appeal to national authorities only in extreme situations.

To the extent that teachers face incentives, they stem from Kenya's national exams. Results on the national primary school leaving exam (KCPE) determine what secondary schools, if any, enroll graduating primary school students. KCPE results are front page news and are often posted in headmasters' offices. Many districts conduct their own exams, closely modeled on the KCPE, called "mocks," to prepare their pupils for the KCPE.

In some schools, teachers offer "preps" during evenings, weekends, and vacation periods to prepare students for the KCPE and district (mock) exams. Parents may pay teachers for these sessions, but, in our sample of rural schools, the amounts are small, usually only 10–20 Kenyan shillings (KES) (US $0.16 to $0.33) per school term, and students who cannot pay are not excluded. Time spent in preps varies widely, but a typical amount is 5 hours per week during the term, and 25 hours per week during vacation periods (5 hours per day for 5 days). Evidence from the United States suggests that test preparation classes can raise scores. Although most US college admission tests try to measure aptitude not achievement, and so should be hard to prepare for, studies (Robert L. Bangert-Drowns, James A. Kulik, and Chen-Li C. Kulik 1983) show gains of 0.15 to 0.40 standard deviations from test preparation. We have no experimental evidence on the impact of prep sessions on test scores in Kenya, but a simple regression of test scores on teacher attendance and prep sessions suggests that the marginal effect of prep sessions is strongly positive and much higher than the marginal effect of teacher attendance (see Glewwe, Ilias, and Kremer 2003).

B. *Program Description*

We evaluate a program conducted by International Child Support (ICS), a Dutch non-governmental organization (NGO) in the Busia and Teso districts of Western Kenya. The program provided gifts to teachers and headmasters in schools where students scored well on district exams (students who did not take the exam were assigned low scores). The program provided in-kind prizes, rather than cash, since this was seen as more culturally acceptable in Kenya. Discussions at schools revealed that teachers valued the prizes.

To promote teacher cooperation within schools, and avoid giving teachers incentives to harm each other's work, prizes were awarded based on the average performance of all students in the school who were in grades 4–8, and, thus, were eligible to take the district exams. Education experts are often more supportive of school-based, rather than teacher-based, incentives (Craig E. Richards and Tian Ming Sheu 1992; Eric A. Hanushek 1996). A potential disadvantage of school level-prizes is that they may lead to free riding by teachers. Yet the typical school in our data had only 200 students and 12 teachers, only half of whom taught in the upper grades, so coordination within schools seems feasible. Teachers are in a repeated game with each other and can, for example, observe each other's daily attendance. Headmasters could also use their powers to reward and punish teachers through their control of teaching assignments, for example, to help enforce cooperative solutions.

All teachers who taught grades 4–8 were eligible for prizes. Teachers of lower grades received a lantern whether or not they belonged to a winning school, since no district-wide exams existed for those grades. Winning schools also received a briefcase for the headmaster, a wall clock, a time keeping clock, and a bell.

ICS gave one set of prizes based on absolute performance and another set based on improvement relative to baseline performance. Since government (district) exam results were unavailable for 1997, 1996 scores were used as the baseline. In each category, three first prizes (a suit worth about $51); three second prizes (plates, glasses, and cutlery worth about $43); three third prizes (a tea set worth about $34); and three fourth prizes (bed linens and a blanket worth about $26) were awarded. Schools could win only one type of prize, so 24 of the 50 program schools received prizes, and most teachers should have felt that they had a chance to win.[1] Prize values ranged from 21 to 43 percent of the typical teacher's monthly salary, similar to other merit pay programs.[2]

The ICS incentives were designed to reduce dropping out. All students enrolled at the beginning of the program were included in the formula used to award prizes. Those not taking the test were assigned very low scores. Not taking the English essay test yielded a score of zero. Not taking the multiple choice tests in other subjects led to scores of 15, below the likely guessing score of 25.

To dissuade schools from recruiting good students for their exam, only students enrolled when the program started (February 1998) were used to calculate mean scores. To discourage teachers from moving to incentive schools to compete for prizes, eligibility was limited to teachers employed (in any grade) in program schools in March 1998. Teacher exit and entry were not significantly different between program and comparison schools, nor is there evidence of differential reassignment of teachers within schools (see Glewwe, Ilias, and Kremer 2003).

We conducted a survey of the headmaster and three teachers in each program school in year 2 to obtain their views of the program. All teachers supported motivating teachers by giving them incentives. Eighty-three percent said that the prizes were justly awarded, and 67 percent reported more teacher cooperation. However, this survey may have been subject to social desirability bias since the questionnaire

[1] Busia and Teso districts had separate district exams, so prizes were offered separately in each district in proportion to the number of schools in those districts.

[2] A Dallas pay program, also based on school-wide performance, awarded $1000 annual bonuses, equivalent to about 39 percent of the average monthly salary of Texas teachers (Charles T. Clotfelter and Helen F. Ladd 1996; American Federation of Teachers 1997). A 1999 Rhode Island program awarded $1,000 annual bonuses (Lynn Olsen 1999; American Federation of Teachers 2000). Programs in Colorado awarded 10–50 percent of a teacher's monthly salary (Wendy Wyman and Michael Allen 2001). In Israel, the bonuses examined by Victor Lavy (2002) were 10–40 percent of the average teacher's monthly salary.

was framed as soliciting feedback from the teachers on the incentive program. One piece of evidence for such bias is that 75 percent of teachers interviewed said they had increased homework assignment in response to the program, but student reports (see Section IV) suggest no difference between program and comparison schools in homework assignment.

C. *School Selection*

In February 1998, 50 schools were offered the program, and all accepted. Initially, the program was for one year, but it was later extended another year. Prizes were awarded at a ceremony at the end of the school year. Henceforth, we denote the last pre-program year (either 1996 or 1997, depending on the type of data) as year zero, the first (1998) and second (1999) years of the program as years one and two, and the post-program year (2000) as year three.

The 50 program schools were randomly selected from 100 schools designated by the Ministry of Education as particularly in need of assistance, but that did not participate in an earlier World Bank textbook program. These 100 schools scored below the district average before ICS assisted them. ICS had provided textbooks or modest grants to 75 of these schools before or during the teacher incentive program as set by random assignment into four groups of 25 schools. By design, schools selected and not selected for the teacher incentives program were divided into equal proportions within Busia and Teso districts, within the geographic divisions in each district, and by whether they had received textbooks or grants earlier. It is unlikely that ICS assistance to the 75 schools seriously affects external validity, since that assistance was small relative to overall school budgets, and the earlier programs had little impact (see Glewwe, Kremer, and Sylvie Moulin 2009). Moreover, many NGOs assist Kenyan schools, and while these 100 schools received more support than average, they were not in the upper tail of the distribution.

The 50 comparison group schools also participated in a program that provided preschool teachers with training, materials, and (conditional on teacher attendance) higher salaries. Unlike primary school teachers, preschool teachers are semi-volunteers, with little formal training, who are hired by parents' committees, not the Ministry of Education. Their pay is from the contributions of parents, which are often irregular. That program did not affect the performance of preschool pupils, so it is unlikely that it affected grade 4–8 outcomes in the time period we study. That program's funds went to preschool teachers or to supplies the preschools would not have bought without the program, and so are unlikely to have leaked to grades 4–8. Lastly, preschools are run locally and are administratively separate from Ministry of Education schools, with separate financial accounts.

II. Analytical Framework, Outcome Measures, and Data

Both proponents and skeptics generally agree that teacher incentives change teacher behavior. They disagree on whether teachers respond by promoting broad human capital acquisition or by narrowly focusing on skills and actions that raise scores on the formulas used to reward teachers. Concerns about narrowly targeted gains are heightened if those gains are focused on realms where labor market rewards are due to signaling rather than human capital acquisition. For example, learning test-taking techniques presumably raises the chance that a student will obtain a place in secondary school, but does not directly affect economic output. Such gains are at others' expense.

As shown in Bengt Holmstrom and Paul Milgrom (1991) and George Baker (2002), and modeled in a previous version of this paper (Glewwe, Ilias, and Kremer 2003), increased rewards for measurable outcomes can lead to either increased or decreased effort on other, unobserved outcomes, depending on whether different types of effort are complements or substitutes in the production of those outcomes. Theoretically, teachers could narrowly direct effort at increasing scores on the formula used to determine teacher rewards at the expense of effort aimed at broader, longer term increases in their students' human capital; increase effort aimed at raising those broader, longer term increases; or take any of a continuum of actions in between.

We use several indicators to assess the extent to which the program affected the formula used to reward teachers as well as other, broader measures of human capital. The first is the score on that formula, which reveals whether the teachers responded to the program's incentives. The second is participation in government exams, the only exams used in the formula to reward teachers, raising that participation is a narrow action that increases scores on the formula.[3] Third are dropout rates, which should reflect important elements of broader human capital acquisition. Fourth, we examine whether the program affects scores on exams linked to the incentives, and whether it affects scores on exams with a different format that were not linked to the incentives. Fifth, we check whether the program disproportionately affects scores on exams with formats that are amenable to coaching (e.g., for multiple choice exams—the format of all government exams—students can be coached to guess instead of leaving blanks). Sixth, we investigate the pattern of score changes across subjects. Seventh, we examine outcomes not only during the life of the program, but also beyond it. Lastly, we directly observe teachers' behavior. Some actions, such as attendance or assigning homework, should broadly raise students' human capital, while others, such as extra prep sessions, are more likely to have a narrow signaling component.

Note that parents and students may value narrow signals of student learning. Certain rents (e.g., high-paying government jobs or rationed places in secondary education) may heavily depend on test scores. Thus, higher test scores can benefit pupils as well as teachers. Yet if such signaling effort raises only test scores, and not underlying human capital, it is socially wasteful.

The data used in this paper were collected from 100 Kenyan primary schools from 1996 to 2000. The data include many education outcomes and related variables, including all the indicators discussed in the previous paragraphs. For more details, see Glewwe, Ilias, and Kremer (2003). The following paragraphs discuss aspects of the data that merit particular attention.

[3] The 32 Teso schools do not have government exam data for year 1 because, to reduce parents' costs, Teso did not offer exams. Thus, analysis of government exams in year 1 is limited to Busia, which had the other 68 schools.

To obtain independent evidence on the program's impact that is not directly tied to the incentives, and to avoid differential attrition problems that may affect the government exams, students were also given free exams prepared by the NGO (ICS), that had no link to the teacher incentive program. Scores on government exams are quite low, so the NGO tests were designed to measure a wider range of student performance and were easier for the typical student. In years zero and one, separate NGO exams were given for each grade. All had a multiple-choice format. To facilitate comparisons across grades, given high repetition rates, the year 2 NGO exams were "multilevel." The same tests were given to all grade 3–8 students. The first questions were easy enough to be answered by all students in those grades, but later questions were progressively harder. Also, the NGO exam questions in year 2 were mostly "fill-in-the-blank," not multiple choice. The NGO tests were offered to grades 3–8 in all 100 schools for years zero, one, and two.[4]

The units of any test score variable are arbitrary, so all scores are normalized for each subject-grade combination, year, and district by subtracting the mean score in comparison schools and dividing by the associated standard deviation for those schools.[5] Thus, a normalized score of 0.1 is 0.1 standard deviations above the comparison school mean. Note that a 0.1 standard deviation increase in a normal distribution moves a student from the fiftieth percentile to the fifty-fourth percentile.

Almost all government exams have a multiple-choice format. Students took those tests for the grade they were in, not the grade they would have been in had they not repeated after the program began. To compare repeaters to nonrepeaters in years two and three, the same test was given to students in adjacent grades in nonsample schools. Students scored about a standard deviation lower on tests designed for the next grade, so we use a one standard deviation correction factor to compute the scores repeaters would have received had they not repeated. As seen below, the differences in dropout and repetition across incentive and comparison schools were insignificant.

Information on prep classes in grades 4–8 was obtained from school headmasters for six time periods: each of the three terms (outside of normal school hours) and the three vacation periods (April, August, and December). In year zero, before the program began, all 100 schools were subject to two random, unannounced visits where the present/absent status of each grade 4–8 teacher was recorded. Five such visits were made in year 1, and three in year 2.[6] In each year, for each teacher, an attendance rate is defined as the proportion of visits for which a teacher was present at school, even if he or she was not teaching in class. The sample included only those

[4] In years zero and one, the NGO administered English, math, and science tests. In year 2, only English and math tests were used. For further description of the NGO tests, see Glewwe, Kremer, and Moulin (2009).

[5] The test score regressions in this paper were also estimated separately for Busia and Teso. In only 1 of 14 regressions did the program impact differ across those districts. This difference was significant only at the 10 percent level.

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

Table 1—Descriptive Statistics

	Teacher incentive schools		Comparison schools	
	Mean	SD	Mean	SD
Student variables				
Age in year 0 (1997)	13.3	1.9	13.3	1.9
Sex (male = 1)	0.54	0.50	0.53	0.50
Dropout rate (before year 1)	0.13	0.34	0.13	0.34
Repetition rate (before year 1)	0.25	0.43	0.26	0.44
Took government exam (year 0)	0.79	0.41	0.79	0.41
Took NGO exam (year 0)	0.84	0.37	0.82	0.38
Student test scores (*year* 0)				
English (govt. exam)	0.26	1.09	0.17	1.02
Math (govt. exam)	0.11	0.98	0.11	1.00
Science/agric. (govt. exam)	0.10	0.96	0.09	1.01
Swahili (govt. exam)	0.31	0.94	0.21	0.94
Geog./hist./Christ. (govt. exam)	0.20	1.03	0.25	1.01
Art/crafts/music (govt. exam)	0.10	0.96	0.21	1.03
Home sci./business (govt. exam)	0.10	0.94	0.17	1.04
English (NGO exam)	0.15	0.99	0.07	1.00
Math (NGO exam)	0.14	0.99	0.07	1.01
Science (NGO exam)	0.13	0.97	0.06	1.00
Score on formula used to award prizes	0.05	1.03	0.00	1.00
Teacher/school variables (*year* 0)				
Teacher attendance rate (percent)	0.83	0.35	0.79	0.37
Teacher present in classroom (percent)	0.65	0.48	0.68	0.47
Teacher used blackboard (percent)	0.95	0.21	0.94	0.24
Teacher used teaching aid (percent)	0.09	0.29	0.12	0.32
Teacher energy (from 1 to 5)	4.24	0.84	4.20	0.86
Teacher assigned homework (percent)	0.31	0.46	0.27	0.44
Prep sessions offered (percent)	0.40	0.49	0.41	0.49

Notes: For student variables, age, sex, and "take NGO exam" are for year 1997, the government exam is for 1996 (that exam was not given in 1997), and dropout and repetition are defined by comparing 1996 and 1997. The sample sizes in 1996 (1997) were 7,200 (7,492) for the teacher incentive schools and 8,024 (8,226) for the comparison schools, although all variables had some missing values (this is particularly true for age, which is missing for about 30 percent of the sample, and for repetition, which is not defined for students who dropped out). See Tables 4 and 5 for sample sizes for the teacher/school variables. Note that teacher absence is an average for each teacher over two visits. Finally, *none* of the differences in the means between the teacher incentive and comparison schools in this table is statistically significant.

teachers assigned to program or comparison schools in year zero. Teachers who changed schools in year 1 or year 2 were assigned to their initial schools.[7]

Table 1 provides descriptive statistics in year zero (before the program) for the main variables of interest. None of the differences in the means between the teacher incentive and the comparison schools is statistically significant.

III. Program Impact of Student Outcomes

This section presents our estimates of the impact of the teacher incentives program on student outcomes. In general, we focus on outcomes in the second year

[7] This could be done only for those teachers who switched schools and remained in the sample of 100 schools. There are no data on teachers who switched to other schools, so they were dropped from the analysis. Movement of teachers out of the 100 schools was similar (no significant differences) for program and comparison schools.

(year 2), because reports from the field suggest that teachers took the program more seriously after they had seen it work in year 1, and thus better understood the incentives. This also seems closer to what one would expect from a scaled up program.

Section IIIA shows that teachers responded to the program. Scores increased on the formula used to reward teachers. That increase could have been caused by more students taking the exam and/or higher test scores for the students taking the exam. Section IIIB assesses the role played by higher exam participation. It shows that program school students were more likely to take the government exams, but this reflects more enrolled students taking those exams, not reduced dropping out. Section IIIC examines whether test scores increased on the government exam, and whether increases in those scores are mostly due to a broad-based increase in human capital or to an increase in more narrow skills that raise test scores and then dissipate quickly. In year 1, the results are inconclusive. In year 2, once teachers had had a chance to learn how the program worked, the program led to gains on the government exams (on which the incentives were based). However, it did not lead to similar gains in year 2 on the NGO tests (which were not tied to incentives and, unlike year 1, had a different format). There is also evidence that the program raised performance on multiple-choice questions relative to fill-in-the-blank questions.

A. *Scores on Formula Used to Reward Teachers*

To see whether teachers responded to the incentives program, Table 2 (panel A) presents estimates of the program's impact on the formula used to reward teachers. To maximize precision, all estimates are based on regressions that aggregate over all grades and all subjects, thus estimating the (weighted) average impact of the program. In addition to the program dummy variable, dummy variables are added for all grade/subject combinations, sex, and seven geographic divisions. (Regressions without these controls yield similar results.) The associated standard errors are robust to heteroscedasticity and allow for unstructured correlation of test scores for students in the same school. Following a standard intention-to-treat (ITT) approach, we use only students who were enrolled in year 1 (February 1998) and assign the few students who later switched schools to their initial schools.

Before the program began (year zero), there was no significant difference in the formula across the program and comparison schools (panel A, column 1). Yet, in the program's first year (column 2), the formula in the program schools was 0.13 standard deviations higher than in the comparison schools, a difference significant at the 10 percent level. By year 2 (column 3), the score on the formula in those schools was 0.22 standard deviations higher, which is significant at the 1 percent level. After the program ended (year 3), there was again no significant difference in the scores on the formula (column 4), suggesting little persistence of program impact.

These results demonstrate that school teachers responded to the program's incentives. They also suggest that over time teachers learned how to respond to the incentives.

Table 2—Program Impacts on Score on Teacher Reward Formula, Exam Participation, and Dropping Out

	Year 0 (Pre-program) (1)	Year 1 (2)	Year 2 (3)	Year 3 (post-program) (4)
Panel A. Dependent variable: score on formula used to reward teachers				
Incentive school	0.036	0.131*	0.215***	0.026
	(0.083)	(0.079)	(0.075)	(0.060)
Observations	63,812	76,509	73,789	57,674
Panel B. Dependent variable: take government exam (linear probability model)				
Incentive school	0.002	0.064*	0.070**	−0.005
	(0.029)	(0.038)	(0.029)	(0.028)
Observations	14,945	9,731	11,651	8,964
Panel C. Dependent variable: take NGO exam (linear probability model)				
Incentive school	0.010	0.019**	0.010	0.032
	(0.012)	(0.008)	(0.028)	(0.036)
Observations	14,921	13,085	12,982	2,277
Panel D. Dependent variable: dropping out (linear probability model)				
Incentive school	0.004	−0.008	−0.008	0.002
	(0.017)	(0.012)	(0.011)	(0.009)
Observations	13,841	13,347	12,007	9,479
Panel E. Dependent variable: take government exam if enrolled (linear probability model)				
Incentive school	0.002	0.061	0.076**	−0.004
	(0.029)	(0.038)	(0.034)	(0.032)
Observations	14,945	9,627	10,032	7,529

Notes: Robust standard errors, based on regressions that permit unstructured correlation within schools, in parentheses.The results in panel A have up to seven observations per student (one for each government test). Government test data are not available for Teso in year 1 (1998). In panel C, the year three NGO exam data are available for only 27 of the 100 schools. In all panels, the sample in column 1 is limited to primary school pupils who were in an upper grade (4–8) in year zero, and the sample in columns 2–4 is limited to students who were in an upper grade in year 1. In panel A, columns 3 and 4 are limited to pupils who did not drop out or transfer to another school in those years, but do include repeaters (estimates that exclude repeaters are very similar).

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

B. *Exam Participation*

Most of the increase in scores on the formula used to reward teachers was due to greater exam participation. Incentives for teachers were based on their students' scores on the government exams administered every October in grades 4–8.[8] Some students do not participate since doing so costs KES 120 (US $2). As explained

[8] These exams, also called district or mock exams, are prepared by the District Education Office of the Ministry of Education. They are designed to prepare students for the KCPE, and so have a very similar format. Results for the KCPE, which is only for grade 8, are similar to the government exam results (see Glewwe, Ilias, and Kremer 2003).

above, to avoid giving teachers an incentive to encourage weak students to drop out, students who did not take the test were assigned very low scores on the formula used to award prizes. This gave program school teachers an incentive to encourage enrolled students to take the exam. Even by guessing, students could get higher scores than the scores assigned to them if they did not take the exams.

Exam participation rates were higher in program schools than in comparison schools for the government exams (on which the incentives were based), but less so for the (nonincentive) NGO exams. Baseline participation (year zero) was about 80 percent on both types of exams (Table 1), with no statistically significant difference across incentive and comparison schools (Table 2, panels B and C, column 1). In year 1, participation in the government exams was 6.4 percentage points higher in the incentive schools (panel B, column 2). By year 2, it was 7.0 percentage points higher and significant at the 5 percent level. In the post-program year, there was no incentive to encourage students to take the test and the participation rate was 0.5 percentage points lower in the incentive schools, an insignificant difference. In contrast, participation in the NGO exams, which had no link to teacher incentives, was only slightly higher in the incentive schools in years one and two, and the only significant difference was 1.9 percentage points (panel C, column 2).[9]

This higher exam participation can be further decomposed into reduced dropping out and increased exam participation conditional on enrollment. Increasing exam participation conditional on enrollment does not raise broad human capital, but reductions in the dropout rate could lead to sizeable increases in human capital acquisition. Panel D of Table 2 shows that the program did not affect dropping out in either year 1 or year 2, and so it did not affect enrollment. Thus, all of the increase in exam participation is due to increased participation among students who are already enrolled. This is confirmed in panel E of Table 2.

C. *Test Scores*

In addition to raising students' exam participation, teachers could increase their probability of receiving a prize by raising student performance on the government exams. Table 3 shows the impact of the teacher incentives program on students' scores on the government and NGO exams.[10] There is no significant difference in pre-program scores on either set of exams between incentive and comparison

[9] The sample size drops over time, primarily because more students complete their primary education and thus leave our sample. The samples in years one, two, and three are limited to students who were enrolled in grades 4–8 in February 1998 (year 1) and had not yet graduated.

[10] The sample sizes in Table 3 vary for several reasons. First, the sample size on the government exams for year zero is much smaller than in year 1 because no district exams were offered in 1997, so the exams are from 1996 and in that year they were given only to students in grades 5–8. Second, the sample size increases somewhat from year 1 to year 2, even though some children dropped out or graduated between those two years because Teso district did not offer government tests in year 1. Third, the drop from year 2 to year three reflects missing test scores for students who were in grades 4–8 in year 1, but graduated, dropped out, or switched schools between years two and three. The same applies for the drop in observations for the NGO tests between years one and two. In addition, for the NGO tests there were three subject tests in years zero and one, but only two in year 2. Finally, data for the NGO exams in year three include only 27 of the 100 schools—those that participated in a de-worming project. Point estimates (not shown in Table 3) are positive, but none of the t-statistics exceeds one.

Table 3—Program Impact on Test Scores and Grade Repetition

	Year 0	Year 1	Year 2	Year 3	Year 1– Year 0	Year 2– Year 0	Year 3– Year 0
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Dependent variable: score on government exam							
Incentive school	0.019	0.052	0.144	0.090	0.048	0.136*	0.077
	(0.111)	(0.096)	(0.087)	(0.086)	(0.061)	(0.071)	(0.071)
Year 0 scores	—	—	—	—	0.427***	0.308***	0.267***
					(0.056)	(0.047)	(0.050)
Observations	25,303	50,694	54,401	33,502	50,635	54,346	33,457
Panel B. Dependent variable: score on NGO exam							
Incentive school	0.048	0.092	0.024	—	0.046	−0.017	—
	(0.093)	(0.086)	(0.101)		(0.041)	(0.064)	
Year 0 scores	—	—	—	—	0.691***	0.615***	—
					(0.042)	(0.047)	
Observations	33,487	39,900	18,736	—	39,900	18,736	—
Panel C. Dependent variable: grade repetition (linear probability model)							
Incentive school	0.006	−0.030	−0.029	—	—	—	—
	(0.021)	(0.023)	(0.026)				
Observations	11,649	10,542	8,907	—	—	—	—

Notes: Robust standard errors, which allow for clustering at the school level, are in parentheses. Repeaters are included. Regressions without repeaters yield similar results. The scores of the repeaters were adjusted by subtracting one from their normalized test scores since, by repeating the same year, they were in effect taking easier exams and their scores were about one standard deviation above the average test scores in comparison schools. Year 1government test results are available only for Busia. Year 2 and year three government test results are available for both Busia and Teso. This leads to the increase in sample size from year 1 to year 2 on the government exam.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

schools (Table 3, panels A and B, column 1).[11] Columns 2–4 of Table 3 show estimated results that do not control for initial test scores in year zero. Conditioning on those test scores at the student level greatly reduces the sample size (see Glewwe, Ilias, and Kremer 2003), so columns 5–7 condition on average test scores in year zero, averaging by school, grade, and subject. This increases the precision of the estimated program effects with almost no loss of sample size.

Both estimates of the impact of the incentives program on government exam scores are small (about 0.05 standard deviations of the test score) and statistically insignificant in year 1 (panel A, columns 2 and 5). In year 2 (columns 3 and 6) both estimates are larger, about 0.14, and the one that conditions on year zero test scores is significant at the 10 percent level.

For the NGO exams, the estimated program effects are generally smaller and never significant. In year 1, when the format of the NGO exams was similar to that

[11] Since grade 2 students had no government exams in 1996, we used 1997 NGO tests, when available, as pretests for government tests of grade 4 students in 1998.

of the government exams, the point estimates of the program impact were similar to those on the government exam, positive, but small and insignificant. In year 2, after teachers had time to learn how the program functioned, and the NGO test format was primarily fill-in-the-blank and thus different from the government test format, the estimated impact on NGO tests is close to zero. Another piece of evidence that the program did not increase broad measures of human capital is that it did not lead to a large reduction in grade repetition (Table 3, panel C). Although the point estimate is negative, it is neither statistically significant nor particularly large.

There remains one more issue to discuss concerning the government exam. Recall from Section IIIB that participation in the government exam increased in incentive schools relative to comparison schools. Theoretically, this could bias scores on that exam in either direction. If teachers try to persuade all students who do not want to take that exam to take it, the addition of marginal students would probably reduce average test scores, since academically weak students are less likely to pay for the government exam. But if teachers focus on convincing potentially high-scoring students and their parents of the test's importance, average scores in the incentive schools may increase. To check for potential bias, we compared year 1 scores on the NGO test for the students in incentive and comparison schools who were eligible to take the government exam in year 2 (over 90 percent of whom took the NGO test in year 1). For the 9,965 students eligible to participate in the year 2 exams who had NGO scores from year 1, we regressed those scores on a dummy variable for incentive schools, a dummy variable for not taking the government exam in year 2, and an interaction term. The interaction indicates whether the 30 percent of students in the comparison schools who did not take the government exam have the same mean NGO scores as the 20 percent of students in the program schools who did not take that test. There should be a large difference if the 10 percent of students in incentive schools induced by the program to take the government test were unusually strong or weak students. The estimated impact of this interaction term was just 0.065 standard deviations, and statistically insignificant (t-statistic of 0.80). We conclude that higher exam participation does not lead to a sizable bias in our estimates of the program's impact on government exam scores.[12]

Summarizing the test score results, it is unclear how much impact the program had in year 1, but by year 2, once teachers had an opportunity to see how the program worked, there is evidence that the program raised scores on the government exam. The estimated impacts on the NGO test were close to zero in year 2, although, due to imprecise estimates, the differences in the estimated impacts on the government and NGO exams are also insignificant.

The results thus far can be used to decompose the impact of the program on the formula used to reward teachers into the impact from higher participation on the government exam and the impact of higher scores on that exam for those who participated. This is done as follows. The government tests had a maximum score

[12] A more rigorous approach to address this bias is to estimate bounds on the estimated impact of the program on test scores that account for its impact on taking the government exam. Attempts to do so using David S. Lee's (2009) method led to statistically insignificant effects, which is not surprising given the weak significance in panel A of Table 3.

of 100. Recall that, for all but the English test, the formula stipulates that a score of 15 be given to any student not taking the test. Let p be the proportion of students in a school who take the government tests. For the six subjects other than English, the formula used to award prizes is $(1 - p) \times 15 + p \times \overline{T}$, where $\overline{T}$ denotes school average test scores. The English tests had an essay format that precluded guessing. Students not taking that exam were given a score of 0, so the formula for those tests is $(1 - p) \times 0 + p \times \overline{T} = p \times \overline{T}$. The formula results in Table 2 and test score results in Table 3 are based on normalized test scores. In year 2, the average mean score across subjects was 38, and the average standard deviation was 12.2. Averaging over the normalized tests gives the formula used to reward teachers:

$$\begin{aligned}
&\frac{6}{7}(1-p)\,\frac{15-\mu}{\sigma} + \frac{1}{7}(1-p)\,\frac{0-\mu}{\sigma} \\
&\qquad + p\left[\frac{(\overline{T}_1-\mu)+(\overline{T}_2-\mu)+\cdots+(\overline{T}_7-\mu)}{7\sigma}\right] \\
&\qquad = \frac{-\mu}{\sigma}(1-p) + (1-p)\left[\frac{6\times 15}{7\times\sigma} + \frac{1\times 0}{7\times\sigma}\right] \\
&\qquad\quad + p\,\frac{1}{\sigma}\left[\frac{\overline{T}_1+\overline{T}_2+\cdots+\overline{T}_7}{7} - \mu\right] \\
&\qquad = \frac{-\mu}{\sigma} + \frac{1}{\sigma}\;(1-p)\left[\frac{6\times 15}{7}\right] \\
&\qquad\quad + p\,\frac{1}{\sigma}\left[\frac{\overline{T}_1+\overline{T}_2+\cdots+\overline{T}_7}{7}\right] \\
&\qquad = \frac{1}{\sigma}\left[-\mu + (1-p)\times\left(\frac{90}{7}\right) + p\times\overline{T}\right],
\end{aligned}$$

where μ is the mean score and σ is the standard deviation. The total derivative of this expression with respect to p and $\overline{T}$ is:

$$d\text{Formula} = (1/\sigma)[\overline{T} - 90/7]\;dp + [(1/\sigma)\times p]\;d\overline{T}.$$

In year 2, the impact of the program on the formula was 0.215, and p was 0.59. From Table 3 (column 6), $d\overline{T} = 1.66$ (0.136×12.2); and from Table 2 (column 3), $dp = 0.070$. Since $\sigma = 12.2$ and $\overline{T} = 38$, $(1/\sigma)[\overline{T} - 90/7] = 2.06$ and $[(1/\sigma) \times p] = 0.048$. Inserting these into the formula above implies that the higher exam participation rate constitutes 0.144 of dFormula in year 2, the change in the scores on the government test constitutes 0.080, and the sum of these (0.224) is very close to the actual dFormula, 0.215. Thus, about two-thirds (64 percent) of the impact of the test scores on the formula is through increases in the number of students taking the

Table 4—Impact of Program on Test Taking Skills and Prep Classes

	Any of last four multiple-choice questions correct? (probit)	Probability of correctly answering a multiple-choice question (linear probability model)	Probability of correctly answering a fill-in-the-blank question (linear probability model)	Relative likelihood of correctly answering multiple-choice and fill-in-the-blank questions (linear probability model)	
				Difference	Ratio
Panel A. Dependent variables: indicators of correctly answering different types of questions (year 2 English test)					
Incentive school	0.243**	0.036*	0.022	0.015**	0.048**
	(0.105)	(0.020)	(0.019)	(0.007)	(0.024)
Observations	8,573	9,395	9,395	9,395	9,371
Panel B. Dependent variable: number of test preparation sessions					
	Year 0	Year 1	Year 2	Year 1–Year 0	Year 2–Year 0
Incentive school	−0.011	0.042	0.074**	0.053	0.085*
	(0.044)	(0.037)	(0.034)	(0.042)	(0.046)
Observations	3,000	3,000	3,000	3,000	3,000

Notes: Robust standard errors, allowing for within-school correlation, are in parentheses. For panel A, all regressions included sex, grade, and division dummy variables as control variables. They also omit grade 4 because grade 4 did not have any multiple-choice questions after the first easy 20 multiple-choice questions. The student is the unit of observation. For panel B, each observation is a prep session in a grade in a school. All regressions include a constant, a grade variable, and dummy variables for geographic regions.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

test, not increases in test scores, and the other one-third is due to the increase in test scores.

There is also evidence that the program improved students' test taking techniques. Prep session teachers instruct students not to leave blanks on multiple-choice questions. We have data on whether students correctly answered each question on the 1999 NGO English test, which had mostly a fill-in-the-blank format (all math test questions had that format). That test began with 20 relatively easy multiple-choice questions, followed by 74 more difficult questions. Of the 74 questions, 4 were multiple choice and 70 were fill-in-the-blank. The data do not indicate whether a student answered a question, but of course unanswered questions received a score of 0.

To see whether the program improved students' test taking skills, we constructed two variables, one indicating the percentage of the four relatively hard multiple-choice questions that were answered correctly, and the other indicating the same percentage for the 70 fill-in-the-blank questions. A random effects probit regression of whether any of the multiple choice questions were answered correctly on a program dummy variable, and sex, grade, and division dummy variables, yields a positive program impact that is significant at the 5 percent level (Table 4, panel A, column 1), suggesting that incentive school students are less likely to leave blanks. In addition, the program had a stronger impact on the probability of correctly answering multiple-choice questions than on correctly answering fill-in-the blank questions (Table 4, panel A, columns 2 and 3), and only the effect on the multiple-choice questions is significant. Indeed, the relative likelihood of correct answers on multiple-choice questions versus fill-in-the-blank questions (columns 4 and 5) is also higher, and statistically significant, in program schools (5 percent level for the ratio of the two constructed variables, and 10 percent level for the difference of those variables).

Next, there is some evidence that the largest program effects were in subjects where memorization is important. Average effects in years one and two were largest for the geography, history, and Christian religion (GHCR) test, the next largest were for science and math, and no other subjects had significant effects (Glewwe, Ilias, and Kremer 2003). GHCR is the subject where memorization is arguably most important, and so it is particularly susceptible to "cramming" before exams. Primary school science also involves substantial memorization, but math likely requires less.[13]

Finally, consider evidence from year three, after the program ended. As mentioned above, the impact of the program on the formula used to reward teachers fell to zero. Test score gains on the government exam were smaller and not statistically significant in year three, after the program ended (Table 3, panel A, column 4), but standard errors are large for year three since much of the sample has attritted, so we cannot reject the hypotheses of full retention of gains, no retention of gains, or normal depreciation in learning over time (see, e.g., Tahir Andrabi et al. 2008). It is worth noting, however, that increases in learning from other programs in the same area of Kenya persisted after those programs ended (Esther Duflo, Pascaline Dupas, and Kremer 2009; Kremer, Edward Miguel, and Rebecca Thornton 2009).

In summary, the program clearly increased scores on the formula to reward teachers, but about two-thirds of that impact is due to the increase in the number of students taking the government exams. The other one-third reflects increased scores on the government tests, at least some of which reflects an increase in narrow test taking skills (strategies for handling multiple-choice questions, or "cramming" for tests that are prone to memorization). Overall, while the program clearly led to gains on narrow outcomes that were the focus of the incentives, we find little evidence of a broader-based increase in students' human capital.

IV. Program Impact on Teacher Behavior

This section examines the impact of the program on teacher behavior. Wherever possible, we focus on objective data on teacher behavior rather than self reports, since self reports may be inaccurate. Following Chaudhury et al. (2006), we also put more weight on teacher presence in school than on teacher presence in class conditional on being in school, since teachers may have been spurred to go from the staff room to the classroom when they saw that they were being observed, and this might be particularly likely in the incentive schools.

[13] ICS staff members familiar with the curriculum suggested that GHCR and home science and business education require the most memorization, science requires a medium amount, and English, math, and Swahili require the least.

As with student outcomes, direct observations of teacher behavior are consistent with the hypothesis that teachers' efforts were fairly narrowly targeted toward increasing scores on the formula used to reward teachers, rather than toward increasing human capital more broadly. First, there is evidence that the program led teachers to offer more prep classes. Before the program, headmasters in incentive schools were slightly less likely to report offering preps (Table 4, panel B, column 1), but after the program started, they reported that teachers conducted more preps (columns 2 and 3). They were 4.2 percentage points more likely to report conducting preps in year 1 and 7.4 percentage points more likely in year 2, and the latter is significant at the 5 percent level. Differenced estimates (columns 4 and 5) are similar. Ideally, we would like direct observations of prep classes, since headmaster reports may be unreliable and because we found evidence of social desirability bias in teacher responses to a special questionnaire asking their views on the program and their responses to it, as discussed above. However, the questionnaire used to collect data from headmasters about prep classes does not even mention the teacher incentive program, so it seems less likely to be subject to social desirability bias. Also, it is not clear that social desirability bias would increase reports of test preparation sessions rather than reduce them.

Second, external observations of teacher behavior that might be thought to promote broad human capital acquisition provide no evidence of changes in teacher attendance, homework assignment, or pedagogy (Table 5). Before the program, the schools later selected to be incentive schools had slightly higher teacher attendance, but this difference is insignificant (Table 5, panel A, column 1).[14] In year 1 of the program, teacher attendance was slightly lower in the incentive schools, while in year 2, it was slightly higher (panel A, columns 2 and 3). Yet both differences are completely insignificant,[15] as are difference-in-difference estimates (columns 4 and 5).

There is no evidence that the program affected the presence or behavior of teachers in the classroom. Trained observers watched each teacher each year for one class period, recording several measures of teacher behavior, both objective information on teacher activities and subjective impressions of their energy level and caring for students. Prior to the program, there were no significant differences between the incentive and comparison schools in teacher presence in the classroom or in any pedagogy measure (Table 5, panels B–F, column 1). During the program, differences between incentive and comparison schools in teachers' presence in class (panel B, columns 2 and 3) were not statistically significant. While one could argue that the (insignificant) year 2 point estimate, a 4.4 percentage point increase, is large, teachers can manipulate this more easily on days they are observed than they can manipulate presence at the school, for which estimated program impacts are very small. We also find no significant differences during the program period (years

[14] These results are robust to a specification where each visit is treated as a binary opportunity for attendance and the month of visit is controlled for. The samples in panel A are smaller than those in panel B because teacher attendance data exist only for teachers in upper grades.

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

Table 5—Program Impacts on Teacher Attendance and Pedagogy

	Year 0 (1)	Year 1 (2)	Year 2 (3)	Year 1–Year 0 (4)	Year 2–Year 0 (5)
Panel A. Dependent variable: teacher attendance (percent of visits teacher is present)					
Incentive school	0.041 (0.038)	−0.014 (0.021)	0.002 (0.027)	−0.066 (0.047)	−0.036 (0.056)
Observations	454	407	349	330	289
Panel B. Dependent variable: teacher present in the classroom (linear probability model)					
Incentive school	−0.020 (0.050)	−0.008 (0.042)	0.044 (0.058)	0.009 (0.065)	0.081 (0.082)
Observations	631	826	481	380	373
Panel C. Dependent variable: use of blackboard (linear probability model)					
Incentive school	0.018 (0.024)	−0.028 (0.022)	0.047 (0.040)	−0.048 (0.031)	0.069 (0.063)
Observations	404	598	237	246	142
Panel D. Dependent variable: use teaching aid (linear probability model)					
Incentive school	−0.028 (0.028)	−0.004 (0.030)	0.015 (0.040)	0.027 (0.57)	0.050 (0.070)
Observations	399	567	235	241	140
Panel E. Dependent variable: teacher energy (1 to 5: 1 = energetic)					
Incentive school	0.010 (0.094)	0.031 (0.069)	−0.129 (0.106)	−0.026 (0.143)	−0.048 (0.181)
Observations	383	570	233	239	139
Panel F. Dependent variable: homework assignment (linear probability model)					
Incentive school	0.054 (0.036)	−0.045 (0.045)	−0.008 (0.045)	−0.079 (0.057)	−0.020 (0.058)
Observations	1,666	1,676	2,371	401	385
Panel G. Aggregating over all variables					
Incentive school	0.007 (0.060)	−0.057 (0.059)	0.003 (0.061)	−0.091 (0.086)	0.045 (0.103)
Observations	500	500	500	500	500

Notes: Robust standard errors, allowing within-school correlation, in parentheses. In panels A–E, each observation in columns 1–3 is either a teacher or a classroom visit. In panel F, each observation in columns 1–3 represents a student asked about homework assigned the previous day. All estimates in columns 4–5 of panels A–F use grade in a school as the unit of observation since classrooms, teachers, and pupils cannot be matched over years. In panel G, each observation represents a grade in a school. All regressions include a constant, a grade variable, and dummy variables for geographic regions (teacher attendance also inludes a teacher sex variable).

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

one and two) between the two school groups in any pedagogical practices (panels C–E, columns 2 and 3).[16]

[16] We examined many pedagogy measures, but present results for only two objective measures (teaching aids and use of blackboards) and one subjective one (teacher energy). Point estimates are near zero for all measures. In

There is also no evidence that the program raised homework assignment. In grades 4–8, information was collected for each school from a random subset of students on whether they were assigned homework the previous day. Incentive schools assigned slightly more homework than comparison schools in year zero, but the difference is insignificant (panel F, column 1). During the program, incentive schools assigned slightly less homework, yet the gap was never statistically significant in either levels or differences (columns 2, 3, 4, and 5). Last, we used a method similar to that of Jeffrey R. Kling, Jeffrey B. Liebman, and Lawrence F. Katz (2007) to jointly test whether the program had any impact on the teacher attendance and pedagogy variables (all variables in panels A–F). We could not reject the joint hypothesis that all six estimated effects were zero (panel G).

A final issue is that teacher incentives can stimulate teacher cheating, as discussed by Brian A. Jacob and Stephen D. Levitt (2003). In Kenya, as in most countries, outside monitors, often teachers from other schools, supervise government exams. In one program school teachers colluded with those monitors in year 1 to facilitate student cheating. That school was disqualified in year 1, but allowed to participate in year 2. Its scores are excluded in the year 1 analysis, but included in year 2. No cheating was found in comparison schools. Yet we doubt that teacher cheating had a major impact on program schools' test scores. Analysis of item responses to detect cheating using the techniques of Jacob and Levitt (2003) provides little evidence of suspicious strings of questions for which all students in a class got the right answer. Also, the similar program impact on the regular government exams and on the heavily monitored KCPE (see Glewwe, Ilias, and Kremer 2003) suggests little or no cheating.

V. Summary, Discussion, and Policy Implications

Schools randomly selected to participate in this teacher incentive program in Kenya scored significantly higher on the formula used to determine teacher awards. The estimated impact on this formula grew between the first and second years of the program, suggesting that teachers responded more effectively after they had time to see how the program worked. Indeed, anecdotal evidence from the first year's award ceremony indicates that prior to that ceremony, some teachers did not fully realize that students who drop out or do not take the test reduced their chance of winning a prize. Students in program schools were also more likely to take the exams linked to incentives, and by the program's second year they had higher scores on those exams.

Yet there is little evidence that teachers in the program schools increased efforts to reduce dropouts or promoted broad acquisition of human capital. Scores on exams not linked to incentives did not increase significantly. Teachers in program schools were neither more likely to be in school nor more likely to assign homework. Pedagogical practices and student dropout rates were similar in program and comparison schools. Instead, there is evidence of increased test preparation sessions and increased test-taking among students enrolled in program schools, and a

the difference-in-difference estimates (columns 4 and 5) the unit of observation is grades within schools as teachers cannot be matched across years. These estimates are also close to zero and far from statistically significant.

decomposition analysis suggests that two-thirds of the increase in the formula used to reward teachers in those schools is due to increased test-taking among currently enrolled students. Lastly, students in program schools also did better on multiple-choice questions relative to fill-in-the-blank questions.

Our interpretation that the program had little or no effect on broad measures of human capital does have three caveats. First, some of our estimates are imprecise. We cannot reject the possibility that there were moderate gains on some measures of human capital, and that some of those gains persisted after the program ended. Second, teacher incentives may work as much by encouraging potentially good teachers to enter the profession as by eliciting higher effort from those who would be teachers in any case. Yet given the queuing for teaching posts in Kenya, it is unlikely that people with either teaching jobs or the academic qualifications needed to enter teacher training colleges (but not universities) are currently opting out of that profession. Thus, any effect on this margin in Kenya, or in other developing countries with queues for teaching jobs, is likely to be small.

Third, an alternative incentive design could have had a more favorable impact. For example, the program was explicitly temporary. A permanent program might have led teachers to invest in boosting long-run learning. On the other hand, Kenyan teachers often transfer between schools. Moreover, being temporary allowed the program to base incentives on improvements over baseline performance, to include incentives to prevent students from dropping out, and to restrict the program to teachers already in school and avoid giving teachers incentives to transfer to schools with better students. To take another example, while the incentives were similar in magnitude to those in most US programs and in the Israeli program analyzed by Lavy (2002), perhaps larger incentives or teacher-specific incentives would have led to more efforts focused on broad acquisition of human capital. Of course, larger incentives could also induce wasteful or even harmful signaling effort, such as cheating on tests or forcing weak students to drop out, and individual-level teacher incentives might undermine cooperation within schools. Yet we cannot rule out that incentives based on different tests, or that treated students taking the exam at the same rate as comparison students, could have generated more favorable outcomes.

Several papers have used nonexperimental data to examine teacher incentive programs in developed countries. Some argue that the programs promoted broad-based learning, but many others find gains that are concentrated on narrow measures of incentivize-specific skills, and still others find potentially counterproductive effects on teacher behavior. Lavy (2002) finds that rewarding Israeli teachers based on average school (rather than individual teacher) performance raised test scores, but not performance on matriculation exams. Analyzing another Israeli program, based on individual teacher performance and with much larger prizes, Lavy (2004) concludes that pass rates of weak students on the high school matriculation exam rose by 7–18 percentage points. Jacob (2005) finds that a Chicago program that did not promote low scoring students, and put their schools and teachers on probation, raised students' scores, though the gains were largest for skills used on the high-stakes exam, and some schools raised scores by putting more pupils into special education. Daniel M. Koretz (2002) estimates that a Kentucky teacher incentive program had large positive impacts (0.5 to 0.6 standard deviations) on the test used to decide teacher rewards, but much smaller effects on another test not tied to the rewards. Some researchers find that high-stakes testing can lead teachers or administrators to manipulate test results. David N. Figlio and Joshua Winicki (2005) show that Virginia school districts increase calories in school lunches on days when students take high-stakes tests, artificially inflating test scores. Jacob and Levitt (2003) estimate that 4–5 percent of Chicago elementary school teachers help their pupils cheat, and that cheating increased after high-stakes testing was introduced.

Turning to evidence from developing countries, Karthik Muralidharan and Venkatesh Sundararaman (2009) report on a randomized evaluation of a teacher incentive program in India. As in this study, they find that paying teachers for test score increases in India raises test scores and exam preparation sessions, but does not improve teacher attendance. However, while we find no spillovers to exams that had a different format and were not linked to the incentives, Muralidharan and Sundararaman (2009) do find spillovers to questions with an unfamiliar format that were designed to measure conceptual understanding. They also find gains on subjects not included in the incentives. This may represent an increase in underlying learning, but as some of the questions were multiple choice, it is possible that test-taking skills acquired during exam preparation sessions were useful for these unfamiliar formats and for other subjects. Whether the program in India leads teachers to change behavior in ways that promote broad acquisition of human capital, or merely to target narrower skills that only raise test scores, will become clearer after that program ends. At this point, it is difficult to determine whether the differences in our study and that study reflect differences between the Kenyan and Indian contexts, different measurement approaches, or differences in these two teacher incentive programs. For example, the impact of incentives may depend on teacher training, which could differ between the Kenyan and Indian contexts. Another example is that teachers may respond to limited incentives, based on schools' average performance, with low-cost actions that boost scores on the formula but do little to increase broader learning, while strong incentives, based on individual teacher performance, may induce them to take actions that increase long-run learning.

While we do not see measurable improvements in broader indicators of achievement or human capital acquisition, we also do not see negative effects on these variables, as may be the case if teacher effort on narrow skills and broader skills were substitutes. If there were some negative effects, one could argue that they may be smaller than the program's distributional benefits. Students from more privileged backgrounds already have opportunities to take part in test prep activities and mock exams to prepare them for the KCPE exam, and this program may have provided those opportunities to less privileged students.

Finally, our finding that, for the program and context we examine, teacher incentives based on student test scores were insufficient to solve the problem of high teacher absence rates suggests that it is worth exploring other types of reforms to address the problem of weak teacher incentives. Several recent studies have done so. Incentives for teacher attendance proved effective in an Indian NGO setting (Duflo, Rema Hanna, and Stephen Ryan 2007), where they were implemented by monitors outside the school. There seems to be little downside to such incentives, so it would seem desirable to strengthen them. Duflo, Dupas, and Kremer (2009) and

Muralidharan and Sundararaman (2008) find positive effects from devolving control over teachers to local school committees. Joshua Angrist et al. (2002) and Angrist, Eric Bettinger, and Kremer (2006) and Kremer, Miguel, and Thornton (2009) find positive effects from programs that reward students, as opposed to teachers, for their academic performance and/or allow parents to choose schools and tie school finance to their decisions. Much more remains to be learned, but these findings show that it is possible to find solutions to the problem of weak teacher incentives in developing countries.

REFERENCES

American Federation of Teachers. 1997. "Survey Analysis of Salary Trends 1997." http://archive.aft.org/salary/1997/download/salarysurvey97.pdf.

Andrabi, Tahir, Jishnu Das, Asim I. Khwaja, and Tristan Zajonc. 2008. "Do Value-Added Estimates Add Value? Accounting for Learning Dynamics." Unpublished.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *American Economic Review*, 92(5): 1535–58.

Angrist, Joshua, Eric Bettinger, and Michael Kremer. 2006. "Long-Term Educational Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia." *American Economic Review*, 96(3): 847–62.

Baker, George. 2002. "Distortion and Risk in Optimal Incentive Contracts." *Journal of Human Resources*, 37(4): 728–51.

Bangert-Drowns, Robert L., James A. Kulik, and Chen-Lin C. Kulik. 1983. "Effects of Coaching Programs on Achievement Test Performance." *Review of Educational Research*, 53(4): 571–85.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers. 2006. "Missing in Action: Teacher and Health Worker Absence in Developing Countries." *Journal of Economic Perspectives*, 20(1): 91–116.

Clotfelter, Charles T., and Helen F. Ladd. 1996. "Recognizing and Rewarding Success in Public Schools." In *Holding Schools Accountable: Performance-Based Reform in Education*, ed. Helen F. Ladd, 23–64. Washington, DC: Brookings Institution Press.

Duflo, Esther, Pascaline Dupas, and Michael Kremer. 2009. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." http://www.povertyactionlab.org/sites/default/files/publications/61_Duflo_Peer_Effects_and_Tracking.pdf.

Duflo, Esther, Rema Hanna, and Stephen Ryan. 2007. "Monitoring Works: Getting Teachers to Come to School." Bureau for Research and Economic Analysis of Development (BREAD) Working Paper 103.

Figlio, David N., and Joshua Winicki. 2005. "Food for Thought: The Effects of School Accountability Plans on School Nutrition." *Journal of Public Economics*, 89(2–3): 381–94.

Glewwe, Paul, Nauman Ilias, and Michael Kremer. 2003. "Teacher Incentives." National Bureau of Economic Research Working Paper 9671.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. 2009. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics*, 1(1): 112–35.

Hanushek, Eric A. 1996. "Outcomes, Costs, and Incentives in Schools." In *Improving America's Schools: The Role of Incentives*, ed. Eric A. Hanushek and Dale W. Jorgenson, 29–52. Washington, DC: National Academy Press.

Holmstrom, Bengt, and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design." *Journal of Law, Economics, and Organization*, 7(S): S24–52.

Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics*, 89(5–6): 761–96.

Jacob, Brian A., and Steven D. Levitt. 2003. "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating." *Quarterly Journal of Economics*, 118(3): 843–77.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica*, 75(1): 83–119.

Koretz, Daniel M. 2002. "Limitations in the Use of Achievement Tests as Measures of Educators' Productivity." *Journal of Human Resources*, 37(4): 752–77.

Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2009. “Incentives to Learn.” *Review of Economics and Statistics*, 91(3): 437–56.

Lavy, Victor. 2002. “Evaluating the Effect of Teachers’ Group Performance Incentives on Pupil Achievement.” *Journal of Political Economy*, 110(6): 1286–1317.

Lavy, Victor. 2004. “Performance Pay and Teachers’ Effort, Productivity, and Grading Ethics.” National Bureau of Economic Research Working Paper 10622.

Lee, David S. 2009. “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects.” *Review of Economic Studies*, 76(3): 1071–1102.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2008. “Contract Teachers: Experimental Evidence from India.” http://econ.ucsd.edu/~kamurali/contract%20teachers.pdf.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2009. “Teaching Incentives in Developing Countries: Experimental Evidence from India.” Unpublished.

Olsen, Lynn. 1999. “Pay-Performance Link in Salaries Gains Momentum.” *Education Week*, October 13.

Richards, Craig E., and Tian Ming Sheu. 1992. “The South Carolina School Incentive Reward Program: A Policy Analysis.” *Economics of Education Review*, 11(1): 71–86.

Wyman, Wendy, and Michael Allen. 2001. “Pay-for-Performance: Key Questions and Lessons from Five Current Models.” Education Commission of the States (ECS) Issue Paper 2830. http://www.ecs.org/clearinghouse/28/30/2830.htm.

This article has been cited by:

1. Mark Gius. 2014. Using a Difference-in-Differences Approach to Estimate the Effects of Teacher Merit Pay on Student Performance. *Eastern Economic Journal* **39**:1, 111-120. [CrossRef]
2. A. Leigh. 2013. The Economics and Politics of Teacher Merit Pay. *CESifo Economic Studies* **59**:1, 1-33. [CrossRef]
3. Alejandra Mizala, Ben Ross Schneider. 2013. Negotiating Education Reform: Teacher Evaluations and Incentives in Chile (1990-2010). *Governance* n/a-n/a. [CrossRef]
4. Gadi Barlevy,, Derek Neal. 2012. Pay for Percentile. *American Economic Review* **102**:5, 1805-1831. [Abstract] [View PDF article] [PDF with links]
5. Benjamin A. Olken, Rohini Pande. 2012. Corruption in Developing Countries. *Annual Review of Economics* **4**:1, 479-509. [CrossRef]
6. Esther Duflo,, Rema Hanna,, Stephen P. Ryan. 2012. Incentives Work: Getting Teachers to Come to School. *American Economic Review* **102**:4, 1241-1278. [Abstract] [View PDF article] [PDF with links]
7. Sonja Fagernäs, Panu Pelkonen. 2012. Preferences and skills of Indian public sector teachers. *IZA Journal of Labor & Development* **1**:1, 3. [CrossRef]
8. Uri Gneezy,, Stephan Meier,, Pedro Rey-Biel. 2011. When and Why Incentives (Don't) Work to Modify Behavior. *Journal of Economic Perspectives* **25**:4, 191-210. [Abstract] [View PDF article] [PDF with links]
9. Tahir Andrabi,, Jishnu Das,, Asim Ijaz Khwaja,, Tristan Zajonc. 2011. Do Value-Added Estimates Add Value? Accounting for Learning Dynamics. *American Economic Journal: Applied Economics* **3**:3, 29-54. [Abstract] [View PDF article] [PDF with links]
10. Karthik Muralidharan, Venkatesh Sundararaman. 2011. Teacher Performance Pay: Experimental Evidence from India. *The Journal of Political Economy* **119**:1, 39-77. [CrossRef]
11. Derek NealThe Design of Performance Pay in Education **4**, 495-550. [CrossRef]