Incentives to Learn

Michael Kremer[*]

Edward Miguel[**]

Rebecca Thornton[***]

October 2005

Abstract: We report results from a randomized evaluation of a merit scholarship program for adolescent girls in Kenya, and discuss their implications for understanding educational production and for the policy debate surrounding merit awards. Girls who scored well on academic exams had their school fees paid and received a large cash grant. Girls eligible for the scholarship showed substantial gains in exam scores and gains persisted in the years following the competition. Both student and teacher school attendance increased in the program schools. Our results suggest not only that study effort is responsive to incentives but also that there are positive externalities: boys, who were ineligible for the award, also experienced exam gains, as did girls with low pretest scores (who were very unlikely to win). These large externalities address some of the equity concerns raised by critics of merit awards, and provide further rationale for public education subsidies.

[*] Dept. of Economics, Harvard University, The Brookings Institution, and NBER. Littauer 207, Harvard University, Cambridge, MA 02138, USA; mkremer@fas.harvard.edu.
[**] Dept. of Economics, University of California, Berkeley and NBER. 549 Evans Hall #3880, University of California, Berkeley, CA 94720-3880, USA; emiguel@econ.berkeley.edu.
[***] Dept. of Economics, Harvard University, Littauer 207, Cambridge, MA 02138, USA; rlthornt@fas.harvard.edu.

The authors thank ICS Africa and the Kenya Ministry of Education for their cooperation in all stages of the project, and would especially like to acknowledge the contributions of Elizabeth Beasley, Pascaline Dupas, James Habyarimana, Sylvie Moulin, Robert Namunyu, Petia Topolova, Peter Wafula Nasokho, Owen Ozier, Maureen Wechuli, and the GSP field staff and data group, without whom the project would not have been possible. George Akerlof, David Card, Rachel Glennerster, Brian Jacob, Matthew Jukes, Victor Lavy, Michael Mills, Antonio Rangel, Joel Sobel, Doug Staiger, and many seminar participants have provided valuable comments. We are grateful for financial support from the World Bank and MacArthur Foundation. All errors are our own.

1. Introduction

This paper estimates the impact of a merit scholarship program in Kenyan primary schools. The scholarship schools were randomly selected from among a group of candidate schools, allowing differences in educational outcomes between program and comparison schools to be attributed to the scholarship. In our main result, we find that girls in the program schools had significantly higher test scores than comparison school girls. Moreover, test scores in program schools were higher for boys, who were ineligible for the scholarship, as well as for girls with low pretest scores (who were very unlikely to win). Both student and teacher attendance increased significantly in program schools.

These results have implications for understanding the nature of educational production. While most education research focuses on the effect of material inputs, class size, or school organization, the most important input in the education production function may be student effort. Our results suggest not only that effort is responsive to incentives but also that there are positive externalities from student effort, and perhaps strategic complementarity in effort choices.

The findings also speak to the current debate over merit scholarships. Merit scholarship funds have grown by almost 50% in the U.S. during the past five years (College Board 2002), but many educators have opposed them on equity grounds, fearing that benefits would disproportionately flow to well-off pupils and thus exacerbating inequality (Orfield 2002). On the other hand, our findings indicate that the benefits to offering merit-based scholarships may be large, especially in less developed countries where educational needs are great but funding is extremely limited. Recent education finance reforms in many African countries have focused on reducing the cost of education across the board by eliminating primary school fees (United Nations 2003) or by subsidizing primary school attendance, as in Mexico's PROGRESA program. However, for many of the poorest countries, it is impossible to provide free universal schooling in the short run; in Kenya, for example, we estimate financing secondary schooling to the entire age cohort at 18% of GDP (based on figures in World Bank 2004).

The evidence of positive externalities to individual study effort creates a new rationale for merit scholarships, and addresses some concerns regarding equity. In our Kenyan context, the existence of large externality benefits for students at the bottom of the baseline test score distribution implies that merit scholarships could be justified even under a social welfare function that put no weight on gains for students in the top half of the test distribution. Moreover, human capital externalities in production are often cited as a justification for government education subsidies more generally since private education investment in the absence of such subsidies would be inefficiently low (Lucas 1988). Although recent empirical studies find that human capital externalities in the labor market are small if they exist at all (Acemoglu and Angrist 2000, Moretti 2004), our results suggest it may be that the largest positive externalities from education investments occur earlier within the classroom, providing further justification for government education subsidies.

We find little evidence supporting other common criticisms of merit scholarships. There is no evidence that program incentives weakened the intrinsic motivation to learn in school, based on surveys of students. While standard economic models suggest incentives should increase individual study effort, an alternative theory from psychology asserts that extrinsic rewards may interfere with intrinsic motivation and could reduce effort in some circumstances.[1] A weaker version of this view is that incentives lead to better performance in the short-run, but have negative effects after the incentive is removed by weakening intrinsic motivation. We find no evidence for this when we examine test scores in the years following the scholarship competition, or at least we find that any reduction in intrinsic motivation was offset by other factors. Similarly, there are no statistically

[1] See Benabou and Tirole (2003). Early experimental psychology research in education supported the idea that reward-based incentives lead to increased effort in students (Skinner 1961). However, laboratory research conducted in the 1970's studied behavior before and after pupils received "extrinsic" motivational rewards and found that these external rewards produced negative impacts in some situations (Deci 1971; Kruglanski et al. 1971; Lepper et al. 1973). Later laboratory research attempting to quantify the effect of external factors on intrinsic motivation has yielded mixed conclusions: Cameron et al. (2001) conducted meta-studies of over 100 experiments and found that the negative effects of external rewards were limited and could be overcome in certain settings – such as for high-interest tasks – but in a similar meta-study Deci et al. (1999) conclude that there are often negative effects of rewards on task interest and satisfaction. The current study differs from much of the existing work by estimating impacts in a real-world context rather than the laboratory, and by exploring spillover effects on third parties.

significant changes in students' self-expressed attitudes toward school, toward their own academic ability, or in their time use outside of school.[2]

The scholarship program we study also does not appear to have led students to focus on test performance at the expense of other dimensions of learning. This stands in sharp contrast to another project conducted in Kenya which provided incentives for teachers based on students' test scores. That teacher incentive program had no measurable effect on either student or teacher attendance, but increased the frequency of extra test preparation tutoring sessions (Glewwe et al. 2003). Students' scores increased on the exam for which the teacher incentives were provided but did not remain high afterwards. In contrast, in the student merit scholarship program both student school participation and teacher school attendance increased in program schools, test score gains remained large in the year following the competition, and there was no increase in test preparation tutoring.

In the work most closely related to the current study, Angrist and Lavy (2002) find that cash awards raised test performance among high school students in Israel.[3] They examine a scholarship program that provided cash for good matriculation exam performance in twenty schools. Students offered the merit award were 6-8 percentage points more likely to pass exams than comparison

[2] There is some evidence that in other contexts students may change their curriculum in response to merit scholarships, but in the setting of this study, the curriculum is fixed. Binder et al. (2002) show that while scholarship eligibility in New Mexico increased student grades, the number of credit-hours students completed decreased, suggesting that students took fewer courses in order to keep up their grades. Similarly, after the HOPE college scholarship program was introduced in Georgia, the average SAT score for high school seniors rose almost 40 points (Cornwell et al. 2002), but there was a 2% average reduction in completed college credits, a 12% decrease in full course-load completion, and 22% increase in summer school enrollment (Cornwell et al 2003).

[3] Leuven et al (2003) also use an experimental design, to estimate the effect of a financial incentive on the performance of Dutch university students. They similarly estimate large positive effects, but their small sample size limits statistical precision, complicating inference. Ashworth et al. (2001) study Education Maintenance Allowances (EMA), weekly allowances given to 16-19 year old students from low-income U.K. households based on school enrollment and academic achievement. Initial findings indicate that EMA raised school enrollment among eligible youth by 6 percentage points and by 4 percentage points among the ineligible, suggesting externalities. It is unclear how much of these impacts are due to rewarding students for enrollment versus achievement. Since program areas were not randomly selected – EMA was targeted to poor urban areas – the authors resort to propensity score matching to estimate impacts. Croxford et. al. (2002) find similar EMA impacts in Scotland. Angrist et al (2002) show that a Colombian program that provided private school vouchers to students conditional on their maintaining a satisfactory level of academic performance led to academic gains, although it is unclear how much of this impact came from the expanded range of school choice participants experienced, and how much from the incentives. A number of studies suggest university merit scholarships increase enrollment (see Dynarski 2003).

students in a pilot program that randomized awards among schools, with the largest effects among the top quartile of students. A second pilot that randomized awards at the individual level within a different set of schools did not produce significant impacts.

Our study differs from the Israel study in several important ways. First, by examining a program in which scholarships were randomized at the school level, we are able to estimate externality impacts of increased student effort. Our results suggest the possibility that the weak program impacts estimated in Angrist and Lavy's second pilot comparing students within schools could in part be due to positive classroom externalities in student effort, which would contaminate the experiment. Second, due to political and logistical issues, the program in Israel and its evaluation, which was meant to run for three years, were discontinued after the first year, making it impossible to estimate longer-term impacts and impacts once the incentive was removed. Third, the sample in our study includes more than three times as many schools, and our program and comparison groups are more balanced on observable characteristics. Finally, in addition to test score outcomes, we collected data on student school attendance, teacher attendance, purchases of school supplies, student time use, and a range of student attitudes which allow us to explore potential mechanisms through which merit scholarships affect test scores.

The paper proceeds as follows. Section 2 provides information on schooling in Kenya and the scholarship program. Section 3 discusses incentives, externalities, and study effort. Section 4 presents the data and estimation strategy, section 5 discusses the results, and section 6 compares the cost effectiveness of scholarships to other programs. The final section concludes.

2. The Girls Scholarship Program

2.1 Schooling in Kenya

Schooling in Kenya consists of eight years of primary school followed by four years of secondary school. While approximately 85% of children of primary school age in western Kenya are enrolled in

(Central Bureau of Statistics 1999), there are high dropout rates in grades 5, 6, and 7 and only about one-third finish primary school. Dropout rates are especially high for teenage girls.[4]

Secondary school admission depends on performance on the grade 8 government Kenya Certificate of Primary Education (KCPE) exam. To prepare, grade 4-8 students take exams at the end of each school year. These exams are standardized for each district and test students in English, geography/history, mathematics, science, and Swahili. Students must pay a fee to take the exam, US$1-2 depending on the year; we discuss implications of this fee below. Kenyan district education offices have a well-established system of exam supervision, with outside monitors for the exams and teachers from each school playing no role in either supervision or grading. Exam monitors document and punish any instances of cheating, and report these cases to the district office.

During the period when the scholarship program we study was introduced, primary schools charged school fees to cover their non-teacher costs, including textbooks, chalk, and classroom maintenance. These fees averaged approximately US$6.40 (KSh 500)[5] per family each year. In practice, while these fees set a benchmark for bargaining between parents and headmasters, most parents did not pay the full fee. In addition to this fee, there were also fees for activities, such as taking exams, as well as costs of school supplies, certain textbooks, and uniforms (approximately US$6.40). The scholarship project was introduced by the NGO in part to assist families of high-scoring girls to cover these costs.

In late 2001, then-president Daniel Arap Moi announced a national ban on primary school fees, but the central government did not provide alternative sources of school funding and other policymakers made unclear statements on whether schools could impose "voluntary" fees. Schools varied in the extent to which they continued collecting fees in 2002, but this is difficult to quantitatively assess. Mwai Kibaki became president of Kenya following the December 2002

[4] For instance, girls in our baseline sample (in comparison schools) had a dropout rate of 9% from January 2001 through early 2002, versus 6% for boys.
[5] One US dollar was worth 78.5 Kenyan shillings (KSh) in January 2002 (Central Bank of Kenya 2002).

elections and eliminated primary school fees in early 2003. This policy was followed by primary school committees, in part because the national government made substitute payments to schools to replace local fees. Our study focuses on program impacts in 2001 and 2002 before primary school fees were eliminated by the Kibaki reforms.

2.2 Project Description and Timeline

The Girls Scholarship Program (GSP) we study was carried out by a Dutch non-governmental organization (NGO), called ICS Africa, in two rural districts in western Kenya, Busia and Teso. Busia district is mainly populated by a Bantu-speaking ethnic group (Luhyas) with agricultural traditions while Teso district is populated primarily by a Nilotic-speaking group (Tesos) with pastoralist traditions. These groups differ in language, history, and certain present-day customs, although not along observed household assets.

Half of the 127 sample primary schools were randomly invited to participate in the program in March 2001. The randomization first stratified schools by administrative divisions[6] and by participation in a past NGO assistance program which provided classroom flip charts.[7] Randomization was then carried out using a computer random number generator which successfully created treatment groups comparable along observable characteristics (discussed below).

The program provided incentives for students to excel on academic exams. Scholarship winners from grade 6 were chosen based on their total score across the five subject tests. The NGO awarded scholarships to the highest scoring 15% of grade 6 girls in the program schools within each district (amounting to 110 girls in Busia and 90 in Teso). Schools varied considerably in the number of winners, and 57% of program schools (36 of 63 schools) had at least one 2001 winner; among

[6] Divisions are subsets of districts, with eight divisions in all within Busia and Teso districts.
[7] All GSP schools had previously participated in an evaluation of a flip chart program, and are a subset of that sample. These schools are representative of local primary schools along most dimensions but exclude some of the most advantaged as well as some of the worst off – see Glewwe et al. (2004) for details on the sample. The flip chart program did not affect any measures of educational performance.

schools with at least one winner, there was an average of 5.6 winners per school. In January 2002 the NGO held school assemblies for students, parents, teachers, and local government officials to announce and publicly recognize the 2001 winners.

The scholarship program provided winning grade 6 girls with an award for the next two academic years, grades 7 and 8 (through the end of primary school). In each year, the award consisted of: (1) a grant of US$6.40 (KSh 500) intended to cover the winner's school fees, paid to her school; (2) a grant of US$12.80 (KSh 1000) intended for school supplies paid directly to the girl's family; and (3) public recognition at the school awards assembly. Although there was no enforcement to make sure that parents spent the grant on school supplies, the fact that the money was presented to parents in a public ceremony may have generated some community pressure to use the money in ways that benefited their daughter's education.[8] Since many parents would not otherwise have fully paid school fees, schools with winners benefited to some degree from the award money paid directly to the school. Some of these funds may have also benefited teachers if they were used to improve the staff room, for instance, although the amounts dedicated to this were likely small.

Two cohorts of grade 6 girls competed for scholarships. Girls registered for grade 6 in January 2001 in program schools were the first eligible cohort (cohort 1) and those registered for grade 5 in January 2001 made up the second cohort (cohort 2), competing in 2002. The NGO restricted eligibility to girls already enrolled in program schools in January 2001 before the program was announced. Thus there was no incentive for students to transfer schools, and incoming transfer rates were low and nearly identical in program and comparison schools (not shown).

Cohort 1 students took end-of-year grade 5 exams in November 2000, and these are used as baseline test scores in the analysis.[9] In March 2001, the NGO met with the headmasters of schools

[8] It is impossible to formally test how the funding was actually spent without detailed household consumption expenditure data, which we do not have.

[9] A detailed project timeline is presented in Appendix Table A. Unfortunately, there is incomplete 2000 baseline exam data for cohort 2 (when these students were in grade 4), especially in Teso district where most schools did not

selected for the program to give each school community the choice to participate. Headmasters were asked to relay information about the program to parents via a school assembly and in September and October the NGO held additional community meetings to reinforce knowledge about program rules in advance of the November 2001 district exams. After these meetings, NGO enumerators began collecting school attendance data during unannounced visits.

It is worth briefly characterizing the students who performed well on the 2001 exam. The baseline 2000 test score is a very strong predictor of being a top 15% performer on the 2001 test in both program and comparison schools as expected. Children whose parents have more schooling are also more likely to be in the top 15% of test performers in program schools than in comparison schools: average years of parent education are nearly three years greater for scholarship winners than losers (7.7 years for winners versus 4.8 years for non-winners), and this large effect is statistically significant at 99% confidence. The link between family education background and test score is somewhat stronger in the program schools (regressions not shown). These patterns are consistent with a model in which test scores depend on effort and family background (as well as other factors) and there is considerable heterogeneity in effort levels in the absence of a merit scholarship, but much less heterogeneity under a scholarship program. However, note there is no statistically significant difference between winners and non-winners in household ownership of iron roofs or latrines (regressions not shown), suggesting that children from wealthier households in terms of asset ownership were no more likely to win.

Official exams were again held in late 2002 in Busia district. The 2002 exams in Teso district were canceled because of possible disruptions in the run-up to upcoming 2002 national elections, so the NGO instead administered its own standardized exams in February 2003. Thus the second cohort of winners were chosen in Busia based on the official 2002 district exam, while Teso winners were

---

offer a grade 4 exam, and thus baseline comparisons focus on cohort 1. School average 2000 scores for cohort 1 students are used to control for baseline differences in some specifications. The test score data is discussed in greater detail in section 4.1 below.

chosen based on the NGO exam. In this second round, 70% of program schools (44 of 63 schools) had at least one winner, an increase over 2001, and in all 78% of program schools had at least one winner in either 2001 or 2002.

In terms of data collection, the NGO visited all schools during 2002 to conduct unannounced attendance checks and administer questionnaires to grade 5-7 students, collecting information on study effort, habits, and attitudes toward school. The student survey data indicates that most girls understood program rules: 89% of cohort 1 and 2 girls in Busia district claimed to have heard of the program, and knowledge levels were similar in Teso district (86%). Girls had somewhat better knowledge about program rules governing eligibility and winning than boys: Busia girls were 7 percentage points more likely than boys to know that "only girls are eligible for the scholarship" (86% for girls versus 79% for boys), although the proportion among boys is still high, suggesting that the vast majority of boys knew that they were ineligible, and patterns are again similar in Teso (not shown).[10] Girls were very likely (70%) to report that their parents had mentioned the program to them, suggesting some parental encouragement.

2.3 Busia and Teso Districts

In June 2001 lightning struck a Teso district primary school not in our sample, severely damaging the school, killing seven students, and injuring 27 others. Because the NGO had been involved with another assistance program in that school, and due to strange coincidences – for instance, the names of several victims were the same as NGO staff members who had recently visited the school – the deaths were associated with the NGO in the eyes of some community members, and the incident led some schools to pull out of the girl's scholarship program: of the original 58 sample schools in Teso

[10] Note that random measurement error is likely to be reasonably common for these survey responses, since rather than being filled in by an enumerator who individually interviewed students, the surveys were filled in by students with the enumerator explaining the questionnaire to the class as a whole; thus values of 100% are unlikely even if all students had perfect program knowledge.

district, five pulled out at that time, and one Busia school located near the Teso district border also pulled out. Figure 1 presents the location of the lightning strike and of the schools that pulled out, four of which are located very near the lightning strike. Three of the six schools that pulled out of the program were treatment schools, and three were comparison schools. Moreover, one girl in Teso district who won the ICS scholarship in 2001 refused the scholarship award, reportedly because of negative views toward the NGO. We discuss implications for econometric inference in Section 4.

Structured interviews were conducted during June 2003 with a representative sample of 64 teachers in 18 program schools, and these confirm the stark differences in program reception across Busia and Teso districts, probably in part due to the lightning strike. When teachers were asked to rate local parental support for the program, 90% of the Busia teachers claimed that parents were either "very positive" or "somewhat positive" but the analogous rate in Teso was only 58%, and this difference across the districts is statistically significant at 99% confidence.

Historically, Tesos were educationally disadvantaged relative to Luhyas, with fewer Teso secondary school graduates, for example. Project survey data confirms this disparity between the districts: Teso district parents have 0.4 years less schooling than Busia parents on average. There is also a tradition of suspicion of outsiders in Teso district, and this has at times led to misunderstandings between NGO's and some people there. It has been claimed that indigenous religious beliefs, traditional taboos and witchcraft practices remain stronger in Teso than in Busia (Government of Kenya 1986), and this underlying cultural environment likely exacerbated the negative local reaction following the deadly lightning strike.

3. Incentives, Externalities, and Study Effort

A stylized framework illustrates several channels through which merit awards could affect academic test scores. The key behavioral change induced by a merit award is likely to be increased study effort. The program we study directly boosted incentives for girls to exert study effort by providing the

award. Although the monetary value of the award was identical everywhere, note that local social prestige associated with winning may have differed across communities.[11]

Beyond individual effort, academic performance may also be a function of the study effort of other students in the class, and a function of teacher effort. (We ignore other inputs into educational production, e.g., textbooks, below for simplicity.) The efforts of a student and her classmates, and a student and her teacher could theoretically be either complements or substitutes. Similarly, own effort and current academic ability may be either complements or substitutes, and thus own effort at one point in time may complement or substitute effort at other times. It is plausible that own effort, effort of other students, and teacher effort are complements in practice, in which case programs increasing effort for some students would generate multiplier effects in average individual effort. This allows for the possibility of multiple classroom equilibria, some with high effort by students and others with a poor overall learning environment.

Related arguments suggest that teachers in program schools could also exert more effort than those in comparison schools. If teachers experience benefits from having more scholarship winners in their class – e.g., ego rents, social prestige in the community, or even gifts from parents – then they should increase their work effort. This might take the form of attending school more regularly or improving their lesson plans. If there are student-teacher complementarities to effort in educational production, teachers might also find additional effort more attractive when their students are putting more effort into their studies.

Another form of extra motivation for teachers in scholarship program schools could be social sanctions against shirking teachers on the part of parents or the headmaster. Such sanctions might differ across communities as a function of local parent support for the program. In that case, the merit

[11] Field interviews conducted by the authors in July 2002 indicate that girls actively competed for the scholarship. One headmaster reported that the program "awakened our girls and was one step towards making the girls really enjoy school." One winning girl who was asked about her own performance versus those students who did not win remarked, "they tried to work hard for the scholarship but we defeated them." It is plausible that this spirit of competition drove some girls to work harder, providing utility benefits beyond the monetary awards.

award would generate larger gains where parents are more supportive, and this may have led to differences across Busia and Teso districts where program reception was very different. The June 2003 structured interviews with teachers provide some evidence on how parental support may have contributed to program success. One Busia teacher mentioned that after the program was introduced, parents began to "ask teachers to work hard so that [their daughters] can win more scholarships." A teacher in a different Busia school asserted that parents visited the school more frequently to check up on teachers, and to "encourage the pupils to put in more efforts."

Pupils in program schools who are ineligible for awards (i.e., boys), or who are eligible but have little realistic change to win (i.e., girls with low initial academic ability), might benefit from the program through several channels. First, greater effort by classmates competing for the merit award could improve the overall classroom learning environment and boost scores directly through a peer effect. Second, all students could benefit from increased teacher effort, to the extent this effort is not targeted solely at the girls with a chance of winning.[12] Of course, it is possible that program school teachers responded to the program by diverting effort away from ineligible students toward students who were eligible – for example, by calling on girls more than boys in class – but we find no evidence of this below. Third, to the extent that student effort complements classmates' and the teacher's effort in educational production, these children might optimally exert additional effort even without direct incentives (as in Lazear 2001).

4. Data and Estimation

4.1 The Dataset

Test score data were obtained from the District Education Offices (DEO) in Busia district and Teso district. Test scores were normalized in each district such that scores in the comparison sample (girls

[12] The July 2002 field interviews conducted by the authors suggest that a desire to compete with girls also drove some boys to study harder. To the extent that this "gendered" competition was an important determinant of boys' gains in program schools, it is an open question how large externality gains would be under an alternative program that targeted boys rather than girls, or in which they competed against each other for the same awards.

and boys together) are distributed with mean zero and standard deviation one. The complete dataset with both cohort 1 and cohort 2 students enrolled in school in January 2001 is called the *baseline sample* (Table 1, Panel B). In the main analysis, we focus primarily on students in schools that did not pull out of the program and for which we have mean school baseline 2000 test scores, and call this the *restricted sample* (Panel C). Average test scores are slightly higher in the restricted sample than in the baseline sample, since the students excluded from the sample are typically somewhat below average in academic achievement, as discussed below. The *longitudinal sample* contains the restricted sample cohort 1 students with individual 2000 baseline test scores. Note that 2000 test scores are missing for most cohort 2 students in Teso district because many schools there did not offer grade 4 exams, so for cohort 2 we focus on the 2002 exam.

Following the lightning strike, six of the 127 schools invited to participate decided to pull out of the program, leaving 121 schools. Five additional schools, three in Teso district and two in Busia, have incomplete exam scores for 2000, 2001 or 2002, leaving 116 schools and 7,258 students in the restricted sample. The restricted sample thus contains data for 91% of schools in the baseline sample. Program school students account for 51% of the restricted sample.

School participation data are based on four unannounced checks, one conducted in September or October 2001 and one in each of the three terms of the 2002 academic year. Collected by NGO enumerators, these data record baseline students who were actually in school on the day of the unannounced check. School participation rates are somewhat below 80% for the baseline sample and approximately 85% for the restricted sample (Table 1, Panels B and C). We use data from these unannounced checks rather than official school attendance registers, since registers are often unreliable. Finally, student surveys were collected in 2002 from all cohort 1 and cohort 2 students present in school at the time of data collection.

4.2 Estimation Strategy

We focus on reduced form estimation, in other words, the impact of the program on test scores. We also estimate program impacts on multiple channels possibly linking individual behavior to test scores – in particular, measures of student and teacher effort – to better understand the mechanisms underlying the reduced form test score estimates. The main estimation equation is:

$$TEST_{ist} = Z_{ist}'\beta + (Z_{ist} * T_s)'\gamma + \delta X_{ist} + \mu_s + \varepsilon_{ist} \quad (1)$$

$TEST_{ist}$ is the normalized test score for student i in school s in year t. Z_{ist} is a vector of demographic indicator variables for gender, or for each cohort and year (i.e., cohort 1 in year 1, cohort 1 in year 2, etc.), and T_s is the program school indicator. In specifications where the goal is to estimate the overall program impact across all cohorts and years, we exclude the $Z_{ist}*T_s$ terms and instead include the program indicator. X_{ist} denotes the average school baseline (2000) test score when we use the restricted sample, and denotes the individual baseline score for the longitudinal sample. The error terms consist of μ_s, a school effect perhaps capturing common local or headmaster characteristics, and an idiosyncratic term, ε_{ist}, which captures unobserved student ability or shocks.

In addition to equation (1), we use non-parametric locally weighted regressions (Fan 1992) to estimate average program impacts across pupils with different baseline scores, as well as to analyze sample attrition patterns.

4.3 Sample Attrition

One potential threat to the validity of the analysis is the effect of the program on sample attrition. We find large differences in attrition across program and comparison schools in Teso district, but not in Busia district. In Busia differences between program and comparison schools are small and not statistically significant: for cohort 1 79% of girls (76% of boys) in Busia program schools and 78%

of girls (77% of boys) in Busia comparison schools took the 2001 exam (Table 2). Among cohort 2 students in Busia, there is again almost no difference between the program school students and comparison school students in the proportion who take the 2002 exam (50% versus 48% for girls, and 50% versus 52% for boys). For both program and comparison students, there is more attrition by 2002 as students drop out of school, transfer to other schools, or decide not to take the exam.

Attrition patterns in Teso district schools, however, are strikingly different. For cohort 1, 53% of girls in program schools (54% of boys) took the 2001 exam, but the rate for comparison school girls is much higher, at 65% (and similarly high for boys, at 66%, Table 2). [13] There are also attrition gaps across program and comparison schools in Teso among cohort 2, although these are smaller than among cohort 1 students.[14]

Non-parametric Fan locally weighted regressions – presenting the proportion of cohort 1 students taking the 2001 exam as a function of their baseline 2000 test score – indicate that Busia district students across all levels of initial academic ability have a similar likelihood of taking the 2001 exam and remaining in the sample (Figure 2, Panels A and B). Although theoretically the introduction of a scholarship could have induced poor but high-achieving students to take the exam in program schools, leading to an upward bias in estimated program impacts, we do not find evidence of this in either Busia district or Teso district. Rather students with low initial achievement are somewhat more likely to take the 2001 exam in Busia program schools relative to comparison schools, and this difference is statistically significant in the left tail of the baseline 2000 distribution. This slightly lower attrition rate among low achieving Busia program school students may lead to a downward bias (toward zero) in estimated program effects there, but these figures suggest any bias is likely to be small.

[13] The rates in Table 2 exclude schools that pulled out of the program entirely from the calculation, but differential attrition patterns are even more pronounced using that data (not shown). We present program impact estimates using all 127 baseline schools in Table 5 Panel B below.

[14] Lower 2002 attrition in Teso is likely because the NGO administered its own exam there in early 2003 and students did not need to pay a fee to take the exam, unlike the government test in earlier years (see main text above).

In contrast, again in Teso district not only were attrition rates high and unbalanced across treatment groups for cohort 1, but significantly more high-achieving students took the 2001 exam in comparison schools relative to program schools, and this is likely to strongly bias estimated program impacts toward zero in Teso (Figure 3, Panels A and B). To illustrate, among high ability girls in Teso with a score of at least +1 standard deviation on the baseline 2000 exam, comparison school students were over 20 percentage points more likely to take the 2001 exam than program school students, and this difference is statistically significant at 95% confidence. The comparable gap among high ability Busia girls is near zero and not statistically significant. There are similar though less pronounced gaps between comparison and program schools for Teso district boys (Panel B). Pooling boys and girls, in Teso program schools students who did not take the 2001 exam scored 0.05 standard deviations lower at baseline on average (on the 2000 test) than those who took the 2001 exams, but the difference is 0.57 standard deviations lower in the Teso comparison schools, and the estimated difference-in-differences is significant at 95% confidence (regression not shown). These attrition patterns in Teso may be due to some high-achieving pupils in program schools feeling especially "vulnerable" to the program in communities where there was mistrust of the NGO, since they were more likely to win an award, as well as to the fact that several schools that pulled out of the program had high average baseline 2000 test scores.

Pupils with high baseline 2000 test scores were much more likely to win an award in 2001 in both Busia and Teso districts, as expected: cohort 1 girls with below average baseline test scores had a miniscule change of winning (Figure 4), but the likelihood of winning rises monotonically and rapidly with the baseline score. The proportion of cohort 1 program school girls taking the 2001 exam as a function of the baseline score (Figure 2 Panel A and Figure 3 Panel A) does not correspond closely to the likelihood of winning an award in either study district. This pattern, together with the very high rate of 2001 test taking for boys and for comparison school girls, all suggest that competing for the NGO award was not the main reason most students took the test.

To summarize, Teso district primary schools had higher rates of sample attrition than Busia schools in 2001, the gap in attrition across the program versus comparison schools was large in Teso district but zero in Busia, and a much higher proportion of high ability students (according to baseline exam scores) took the exam in Teso district comparison schools than Teso program schools, likely biasing program impact estimates toward zero. Thus in much of the following analysis we focus only on Busia district where attrition bias is minimal.

4.4 Other Estimation Issues

Household characteristics are similar across Busia district program and comparison schools (Table 3): there are no significant differences in parent education, number of siblings, proportion of ethnic Luhyas or the ownership of a latrine, iron roof, or mosquito net, using data from the 2002 student surveys, indicating that the randomization was largely successful in creating comparable groups. While there is no comparable pre-program survey data, it is reasonable to assume that the characteristics we examine (e.g., parent education) were stable between 2001 and 2002 and not affected by the program. The 2000 baseline test score distributions provide further evidence on the comparability of the program and comparison groups. Formally, we cannot reject the hypothesis that mean baseline test scores are the same across program and comparison schools for either girls or boys. The two distributions are similar graphically (Figure 5), and we cannot reject equality of the distributions using the Kolmogorov-Smirnov Test (p-value = 0.33 for cohort 1 Busia girls).

Baseline characteristics are also similar across program and comparison schools in the Teso district restricted sample, but there are certain statistically significant differences, including the mean 2000 test score (results not shown), likely due to the different attrition patterns across Teso program and comparison schools.

Another concern related to the interpretation of our findings is the possibility of cheating on the exams but this appears unlikely. First, district records from the outside exam monitors indicate

there were no documented instances of cheating in any sample school during either 2001 or 2002. Several findings reported below also argue against cheating: test score gains among cohort 1 students in scholarship schools persisted a full year after the exam competition when there was no longer any direct incentive to cheat, and there were substantial gains among program school boys ineligible for the scholarship who had no direct program benefits from cheating (although cheating by teachers could still potentially explain that pattern). There are also program impacts on several objective measures of student and teacher effort, most importantly school attendance.

A final issue is the Hawthorne effect, namely, an effect since students knew they were being studied rather than due to the intervention per se, but this too is unlikely for at least two reasons. First, both program and comparison schools were visited frequently on unannounced days to collect data, and thus mere contact with the NGO and enumerators alone cannot explain effects. Moreover, five other primary school program evaluations have been carried out in the study area (as discussed in Section 6) but in no other case did a program lead to such substantial test score gains.

5. Empirical Results

5.1 Academic Test Score Impacts

We find large and statistically significant test score gains for girls in program schools and these gains continue past the scholarship competition year. There are also significant gains among those who had no chance of winning the scholarship – boys and poorly-performing girls at baseline – evidence of positive program spillovers. We first present graphical evidence on program impacts and then the regression estimates.

Baseline 2000 test score distributions are similar across program and comparison schools in Busia district, for both girls and boys (Figure 5). The test score distribution in program schools shifts markedly to the right for cohort 1 girls and boys in the first year of the program (Figure 6), cohort 1 girls and boys in the first year post-competition (Figure 7), and for cohort 2 girls and boys in year 2

when they were competing for the scholarship (Figure 8).[15] The vertical lines in these figures indicate the minimum score necessary to win an award each year. The sample for Figures 5, 6, and 7 is the cohort 1 longitudinal sample, namely, those restricted sample students who have 2000 individual test scores – and thus the samples in Figures 5 and 6 are identical, while sample size falls somewhat for the 2002 results in Figure 7. Figure 8 presents cohort 2 restricted sample students.

In the case of Busia girls, gains are most pronounced for students at two parts of the distribution: first, for those near the minimum winning score threshold – consistent with the view that students exerting the most additional effort were those who believed extra effort would make the greatest difference to their chances of winning – and second, in the left tail of the distribution (Figure 6 Panel A). Test gains are not as visually large for boys, but there are perceptible shifts in both the left and right tails of the program school distributions (Figure 6 Panel B).

The regression analysis first examines the cohort 1 longitudinal sample (the same sample in Figures 5 and 6). The program raised test scores by 0.12 standard deviations on average among girls and boys in 2001 and 2002, among all Busia and Teso district students from both cohorts (Table 4, regression 1). The average impact rises slightly to 0.13 standard deviations (standard error 0.06, regression 2) and becomes statistically significant at 95% confidence when the individual baseline 2000 test score is included as an explanatory variable, as this baseline control reduces residual variation. The 2000 test score is strongly related to the 2001 test score as expected (point estimate 0.80, standard error 0.02). Estimated program impacts are nearly identical for girls and boys (regression 3). Note that boys score much higher than girls on average in the longitudinal sample, with a gender gap of 0.16 standard deviations (standard error 0.04) even with the inclusion of the individual baseline test control.

While the pooled Busia and Teso effect is positive and significant, the estimated impact in the longitudinal sample is considerably larger for Busia district (0.19 standard deviations, standard

[15] These figures use an epanechnikov kernel and a bandwidth of 0.7.

error 0.08, Table 4, regression 4) than for Teso district (-0.02 standard deviations, regression 5). This is consistent with the direction of the hypothesized sample attrition bias in Teso district. Program impact point estimates increase slightly in both districts when the individual baseline test control is not included as an explanatory variable (in a specification analogous to Table 4, regression 1) but standard errors increase sharply: the estimate for Busia district schools is 0.22 standard deviations (standard error 0.19) and for Teso becomes positive but small at 0.08 standard deviations (standard error 0.15 – regressions not shown).

We next construct non-parametric bounds on program effects using the trimming method in Lee (2002). The bounds for Busia schools are tight since there was essentially no differential attrition across program groups there (Table 2), but the bounds for cohort 1 girls in Teso district are wide, ranging from -0.24 standard deviations as a lower bound up to 0.22 standard deviations as an upper bound, in a specification analogous to Table 4, regression 5. Using this conservative method, it is impossible to draw firm conclusions about program treatment effects in Teso district in the presence of sample attrition.

In a second attempt to better characterize the likely bias due to sample attrition in Teso, we impute missing 2001 test scores among longitudinal sample students as a linear function of their 2000 score. This suggests that in the absence of attrition, program impacts for cohort 1 girls in Teso would have been positive and reasonably large: the estimated impact for Teso district girls using this imputation becomes 0.12 standard deviations (standard error 0.14 – regression not shown) when the individual baseline 2000 score is not included as a control. However, this estimate is of course only suggestive given possible omitted variable biases. Another approach for addressing sample attrition bias in Teso district is to focus on impacts for cohort 2 girls, since attrition rates are similar for them in both program and comparison schools (Table 2), although we are unable to determine exact attrition patterns due to their lack of baseline 2000 data. The estimated impact is near zero and not significant (estimate 0.00 standard deviation, standard error 0.11 – regression not shown), evidence

that program impacts were probably negligible in Teso. Whatever interpretation is given to the Teso district results – either unreliable estimates due to attrition or no program impact – the program was less successful in Teso at a minimum in the sense that fewer schools chose to take part.

Program effects at different regions of the initial 2000 test score distribution are estimated using a non-parametric Fan locally weighted regression (with bootstrapped standard errors clustered by school, Figure 9). Busia girls just below the winning threshold had large test score gains, as suggested by the previous figures, and there are also marked and statistically significant gains at the bottom of the baseline distribution (Figure 9 Panel A). This is evidence of positive spillover benefits of the program – girls with below average baseline scores have essentially zero chance of winning and so their gains are unlikely to be the result of attempts to win the award. However, due to limited statistical power it is impossible to reject the hypothesis that program impacts at the bottom of the baseline distribution are the same as gains elsewhere. Busia boys show similar patterns (Panel B) though their gains at the top of baseline distribution are somewhat more pronounced than for girls.

Extending the analysis to the restricted sample for both cohort 1 and cohort 2 girls and boys again pooling Busia and Teso district schools, we find an overall impact of 0.10 standard deviations (Table 5, Panel A, regression 1). The overall program effect remains 0.10 standard deviations (standard error 0.05, regression 2) and becomes statistically significant at 90% confidence when the mean school 2000 test score (computed among students in the restricted sample) is included as an explanatory variable. The average program effect for girls, pooled for Busia and Teso schools, remains large and statistically significant in the restricted sample at 0.14 standard deviations (standard error 0.06, regression 3), but the average effect for boys falls to 0.07 standard deviations – the most noteworthy difference between the results for the longitudinal sample (Table 4) versus the restricted sample (Table 5, Panel A).

The average program impact for Busia district girls is 0.25 standard deviations (standard error 0.07, statistically significant at 99% confidence – Table 5, Panel A, regression 4) [16], again much larger than the estimated effect for Teso girls, at -0.02 standard deviations (regression 5). The estimated effect for Busia boys is reasonably large and marginally statistically significant at 0.13 standard deviations (standard error 0.07, significant at 90% confidence), while the analogous effect for Teso boys is near zero. The externality effects for Busia boys suggest that merit award programs that randomize eligibility within schools (as in one Angrist and Lavy 2002 pilot) could systematically understate impacts. For instance, here the gap between girls and boys is only 0.12 standard deviations, less than half the magnitude of our preferred program impact estimate for Busia girls.

Point estimates are broadly unchanged using the full baseline sample containing test score data for all 127 of the original schools, in an intention to treat (ITT) analysis. These regressions do not include the mean school 2000 test control as an explanatory variable, however, since that data is missing for several schools, and thus standard errors are considerably larger in these specifications[17]. The overall point estimate is 0.12 standard deviations (Table 5, Panel B, regression 1), and is larger for girls at 0.19 standard deviations (standard error 0.12, regression 2) than for boys (0.07 standard deviations). The average program impact for Busia girls is 0.27 standard deviations and nearly statistically significant at 90% confidence (standard error 0.17, regression 3), and smaller for boys at 0.10 standard deviations. The ITT estimated program impact for Teso girls is again positive but not statistically significant at 0.06 standard deviations (standard error 0.15, regression 4). Thus using the larger baseline sample leads to somewhat more positive average estimated program impacts in both Busia and Teso districts, consistent with the hypothesized downward sample attrition bias discussed above (in Section 4.3), but standard errors are much larger in the absence of the baseline test control.

[16] Among Busia restricted sample girls, impacts are somewhat larger for mathematics, science, and geography / history than English and Swahili, but differences by subject are not statistically significant (regression not shown).
[17] A number of the schools added to these specifications have extensive missing data for either the 2001 or 2002 exams (and hence their exclusion from the restricted sample).

Focusing on the Busia district restricted sample we next separately estimate effects for girls and boys across both cohorts and years. The program effect for cohort 1 girls in 2001 – the year these girls were competing for the merit award – is 0.28 standard deviations (standard error 0.10, statistically significant at 99% confidence, Table 6, regression 1), and the effect for cohort 2 in 2002, when they were competing for the award, is 0.21 (standard error 0.10, significant at 95% confidence). These are large impacts: to illustrate with previous findings from Kenya, the average test score for grade 7 students who take a grade 6 exam is approximately one standard deviation higher than the average score for grade 6 students (Glewwe et al 1997). Thus the average estimated scholarship program gain for Busia girls competing for the award corresponds roughly to an additional 0.21-0.28 grades of primary school learning. Notably, these effects are similar to the test score gender gap between boys and girls in the Busia district restricted sample (Table 5). To further illustrate the magnitude, these impacts are similar to the estimated effect of reducing class size by ten students in Israeli primary schools (Angrist and Lavy 1999).

Estimates are unchanged when individual characteristics collected in the 2002 student survey – including student age, parent education, and household asset ownership – are included as additional explanatory variables.[18] Interactions of the program indicator with these characteristics are not statistically significant at traditional confidence levels (regressions not shown), implying that test scores did not increase more on average for students from higher socioeconomic status households.[19] Theoretically, spillover benefits could be larger in schools with more high achieving girls striving for the award. We estimate these effects by interacting the program indicator by various measure of baseline school quality including the mean 2000 test score as well as the proportion of grade six girls in 2000 that were among the top 15% in their district. Neither of these interaction effects are

[18] These are not included in the main specifications because they were only collected for those present in the school on the day of survey administration, thus reducing the sample size and changing the composition of students. Results are also unchanged when school average socioeconomic measures are included as controls (not shown).
[19] Note that although the program had similar test score impacts across socioeconomic backgrounds, students with more educated parents nonetheless were disproportionately likely to win because they have higher baseline scores.

statistically significant at traditional confidence levels (regressions not shown), so we cannot reject the hypothesis that program effects were the same across schools at various academic quality levels.

The program not only raised test scores for cohort 1 girls in Busia district when it was first introduced in 2001 but also continued to boost their scores in 2002: the estimated program impact for cohort 1 girls in 2002 is 0.25 standard deviations (standard error 0.09, statistically significant at 99% confidence, Table 6, regression 1). This suggests that the program had lasting effects on learning, rather than simply being due to cramming for or cheating on exams. The NGO exams administered in February 2003 provide further evidence on post-competition impacts. Although originally administered because 2002 exams were cancelled in Teso district, they were also administered in the Busia sample schools. In the standard specification (like Table 6) the average program impact for cohort 1 Busia girls in early 2003 was 0.19 standard deviations (standard error 0.07, statistically significant at 99% confidence), and the gain for cohort 2 girls is positive and marginally statistically significant at 0.15 standard deviations (standard error 0.08 – regression not shown). Though impacts fall somewhat for cohort 1 over time – from 0.28 standard deviations in the year of the competition (2001), to 0.25 standard deviations the following year, and to 0.19 at the start of the second year after the competition – program impacts are remarkably persistent, and we cannot reject the hypothesis that effects in 2001, the competition year, are equal to the 2002 and 2003 post-competition effects (p-values 0.96 and 0.38, respectively).

As discussed above, boys in Busia program schools also have higher test scores than comparison boys despite not being eligible for the scholarship. The average program impact for cohort 1 Busia boys in 2001 is 0.18 standard deviations (standard error 0.09, statistically significant at 95% confidence, Table 6, regression 2). The survey data suggests that few boys were confused as to whether they too were eligible for the scholarship, so a desire to win the award is unlikely to be driving this. In the second year (2002), there are again positive though not statistically significant

impacts for boys (regression 2), although we cannot reject that effects for boys are the same across both cohorts in 2001 and 2002 at traditional confidence levels.

The focus so far has been program impacts on the first moment of the test score distribution, but program impacts on inequality are also of interest. Point estimates suggest a small overall increase in test score variance for program school girls relative to comparison girls in the restricted sample: the overall variance (including variation both between and within schools) of test scores rises from 0.88 in 2000 at baseline, to 0.94 in 2001 and 0.97 in 2002 for Busia district program school girls, while the analogous variances for Busia comparison girls are 0.92 in 2000, 0.90 in 2001 and 0.92 in 2002; however the difference across the two groups is not statistically significant at traditional confidence levels in any year.[20] The changes in test variance over time for boys in Busia program versus comparison schools are similarly small and not statistically significant (not shown)[21].

5.2 Channels: School Participation, Behaviors and Attitudes

It is useful to explore potential channels for the test score gains since some mechanisms, such as increased test coaching or cramming, might raise test scores without improving actual learning. We consider school participation and test score effects after incentives are removed as two indicators of effort aimed to increase long-run human capital, but treat extra test preparation sessions as having a larger component of effort to increase short-run test scores. While we find evidence that a variety of behaviors change as a result of the program (most notably, student and teacher effort), we are unable to determine which mechanism, if any, is the single leading cause of the test score gains.

[20] The slight (though insignificant) increase in test score inequality in program schools is inconsistent with one particular naïve model of cheating, in which program school teachers simply pass out test answers to their students. This would likely reduce inequality in program relative to comparison schools. We thank Joel Sobel for this point.
[21] One potential concern with these figures is the changing sample size, as different pupils took the 2000, 2001, and 2002 exams. But even if we consider the Busia girls cohort 1 longitudinal sample, where the sample is identical across 2000 and 2001, there are again no significant differences in test variance across program and comparison schools in either 2000 (program girls variance 0.89, comparison 0.92) or in 2001 (0.97 versus 0.89, respectively).

The scholarship program significantly increased school participation measured during unannounced enumerator visits in 2001 and 2002 in Busia district: for cohort 1 and cohort 2 in the restricted sample, the program increased school participation by 4.7 percentage points (standard error 2.5 percentage points, statistically significant at 90% confidence, Table 7, Panel A, regression 1), corresponding to an approximately 30% reduction in absenteeism. Average gains are slightly larger among Busia girls, at 5.0 percentage points (standard error 2.4 percentage points, significant at 95% confidence, regression 2). Since school participation information was collected for all students, even those who did not take the 2001 or 2002 exams, these estimates are less subject to sample attrition bias than test scores. Yet school participation impacts are near zero and not statistically significant in Teso district (estimate -2 percentage points, regression not shown).[22]

The presence of the scholarship program increased average school participation by 6.2 percentage points (standard error 4.2 percentage points, Table 7, Panel A, regression 3) among Busia cohort 1 girls in 2001, and by an even larger 9.4 percentage points (standard error 4.9 percentage points) among cohort 2 in 2001 in a pre-competition effect. School participation gains for Busia girls in 2002 are also positive but smaller (regression 3), though we cannot reject that effects are the same in both years.[23] School participation impacts were not significantly different across school terms 1, 2 and 3 in 2002 (regression not shown), so there is no evidence that gains were concentrated in the immediate run-up to exams due to cramming, for instance. School participation gains are remarkably similar for Busia girls and boys (regressions 3 and 4). We cannot reject that school participation impacts are equal for cohort 1 girls across baseline 2000 test score quartiles (regression not shown).

[22] In the Busia comparison sample, girls with higher baseline test scores have significantly higher average school participation: students above the mean 2000 score were 10 percentage points more likely to be present in school during 2001 than those with below mean scores (standard error 0.01, regression not shown). This is consistent with the view that improved attendance may be a primary channel through which the program generated test score gains.
[23] Although the point estimate goes in the expected direction, there is no significant program effect on the likelihood of dropping out of school by 2002 (regression not shown).

The school participation gains for Busia boys immediately indicate that increases in effort were not simply investments made to increase the probability of winning an award. The observed increase in school participation for low baseline test score girls also allows us to place an upper bound on their expected returns to increased study effort. This bound is extremely low, indicating that any increase in effort (as proxied by school participation) is also unlikely to be due to an attempt to win the award. The probability a program school girl obtained a test score high enough to win an award is a function of her baseline 2000 score. For a girl with a given baseline score, an upper bound on the effectiveness of greater effort in increasing the odds of winning is the probability a girl with that baseline score wins (since the probability of winning cannot fall below zero even at zero effort). Empirically, this upper bound is approximately one percentage point for Busia girls with baseline scores less than the mean of zero (Figure 4 Panel A). Since the scholarship is worth US$38, this means that their expected gain from effort is at most US$0.38. Girls show a 5.0 percentage point average gain in school participation (Table 7), and this translates into roughly 5.0% x 180 school days per year = 9 additional school days. Thus girls competing for the scholarship would only choose to attend school these additional nine days if their productivity at non-school activities was less than US$0.38 / 9 days ≈ US$0.04 per day, and recall that this is an upper bound. This is an implausibly low wage even for teenage girls in rural Kenya, providing further suggestive evidence in favor of externality benefits for low performing girls in program schools.

The estimated program impact on teacher school attendance in Busia was 6.5 percentage points (standard error 2.7 percentage points, statistically significant at 99% confidence, Table 7, Panel B), reducing overall teacher absenteeism by approximately one-third.[24] Note that the mean school baseline 2000 test score is positively but only weakly correlated with teacher attendance

[24] These results are for all teachers in the schools. In our data, it is difficult to distinguish between teacher attendance in grade 6 versus other grades, since the same teacher often teaches a subject (i.e., mathematics) in several different grades, and the data were recorded on a teacher by teacher basis rather than by grade and subject. Thus, it remains possible that average teacher attendance gains would be even larger for grade 6 classes alone.

(point estimate 0.017, standard error 0.015), and results are robust to excluding this term. Estimated program impacts in Busia are not statistically significantly different by teacher's gender or experience (regressions not shown). Program impacts on teacher attendance are near zero and not significant in Teso (regression not shown).

Although student school participation and teacher attendance – two easily observed dimensions of effort – improved in Busia district program schools, there is no evidence that study habits changed significantly in any other dimension we measured in the 2002 student survey. When that survey was administered, cohort 2 girls were competing for the award (cohort 1 girls had already competed), so in what follows we focus on cohort 2.

Program school students in Busia were no more likely than comparison school students to seek out extra tutoring, use a textbook at home during the past week, hand in homework, or do chores at home, and this holds for both girls and boys (Table 8, Panel A). In the case of chores, the estimated zero impact indicates the program did not lead to lost home production, suggesting any increased study effort may have come out of children's leisure time or intensified effort during school hours. Program impacts on classroom inputs, including the number of desks and flipcharts (using data gathered during 2002 classroom observations), are similarly near zero and not statistically significant (regressions not shown).

Program school students were also no more likely than comparison students to report being called on by a teacher in class during the last two days (Table 8, Panel A), and there is no statistically significant difference in how often girls are called on in class relative to boys (regression not shown), indicating that teachers did not divert classroom effort towards girls in program schools. This finding, together with that of increased teacher attendance, provides a plausible explanation for the positive spillovers for Busia boys, namely, greater teacher effort directed to the class as a whole.

There is no statistically significant program impact on the number of textbooks children have at home, the number of new books (the sum of textbooks and exercise books) their household

recently purchased for them (Table 8, Panel B), although there the point estimates for girls are both positive and large, and in the case of textbooks at home, marginally statistically significant (0.27 additional textbooks, standard error 0.17), providing some evidence of increased parental investments in girls' education.[25]

There is no evidence in Busia program schools of the adverse student attitude changes emphasized by some psychologists. We attempted to directly measure "intrinsic motivation" toward education using eight questions where students were asked to compare how much they liked a school activity, for instance, doing homework, compared to a non-school activity, such as playing sports or fetching water. Overall, students report preferring the school activity 72% of the time. There are no statistically significant differences in this index across the program and comparison schools for either girls or boys (Table 8, Panel C). This is evidence against the view that external incentives dampened intrinsic motivation. Similarly, program and comparison school girls and boys are equally likely to think of themselves as a "good student", to think "being a good student means working hard", or to think they can be in the top three students in their class, based on their survey responses.

6. Program Cost-effectiveness

We compare the cost-effectiveness of six programs that have recently been conducted in the study area: the girls' merit scholarship program that is the focus of this paper, the teacher incentive program (Glewwe et al. 2003), a textbook provision program (Glewwe et al. 1997), flip chart program (Glewwe et al. 2004), a deworming program (Miguel and Kremer 2004), and a child sponsorship program that provided a range of inputs, including free uniforms (Kremer et al. 2003). We conclude that providing merit scholarship incentives for students is a highly cost-effective way to

[25] There is a significant increase in textbook use among program girls in cohort 1 in 2002: girls in program schools report having used textbooks at home 6 percentage points (significant at 95% confidence) more than cohort 1 girls in the comparison schools, further suggestive evidence of greater parental investment. However, there are no such gains among cohort 2 students competing for the award in 2002, as shown in Table 8.

improve test scores, and a reasonably cost-effective way to boost school participation. This is true even if one only values benefits for girls with low baseline test scores

The average test score gain in girls' scholarship program schools, among both girls and boys in Busia and Teso districts in both years of the program, is roughly 0.12 standard deviations.[26] The comparable gains for teacher incentive program schools over two years was smaller, at 0.07 standard deviations, and for textbook program schools the average gain was only 0.04 standard deviations. The test gains in the teacher incentive program were concentrated in the year of the competition, and then fell in subsequent years. The deworming, flip chart, and child sponsorship programs did not produce statistically significant test score impacts. Since the cost per test score gain in these three programs is infinite given the zero estimated gain, we do not focus on them below.

One important issue in evaluating cost-effectiveness is whether to treat all payments under the program as social costs or whether to consider some as transfers. We first report "education budget cost effectiveness" (Table 9, column 4) which shows the test score gain per pupil divided by program costs per pupil. This is the relevant calculation for an education policymaker seeking to maximize test gains with a given budget. From the standpoint of a social planner, however, payments to families in the scholarship program, and to teachers in the teacher incentive program, could be considered transfers. If seen as pure transfers, the social cost is simply the deadweight loss involved in raising funds. In calculating "social cost effectiveness" (column 5) we follow a rule of thumb often used in developed countries and treat the marginal cost of raising one dollar as 1.4 dollars (Ballard et al. 1985). To make the education budget and social cost effectiveness figures comparable, we also multiply costs in the education budget calculations by 1.4 to reflect likely tax distortions.

[26] Estimates of the overall gain in Busia and Teso districts include 0.12 standard deviations (Table 4, regression 1), 0.13 standard deviations (Table 4, regression 2), 0.10 standard deviations (Table 5, Panel A), and 0.12 standard deviations (Table 5, Panel B). We use the Table 5, Panel B estimates for Busia and Teso overall, and for Busia alone, in these calculations, although results are nearly identical using the alternative estimates.

It is worth noting that the effective transfer is the net benefit to recipients after making allowances for any disutility of their increased effort. Assuming that students and teachers are rational, the total additional effort exerted should not be greater than the value of the award. Thus the education budget cost effectiveness calculation yields an upper bound on the true social cost of the program (Table 9, column 4), and a lower bound is generated by treating the entire payment as a transfer (as in column 5).

The following calculations use project cost data from the NGO records and exclude research costs. The per pupil cost per 0.1 standard deviation average test score gain under the social cost effectiveness calculation is US$1.41 for the girls scholarship program, and similar at US$1.36 per 0.1 standard deviation average gain for the teacher incentive program, but costs are much higher for the textbook program at $5.61 (Table 9, column 5). In Busia district, where the girls' scholarship program was well-received by residents, the social cost per 0.1 s.d. gain per pupil falls to US$0.71, making student merit awards a much more cost-effective way to boost student scores there than the other programs. Student merit awards and teacher incentives are also cost effective relative to textbook provision, flipcharts, deworming, and the child sponsorship program under the education budget calculation (Table 9, column 4).

The education budget approach also provides a valid measure of cost-effectiveness from the perspective of a social planner who values only the welfare gains for girls with low baseline 2000 test scores, since these girls have little chance of winning an award (Figure 4) and thus expected transfers to them are essentially zero. Low performing girls' parents tend to have less education than average, as discussed above. For this relatively disadvantaged group, the merit award program is a particularly cost-effective way to boost test scores relative to textbook provision, since textbooks only raised scores for students in the top quartile and not elsewhere (Glewwe et al 1997), while merit awards lead to test score gains throughout the baseline distribution (Figure 9).

The estimates for both the girls scholarship program and the teacher incentive program do not include costs associated with administering academic exams in the schools, which are substantial. Including testing costs, the social cost per 0.1 s.d. average test score gain nearly doubles for the girls scholarship schools, from US$1.41 to US$2.78, and more than doubles from US$1.36 to US$3.70 for the teacher incentive program, although these programs still remain far more cost-effective than textbook provision. Once again restricting attention to Busia alone, the per pupil social cost,including testing costs per 0.1 s.d. average test score gain is only US$1.53. Many countries, like Kenya during the study period, already carry out regular standardized testing in primary schools, in which case the additional exam costs are unnecessary and the previous lower estimates are relevant.

Although test score cost effectiveness figures are similar for the girls scholarship and teacher incentive programs, the scholarship program is more attractive in many other dimensions. First, the teacher incentive program did not produce lasting test score impacts, and there is evidence of "teaching to the test" rather than effort directed at human capital acquisition. As a result, the long-term impacts of the teacher incentive program are likely to be smaller than girls scholarships. The girls scholarship program also generated large impacts on school participation, and any future benefits of higher school participation are not considered in the above cost calculations. If scholarship winners have high returns to additional education, then to the extent that winners obtain more education than they would have otherwise, this yields additional benefits. Finally, the distributional impact of the girls scholarship program is likely to be more desirable since it provides rewards to pupils and their families instead of to teachers, who tend to be better-off in rural Kenya.

School participation is a second educational outcome measure important to policymakers. Deworming provision is far more cost effective in this dimension than the other interventions, including merit awards, at an average cost of only US$3.50 per additional year of school participation (Miguel and Kremer 2004). There are no significant school participation gains from teacher incentives, textbook provision, flipcharts, or for merit awards for Busia and Teso districts

overall, so the cost per participation gain is infinite there. However, for Busia district alone the cost per additional year of school participation (using estimates in Table 7 Panel A) is US$4.24 / 0.047 = US$90, making merit awards the second most cost-effective of these six programs. The cost per additional year of school participation for the child sponsorship program was US$99, making it slightly less cost-effective than merit scholarships.[27]

7. Conclusion

Merit-based scholarships have historically been an important part of the educational system in many countries. The evidence we present suggests that such programs can induce students and teachers to exert additional effort, raising test scores not only for the eventual recipients of the scholarships but also for others. A merit scholarship for Kenyan adolescent girls had a large positive effect on test scores in both years of the scholarship competition. There are large, significant and robust average program effects on girls' test scores in Busia district, on the order of 0.2-0.3 standard deviations, but we do not find significant effects in Teso district. Our inability to find these effects may be due to differential sample attrition across Teso program and comparison schools that complicates the econometric analysis, or it may partially reflect the lower value placed on winning the merit award there, especially in the aftermath of the tragic 2001 lightning strike. It remains unclear whether the problems encountered in Teso district would have arisen there in the absence of the lightning tragedy or would arise in other settings.

Initially low-achieving girl students, and boys (who were ineligible for the award), both show considerable test score gains and school participation gains in Busia district, providing evidence of positive program spillovers. Our data are consistent with strategic complementarity between the effort levels of girls eligible for the award, the effort of teachers, and of other students. Such strategic

[27] Another randomized evaluation conducted in this area provided meals to preschool aged children (Vermeersch and Kremer, 2005) and led to participation gains at a cost of only US$36 per additional year of participation (Kremer 2003) – although note that the sample is considerably younger in that case than the current study.

complemenatarity could potentially generate multiple equilibria in the classroom learning environment, or culture. Educators often stress the importance of classroom culture, and Akerlof and Kranton (2003) have recently attempted to formally model these cultures. Most studies find that conventional educational variables – including the pupil-teacher ratio and expenditures on inputs like textbooks – explain only a modest fraction of variation in test score performance, typically with R^2 values on the order of 0.2-0.3 (Summers and Wolfe 1977, Hanushek 2003). While there are many potential interpretations, one possibility is that unobserved classroom culture is driving much of this unexplained variation. In the current study, the divergence in program impacts between Busia and Teso districts, two areas with distinct local ethnic compositions and traditions, is also consistent with multiple equilibria in classroom culture.

The results also suggest that merit scholarships may not only be useful as a way of helping talented students continue their education, but could also help increase average student effort in primary school. A key reservation about merit awards for educators has been the possibility of adverse equity impacts. It is likely that advantaged students gained most from the program we study: we find that scholarship winners come from more educated households, and that the tendency for girls from more educated households to score in the top 15% on tests is somewhat stronger in program schools. However, groups with little chance at winning an award, including girls with low baseline test scores, gained enough from the merit scholarship program to make it cost effective for them, even neglecting the benefits to higher scoring students.

One way to spread the benefits of merit scholarships even more widely would be to restrict the scholarship competition to poor pupils, schools or regions, or alternatively to conduct multiple competitions, each restricted to a small geographic area. For instance, if each Kenyan location – a small administrative unit – awarded merit scholarships to its residents independently of other locations, children would only compete against others who live in the same area, where many households live in comparable socioeconomic conditions. To the extent that such a policy would put

more students near the margin of winning a scholarship, it would presumably generate greater incentive effects and spillover benefits.

There are interesting possibilities for future research related to this project. We have collected detailed contact information for sample pupils and plan to conduct follow-up surveys as they enter adulthood to estimate long-run impacts of increased learning on labor market performance and other life outcomes. We also plan to estimate impacts of winning a scholarship on later outcomes, exploiting the discontinuity created by the sharp winning test score threshold.

References

Acemoglu, Daron, and Joshua Angrist. (2000). "How Large are Human Capital Externalities? Evidence from Compulsory Schooling Laws", *NBER Macroeconomics Annual*, 9-59.

Akerlof, George, and Rachel Kranton. (2003). "Identity and Schooling: Some Lessons for the Economics of Education," *Journal of Economic Literature*, 40, 1167-1201.

Angrist, J. and V. Lavy (2002). "The Effect of High School Matriculation Awards: Evidence from Randomized Trials." *NBER Working Paper #9389*.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. (2002). "Vouchers for Private Schooling in Colombia: Evidence from Randomized Natural Experiments", *American Economic Review*, 1535-1558.

Ashworth, K., J. Hardman, et al. (2001). "Education Maintenance Allowance: The First Year, A Qualitative Evaluation". Research Report RR257, Department for Education and Employment.

Ballard, Charles L., John B. Shoven, and John Whalley. (1985). "General Equilibrium Computations of the Marginal Welfare Cost of Taxes in the United States," *American Economic Review*, 75(1), 128-138.

Benabou, R., and J. Tirole (2004). "Intrinsic and Extrinsic Motivation". *Review of Economic Studies*, 70, 489-520.

Binder, M., P. T. Ganderton, et al. (2002). "Incentive Effects of New Mexico's Merit-Based State Scholarship Program: Who Responds and How?", unpublished manuscript.

Cameron, J., K. M. Banko, et al. (2001). "Pervasive Negative Effects of Rewards on Intrinsic Motivation: The Myth Continues." *The Behavior Analyst* 24: 1-44.

Central Bureau of Statistics. (1999). *Kenya Demographic and Health Survey 1998*, Republic of Kenya, Nairobi, Kenya.

College Board. (2002). *Trends in Student Aid*, Washington, D.C.

Cornwell, C., D. Mustard, et al. (2002). "The Enrollment Effects of Merit-Based Financial Aid: Evidence from Georgia's HOPE Scholarship." *Journal of Labor Economics*.

Cornwell, Christopher M., Kyung Hee Lee, and David B. Mustard. (2003). "The Effects of Merit-Based Financial Aid on Course Enrollment, Withdrawal and Completion in College", unpublished working paper.

Croxford, L., C. Howieson, et al. (2002). "Education Maintenance Allowances (EMA) Evaluation of the East Ayshire Pilot." Research Findings No. 6, Enterprise and Lifelong Learnings Report, Glasgow.

Deci, E. L. (1971). "Effects of Externally Mediated Rewards on Intrinsic Motivation." *Journal of Personality and Social Psychology* 18: 105-115.

Deci, E. L., R. Koestner, et al. (1999). "A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation." *Psychological Bulletin* 125(627-668).

Dynarski, S. (2003). "The Consequences of Merit Aid." *NBER Working Paper #9400*.

Fan, J. (1992). "Design-adaptive Nonparametric Regression." *Journal of the American Statistical Association*, 87, 998-1004.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. (1997). "Textbooks and Test scores: Evidence from a Prospective Evaluation in Kenya", unpublished working paper.

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

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. (2004). "Retrospective v. Prospective Analysis of School Inputs: The Case of Flip Charts in Kenya." forthcoming, *Journal of Development Economics*.

Government of Kenya, Ministry of Planning and National Development. (1986). *Kenya Socio-cultural Profiles: Busia District*, (ed.) Gideon Were. Nairobi.

Hanushek, Erik. (2003). "The Failure of Input-based Schooling Policies", *Economic Journal*, 113, 64-98.

Jacob, Brian, and Steven Levitt. (2002). "Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating", *NBER Working Paper #9413.*

Kremer, Michael. (2003). "Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons", *American Economic Review: Papers and Proceedings,* 93 (2), 102-106.

Kremer, Michael, Sylvie Moulin, and Robert Namunyu. (2003). "Decentralization: A Cautionary Tale", unpublished working paper, Harvard University.

Kruglanski, A., I. Friedman, et al. (1971). "The Effect of Extrinsic Incentives on Some Qualitative Aspects of Task Performance." *Journal of Personality and Social Psychology* 39: 608-617.

Lazear, E.P. (2001). "Educational Production", *Quarterly Journal of Economics*, 116(3), 777-804.

Lee, D. S. (2002). "Trimming the Bounds on Treatment Effects with Missing Outcomes." *NBER Working Paper #T277*.

Lepper, M., D. Greene, et al. (1973). "Undermining Children's Interest with Extrinsic Rewards: A Test of the 'Overidentification Hypothesis." *Journal of Personality and Social Psychology* 28: 129-137.

Leuven, Edwin, Hessel Oosterbeek, Bas van der Klaauw. (2003). "The Effect of Financial Rewards on Students' Achievement: Evidence from a Randomized Experiment", unpublished working paper, University of Amsterdam.

Lucas, Robert E. (1988). "On the Mechanics of Economic Development", *Journal of Monetary Economics*, 22, 3-42.

Miguel, Edward, and Michael Kremer. (2004). "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities", *Econometrica*, 72(1), 159-217.

Moretti, Enrico. (2004). "Workers' Education, Spillovers and Productivity: Evidence from Plant-level Production Functions", *American Economic Review*, 94(3).

Nurmi, J. (1991). "How Do Adolescents See Their Future? A review of the development of future orientation and planning", *Developmental Review*, 11:1-59.

Orfield, Gary. (2002). "Foreward", in Donald E. Heller and Patricia Marin (eds.), *Who Should We Help? The Negative Social Consequences of Merit Aid Scholarships* (Papers presented at the conference "State Merit Aid Programs: College Access and Equity" at Harvard University). Available online at: http://www.civilrightsproject.harvard.edu/research/meritaid/merit_aid02.php.

Skinner, B. F. (1961). "Teaching Machines." *Scientific America* November: 91-102.

Summers, Anita A., and Barbara L. Wolfe. (1977). "Do Schools Make a Difference?" *American Economic Review*, 67(4), 639-652.

United Nations. (2003). *The Right to Education*, Economic and Social Council Special Rapporteur Katarina Tomasevski. Available online at: (http://www.right-to-education.org/content/unreports/unreport12prt1.html#tabel1).

World Bank. (2002). *World Development Indicators* (www.worldbank.org/data).

World Bank. (2004). *Strengthening the Foundation of Education and Training in Kenya: Opportunities and Challenges in Primary and General Secondary Education*. Nairobi.

Figure 1: Map of Busia District and Teso District, Kenya,
with location of Girls Scholarship Program Schools (legend below)

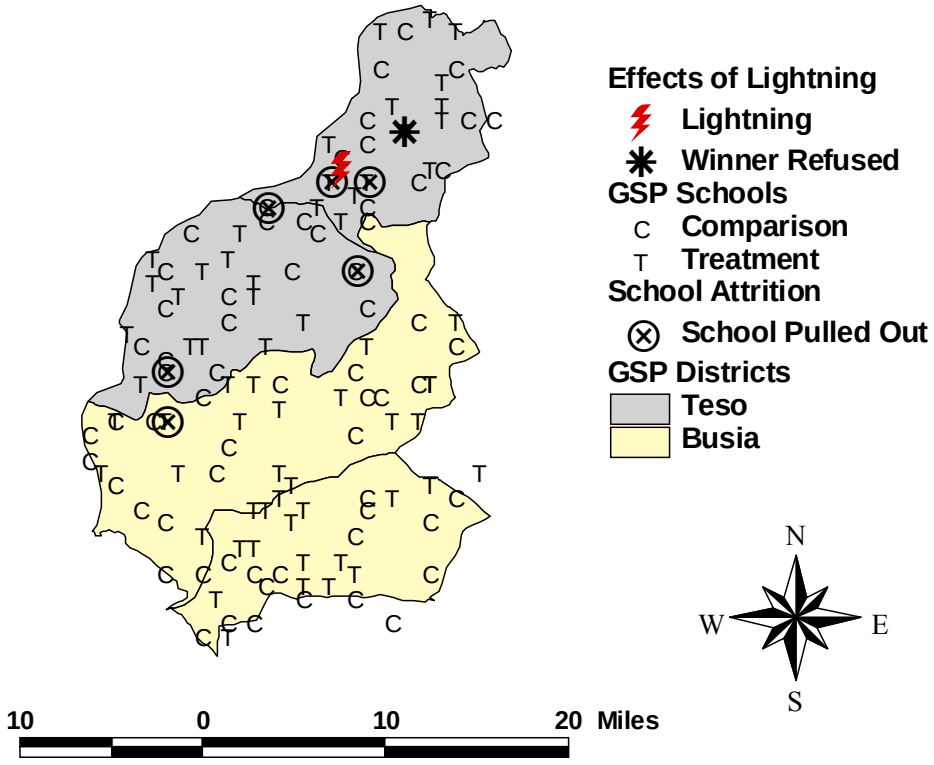

Figure 2: Proportion of Baseline Students in the 2001 Restricted Sample by Baseline (2000) Test Score
Cohort 1 Busia Girls (Panel A) and Busia Boys (Panel B)
(Non-parametric Fan locally weighted regressions)

Panel (A) **Panel (B)**

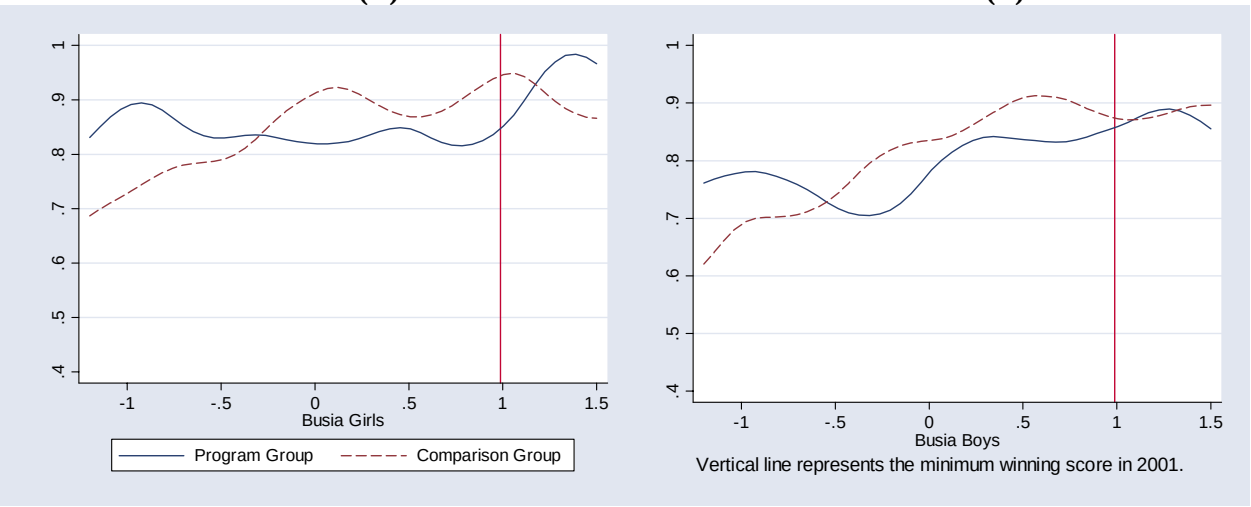

Figure 3: Proportion of Baseline Students in the 2001 Restricted Sample by Baseline (2000) Test Score
Cohort 1 Teso Girls (Panel A) and Teso Boys (Panel B)
(Non-parametric Fan locally weighted regressions)

Panel (A) **Panel (B)**

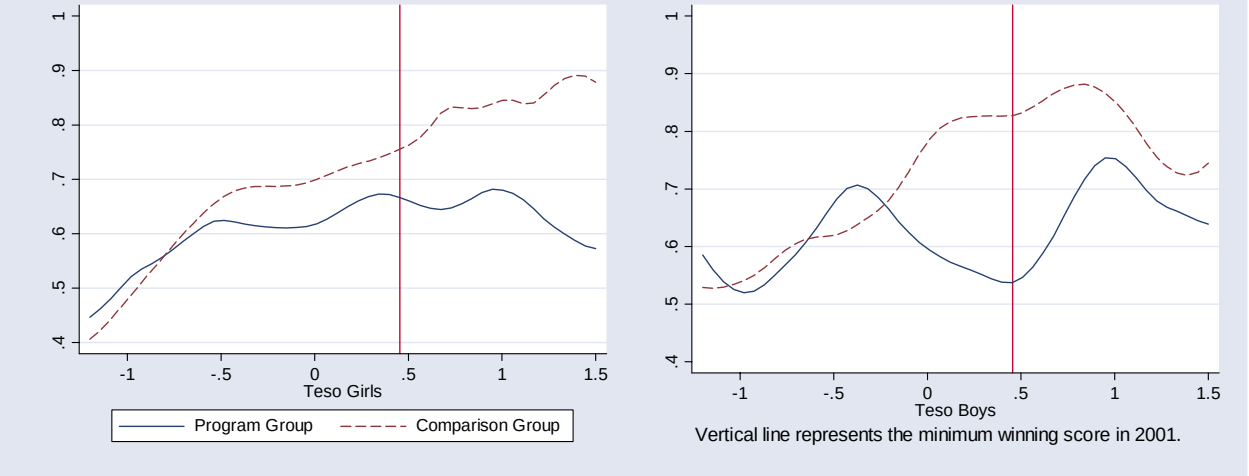

Figure 4: Proportion of Baseline Students Winning the Award in 2001 by Baseline (2000) Test Score
Cohort 1 Busia Program School Girls (Panel A) and Teso Program School Girls (Panel B)
(Non-parametric Fan locally weighted regressions)

Panel (A) **Panel (B)**

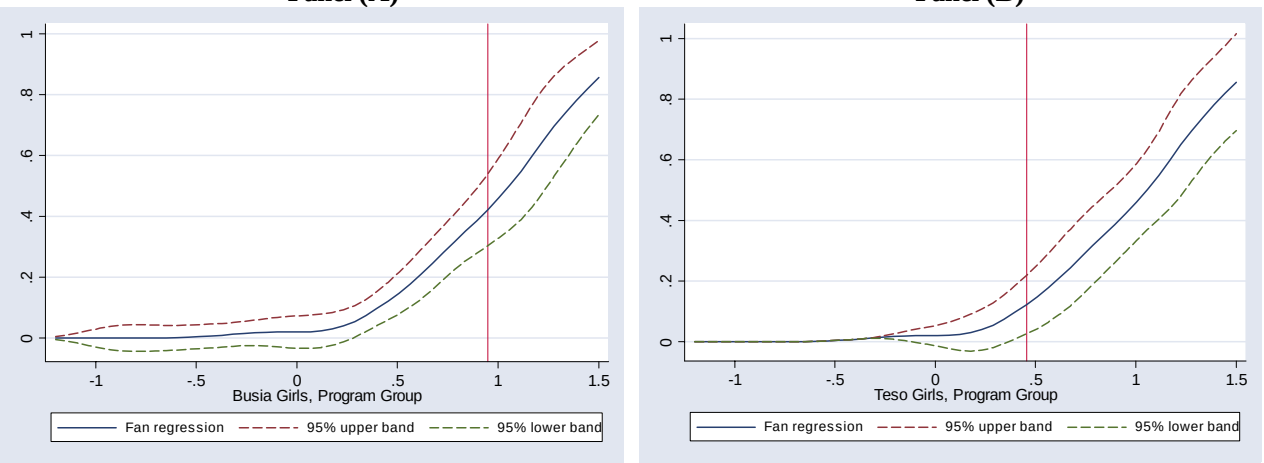

Figure 5: Baseline (2000) Test Score Distribution
Cohort 1 Busia Girls (Panel A) and Busia Boys (Panel B)
(Non-parametric kernel densities)

Panel (A) **Panel (B)**

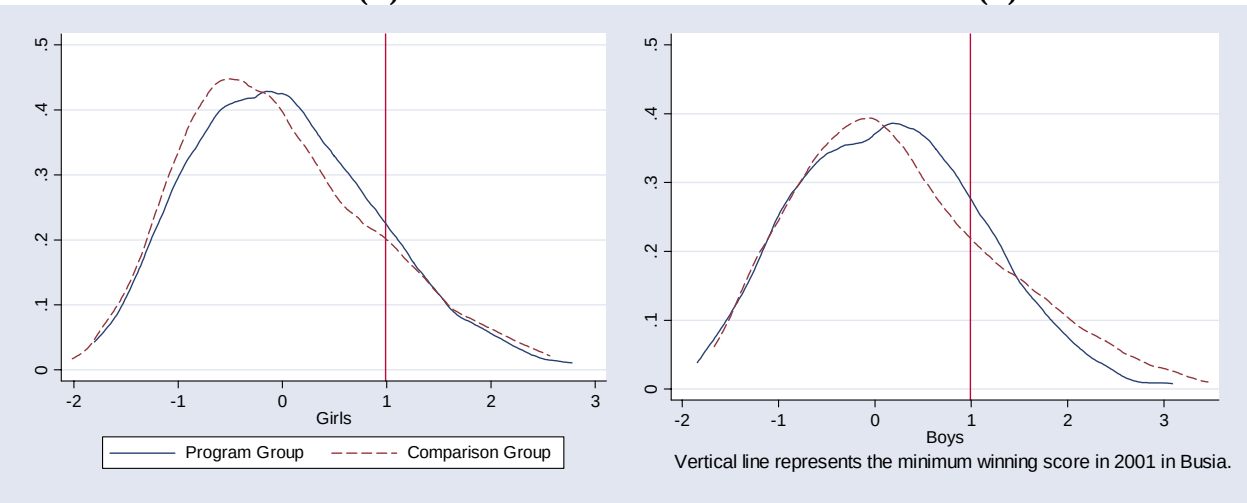

Figure 6: Year 1 (2001) Test Score Distribution
Cohort 1 Busia Girls (Panel A) and Busia Boys (Panel B)
(Non-parametric kernel densities)

Panel (A) **Panel (B)**

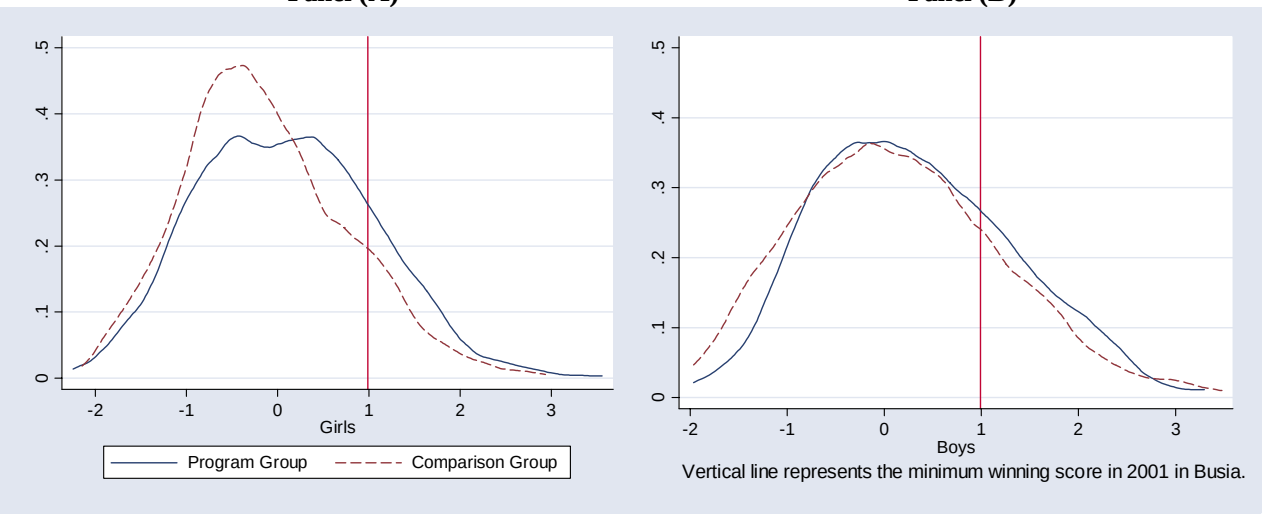

Figure 7: Year 2 (2002) Test Score Distribution
Cohort 1 Busia Girls (Panel A) and Busia Boys (Panel B)
(Non-parametric kernel densities)

Panel (A) **Panel (B)**

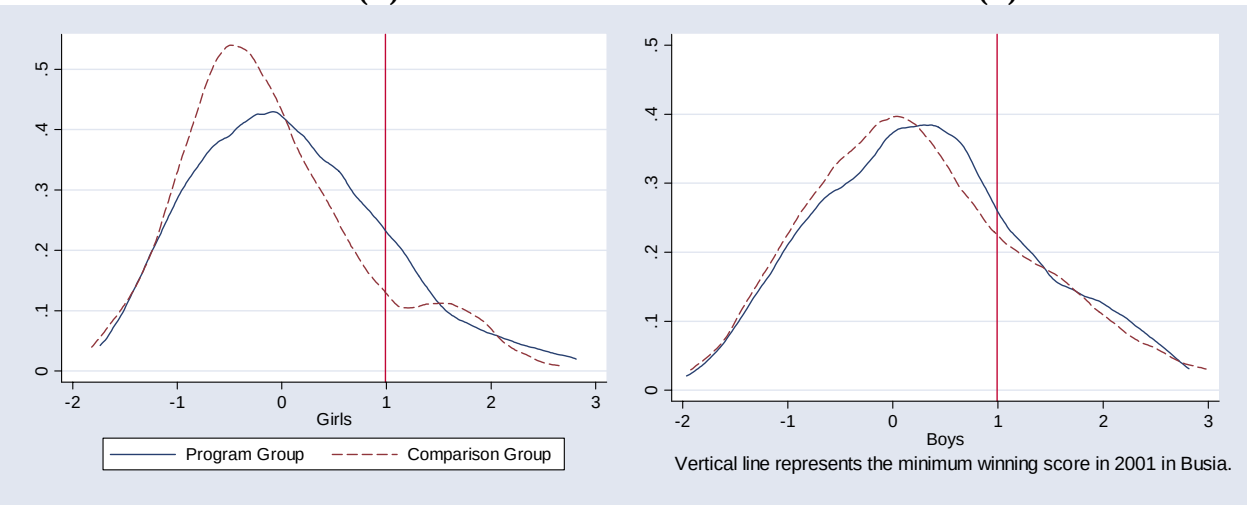

Figure 8: Year 2 (2002) Test Score Distribution
Cohort 2 Busia Girls (Panel A) and Busia Boys (Panel B)
(Non-parametric kernel densities)

Panel (A) **Panel (B)**

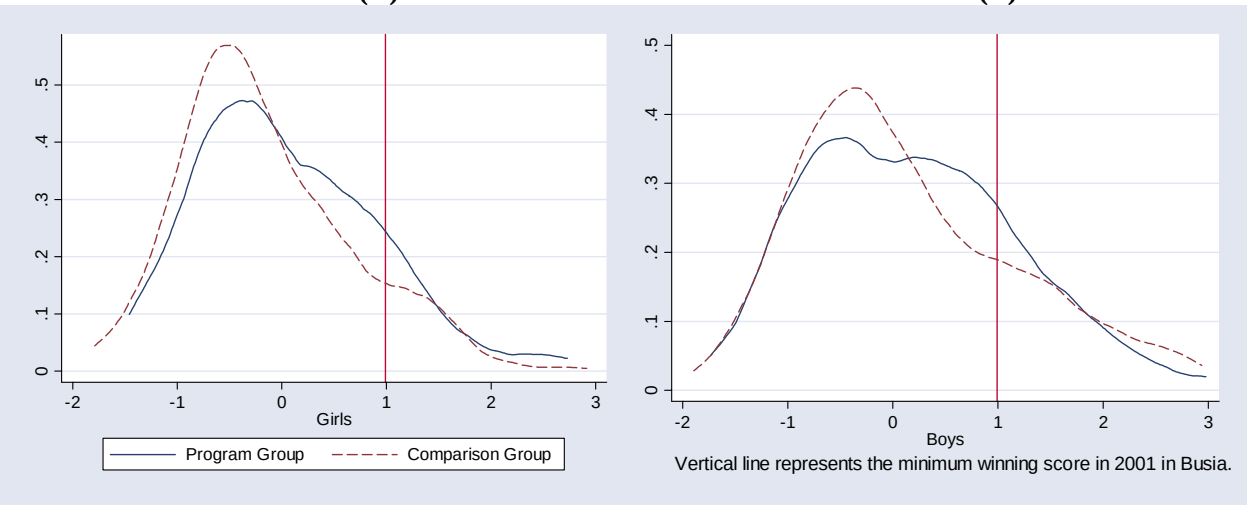

Figure 9: Year 1 (2001) Test Score Impacts by Baseline (2000) Test Score
Difference between Program Schools and Comparison Schools
Cohort 1 Busia Girls (Panel A) and Busia Boys (Panel B)
(Non-parametric Fan locally weighted regression)

Panel (A) **Panel (B)**

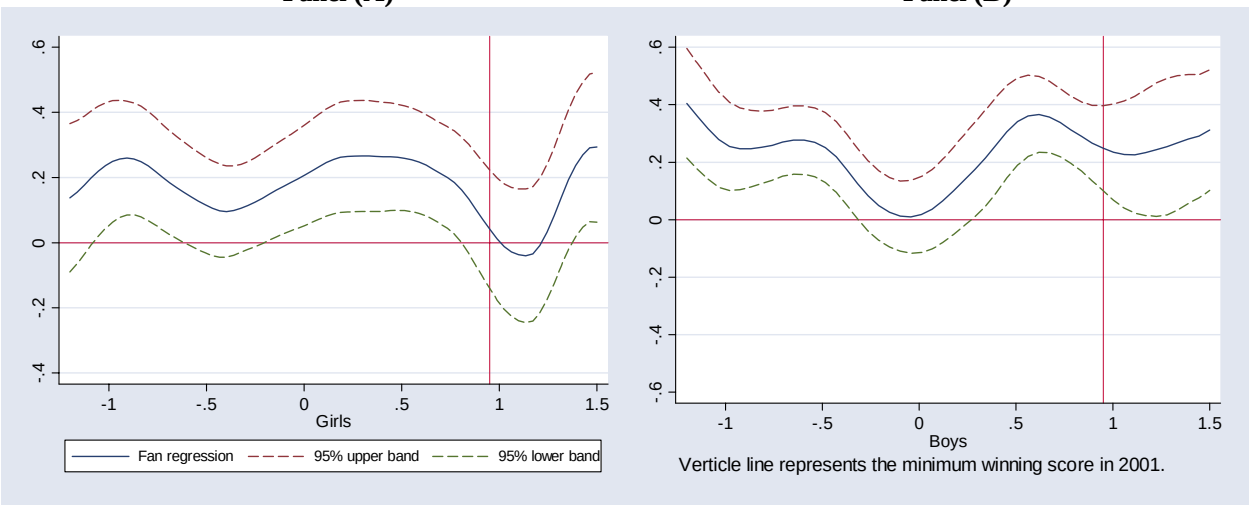

Table 1: Summary Statistics

Panel A: School characteristics	Obs.					
Number of Schools: Program	63					
Number of Schools: Comparison	64					
Number of Schools: Busia	69					
Number of Schools: Teso	58					
Panel B: Baseline sample	--------Cohort 1--------			--------Cohort 2--------		
	Obs.	Mean	Std dev	Obs.	Mean	Std dev
Number of students: Program	2720			3254		
Number of students: Comparison	2638			3116		
Number of students: Busia district	3159			3756		
Number of students: Teso district	2198			2614		
Gender (1=Male)	5358	0.51	0.50	6370	0.52	0.50
Age in 2001	4937	14.3	1.6	5895	13.3	1.6
Test Score 2000	3216	0.06	0.99	-	-	-
Test Score 2001	4040	0.09	0.99	-	-	-
Test Score 2002	3404	0.05	1.01	3620	0.04	1.01
Mean School Test Score 2000	4932	0.12	0.65	5847	0.13	0.65
School Participation 2001	4798	0.79	0.41	5761	0.77	0.42
School Participation 2002	4686	0.77	0.33	5625	0.77	0.32
Panel C: Restricted sample	--------Cohort 1--------			--------Cohort 2--------		
	Obs.	Mean	Std dev	Obs.	Mean	Std dev
Number of students: Program	1827			1783		
Number of students: Comparison	1921			1727		
Number of students: Busia district	2440			1877		
Number of students: Teso district	1308			1633		
Gender (1=Male)	3748	0.51	0.50	3510	0.55	0.50
Age in 2001	3721	14.2	1.5	3498	13.1	1.5
Test Score 2000	2430	0.13	0.97	-	-	-
Test Score 2001	3748	0.09	0.99	-	-	-
Test Score 2002	2810	0.11	1.01	3510	0.05	1.01
Mean School Test Score 2000	3748	0.14	0.64	3510	0.15	0.66
School Participation 2001	3597	0.86	0.35	3384	0.84	0.37
School Participation 2002	3550	0.83	0.27	3503	0.87	0.21

Notes: These statistics are for girls and boys in the sample. A dash (-) indicate that the data are unavailable (for instance, 2000 and 2001 exams for Cohort 2). School participation in 2001 is from a one-time unannounced visit to schools in term 3, 2001. School participation in 2002 is consists of three unannounced visits to schools throughout the school year.

The Baseline sample refers to all students that were registered in grade 6 (cohort 1) or grade 5 (cohort 2) in January 2001. The Restricted sample consists of students who were in the Baseline Sample, in schools that did not pull out of the program, for whom we have mean school test scores in 2000, and who took either the 2001 or 2002 test. The Longitudinal sample contains those cohort 1 Restricted sample students who took the 2000 test. The mean school test score in 2000, used in the analysis, is for those students in the cohort 1 Longitudinal sample.

Table 2: Proportion of Students with 2001 and 2002 Test Scores, Cohorts 1 and 2

Panel A: Cohort 1, with 2001 test (in Restricted sample)

	--------Busia district--------			--------Teso district--------		
	Program	Comparison	Difference (s.e.)	Program	Comparison	Difference (s.e.)
Girls	0.79	0.78	0.01 (0.04)	0.53	0.65	-0.12 (0.09)
Boys	0.76	0.77	-0.01 (0.06)	0.54	0.66	-0.12 (0.09)

Panel B: Cohort 2, with 2002 test (in Restricted sample)

	--------Busia district--------			--------Teso district--------		
	Program	Comparison	Difference (s.e.)	Program	Comparison	Difference (s.e.)
Girls	0.50	0.48	0.02 (0.04)	0.57	0.58	-0.02 (0.09)
Boys	0.50	0.52	-0.02 (0.04)	0.65	0.69	-0.04 (0.08)

Notes: Standard errors in parenthesis. Significantly different than zero at 90% (*), 95% (**), 99% (***) confidence. The denominator for these proportions consists of all grade 6 (cohort 1) or grade 5 (cohort 2) students who were registered in school in January 2001, in schools that did not pull out of the program, and for whom we have mean school test scores for 2000. The relatively low rates of missing data for Teso district students in 2002 is likely the result of the use of ICS exam scores (administered in early 2003), rather than district exam scores; the 2002 Teso district exams were cancelled due to the upcoming Kenyan national elections (as described in Section 2). Cohort 2 data for Busia district students in 2002 is based on the 2002 Busia district exams, which were administered as scheduled in late 2002, and for which students must pay a small fee (unlike the ICS exams, where were free, possibly explaining the lower attrition rate in Teso district in 2002 than in 2001).

Table 3: Demographic and Socio-Economic Characteristics Across Program and Comparison schools Cohort 1 and Cohort 2 Busia Girls and Busia Boys

	----------Girls----------			----------Boys----------		
	Program	Comparison	Difference (s.e.)	Program	Comparison	Difference (s.e.)
Age in 2001	13.5	13.4	0.0 (0.1)	13.9	13.7	0.2 (0.2)
Father's education (years)	5.2	5.2	0.2 (0.5)	4.9	4.9	0.00 (0.5)
Mother's education (years)	4.6	4.6	0.1 (0.4)	4.0	4.2	-0.2 (0.4)
Total children in household	7.0	6.5	0.5 (0.5)	6.3	6.2	0.1 (0.5)
Proportion ethnic Luhya	0.49	0.47	0.03 (0.05)	0.48	0.44	0.03 (0.05)
Latrine ownership	0.96	0.94	0.02 (0.01)	0.95	0.93	0.02 (0.02)
Iron roof ownership	0.77	0.77	0.00 (0.03)	0.72	0.75	-0.02 (0.03)
Mosquito net ownership	0.33	0.33	0.00 (0.03)	0.27	0.26	0.01 (0.04)
Test Scores 2000– Baseline sample (cohort 1 only)	-0.05	-0.12	0.07 (0.18)	0.04	0.10	-0.07 (0.19)
Test Scores 2000 – Restricted sample (cohort 1 only)	0.07	0.03	0.04 (0.19)	0.15	0.28	-0.13 (0.19)

Notes: Standard errors in parenthesis. Significantly different than zero at 90% (*), 95% (**), 99% (***) confidence. Sample includes baseline students in cohort 1 and cohort 2 in 2001 in program and comparison schools in Busia district. Data is from 2002 Student Questionnaire, and from Busia District Education Office records. The sample size is 4,504 questionnaires, 65% of the baseline sample in Busia (the remaining students either had left school by the time of the 2002 survey or were not present in school on the day of the survey).

Table 4: Program Impact on Test Scores
Longitudinal Sample, Cohort 1 Girls and Boys

	Dependent variable: Normalized test scores from 2001 and 2002				
	Busia and Teso districts			Busia district	Teso district
	(1)	(2)	(3)	(4)	(5)
Program school	0.12	0.13**	0.12*	0.19**	-0.02
	(0.13)	(0.06)	(0.07)	(0.08)	(0.09)
Male * Program School			0.01	0.01	0.01
			(0.05)	(0.05)	(0.09)
Male			0.16***	0.09**	0.28***
			(0.04)	(0.04)	(0.07)
Individual test score, 2000		0.80***	0.79***	0.85***	0.69***
		(0.02)	(0.02)	(0.03)	(0.02)
Sample Size	4294	4294	4294	2858	1436
R^2	0.00	0.61	0.61	0.67	0.53
Mean of dependent variable	0.13	0.13	0.13	0.13	0.12

Notes: Significantly different than zero at 90% (*), 95% (**), 99% (***) confidence. OLS regressions, Huber robust standard errors in parenthesis. Disturbance terms are allowed to be correlated across observations in the same school, but not across schools. Test scores were normalized such that comparison group test scores had mean zero and standard deviation one. The Longitudinal sample includes cohort 1 students who were registered in grade 6 in January 2001, in schools that did not pull out of the program, for whom we have individual test score data in 2000, and who took the 2001 test.

Table 5: Program Impact on Test Scores
Cohorts 1 and 2 Girls and Boys

Panel A: Restricted Sample	Dependent variable: Normalized test scores from 2001 and 2002				
	Busia and Teso districts			Busia district	Teso district
	(1)	(2)	(3)	(4)	(5)
Program school	0.10	0.10*	0.14**	0.25***	-0.02
	(0.13)	(0.05)	(0.06)	(0.07)	(0.08)
Male * Program School			-0.07	-0.12*	0.02
			(0.05)	(0.07)	(0.07)
Male			0.31***	0.30***	0.32***
			(0.04)	(0.06)	(0.05)
Mean school test score, 2000		0.77***	0.77***	0.85***	0.66***
		(0.05)	(0.05)	(0.05)	(0.06)
Sample Size	10068	10068	10068	6123	3945
R^2	0.00	0.25	0.27	0.33	0.19
Mean of dependent variable	0.08	0.08	0.08	0.10	0.05

Panel B: Baseline Sample	Dependent variable: Normalized test scores from 2001 and 2002			
	Busia and Teso districts		Busia district	Teso district
	(1)	(2)	(3)	(4)
Program school	0.12	0.19	0.27	0.06
	(0.12)	(0.12)	(0.17)	(0.15)
Male * Program School		-0.12	-0.17*	-0.06
		(0.07)	(0.10)	(0.07)
Male		0.33***	0.33***	0.32***
		(0.05)	(0.08)	(0.05)
Sample Size	11064	11064	6486	4578
R^2	0.00	0.02	0.02	0.02
Mean of dependent variable	0.06	0.08	0.09	0.02

Notes: Significantly different than zero at 90% (*), 95% (**), 99% (***) confidence. OLS regressions, Huber robust standard errors in parenthesis. Disturbance terms are allowed to be correlated across observations in the same school, but not across schools. Test scores were normalized such that comparison group test scores had mean zero and standard deviation one. Indicator variables are included in all specifications for Cohort 1 in 2001, Cohort 1 in 2002, and Cohort 2 in 2002 (coefficient estimates not shown).

The Restricted sample (Panel A) includes students who were registered in grade 6 (cohort 1) or grade 5 (cohort 2) in January 2001, in schools that did not pull out of the program, for whom we have mean school test score data in 2000, and who took the 2001 or 2002 test. The Full sample (Panel B) includes students who were registered in grade 6 (cohort 1) or grade 5 (cohort 2) in January 2001, and who took the 2001 or 2002 test.

Table 6: Program Impact on Test Scores
Restricted Sample, Cohorts 1 and 2 Girls and Boys, Busia District

	Dependent variable: Normalized test scores from 2001 and 2002	
	Busia Girls	Busia Boys
	(1)	(2)
Program impact, Cohort 1 (in 2001)	0.28***	0.18**
	(0.10)	(0.09)
Program impact, Cohort 2 (in 2002)	0.21**	0.11
	(0.10)	(0.13)
Post-competition impact, Cohort 1 (in 2002)	0.25***	0.07
	(0.09)	(0.09)
Mean school test score, 2000	0.83***	0.87***
	(0.05)	(0.06)
Sample Size	2917	3206
R^2	0.36	0.32
Mean of dependent variable	-0.03	0.21

Notes: Significantly different than zero at 90% (*), 95% (**), 99% (***) confidence. OLS regressions, Huber robust standard errors in parenthesis. Disturbance terms are allowed to be correlated across observations in the same school, but not across schools. Test scores were normalized such that comparison group test scores had mean zero and standard deviation one. Indicator variables are included in both specifications for Cohort 1 in 2001, Cohort 1 in 2002, and Cohort 2 in 2002 (coefficient estimates not shown). Restricted sample includes students who were registered in grade 6 (cohort 1) or grade 5 (cohort 2) in January 2001, in schools that did not pull out of the program, for whom we have mean school test score data in 2000, and who took the 2001 or 2002 test.

Table 7: Program Impact on School Participation,
Cohorts 1 and 2 Girls and Boys, Busia District (Panel A),
and Teacher attendance, Busia district (Panel B)

Panel A: Student school participation	Dependent variable: Average Student School Participation (2001, 2002)			
	Busia Girls and Boys		Busia Girls	Busia Boys
	(1)	(2)	(3)	(4)
Program school	0.047*	0.050**		
	(0.025)	(0.024)		
Male * Program School		-0.007		
		(0.017)		
Male		-0.021		
		(0.012)		
Program impact, Cohort 1 (in 2001)			0.062	0.084*
			(0.042)	(0.051)
Program impact, Cohort 2 (in 2002)			0.019	-0.024
			(0.022)	(0.033)
Post-competition impact, Cohort 1 (in 2002)			0.029	0.016
			(0.030)	(0.027)
Pre-competition impact, Cohort 2 (in 2001)			0.094*	0.096*
			(0.049)	(0.060)
Mean school test score, 2000	0.015	0.015	0.014	0.096
	(0.016)	(0.016)	(0.015)	(0.060)
Sample Size	8422	8422	4021	4401
R^2	0.01	0.01	0.90	0.88
Mean of dependent variable	0.85	0.85	0.86	0.84

Panel B: Teacher attendance	Dependent variable: Teacher attendance in 2002, Busia district
Program school	0.065***
	(0.027)
Mean school test score, 2000	0.017
	(0.015)
Sample Size	777
R^2	0.02
Mean of dependent variable	0.87

Notes: Significantly different than zero at 90% (*), 95% (**), 99% (***) confidence. OLS regressions, Huber robust standard errors in parenthesis. Disturbance terms are allowed to be correlated across observations in the same school, but not across schools.

Indicator variables are included in all specifications for Cohort 1 in 2001, Cohort 1 in 2002, Cohort 2 in 2001, and Cohort 2 in 2002 in Panel A (coefficient estimates not shown). The sample in Panel A includes students who were registered in grade 6 (cohort 1) or grade 5 (cohort 2) in January 2001, in schools that did not pull out of the program, and for whom we have school mean test score data. Each school participation observation takes on a value of one if the student was present in school on the day of an unannounced attendance check, zero for any pupil that is absent or dropped out, and is coded as missing for any pupil that died, transferred, or for whom the information was unknown. There was one student school participation observation in the 2001 school year, and three in 2002; the 2002 observations are average in the Panel A regressions, so that each school year receives equal weight. The teacher attendance visits were also unannounced, and actual teacher presence at school recorded.

Teacher attendance data in Panel B were collected similarly during three unannounced school visits in 2002.

Table 8: Program Impact on Education Habits, Inputs, and Attitudes in 2002, Restricted Sample, Cohort 2 Girls and Boys, Busia District

	--------Girls--------		--------Boys--------	
Dependent Variables:	Estimated impact	Mean of dep. var.	Estimated impact	Mean of dep. var.
Panel A: Study/Work habits				
Student went for extra coaching in last two days	-0.02	0.34	-0.07	0.39
	(0.05)		(0.07)	
Student used a textbook at home in last week	-0.01	0.87	0.03	0.84
	(0.04)		(0.04)	
Student did homework in last two days	0.04	0.79	-0.02	0.76
	(0.05)		(0.06)	
Teacher asked the student a question in class in last two days	0.06	0.79	0.03	0.83
	(0.05)		(0.05)	
Amount of time did chores at home [a]	0.01	2.64	-0.05	2.44
	(0.07)		(0.05)	
Panel B: Educational Inputs				
Number of textbooks at home	0.27	1.82	0.05	1.60
	(0.17)		(0.15)	
Number of new books bought in last term	0.35	3.95	-0.02	3.73
	(0.27)		(0.19)	
Panel C: Attitudes towards education				
Student prefers school to other activities (index) [b]	0.01	0.72	0.02	0.73
	(0.02)		(0.02)	
Student thinks s/he is a "good student"	0.01	0.75	0.03	0.74
	(0.05)		(0.04)	
Student thinks that being a "good student" means "working hard"	-0.03	0.75	0.04	0.69
	(0.04)		(0.05)	
Student thinks can be in top three in the class	0.00	0.35	-0.03	0.41
	(0.05)		(0.05)	

Notes: Significantly different than zero at 90% (*), 95% (**), 99% (***) confidence. Marginal probit coefficient estimates are presented when the dependent variable is an indicator variable, and OLS regression is performed otherwise. Huber robust standard errors in parenthesis. Disturbance terms are allowed to be correlated across observations in the same school, but not across schools. Each coefficient estimate is the product of a separate regression, where the explanatory variables are a program school indicator, as well as mean school test score in 2000. Restricted sample includes students who were registered in grade 6 in January 2001, in schools that did not pull out of the program, and for whom we have school average test score data in 2000. The sample size varies from 700-850 observations, depending on the extent of missing data in the dependent variable.

[a] Household chores include fishing, washing clothes, working on the farm and shopping at the market. Time doing chores included "never", "half an hour", "one hour", "two hours", "three hours", and "more than three hours" (coded 0-5 with 5 as most time).

[b] The "student prefers school to other activities" index is the average of eight binary variables indicating whether the student prefers a school activity (coded as 1) or a non-school activity (coded 0). The school activities include: doing homework, going to school early in the morning, and staying in class for extra coaching. These capture aspects of student "intrinsic motivation". The non-school activities include fetching water, playing games or sports, looking after livestock, cooking meals, cleaning the house, or doing work on the farm.

Table 9: Test Score Cost-effectiveness of Various Kenyan Primary School Interventions

Project (article)	Average test score gain, Years 1-2 (1)	Cost / pupil (2)	Cost / pupil per 0.1 s.d. gain (3)	Cost / pupil per 0.1 s.d. gain, adjustment for deadweight loss (4)	Cost / pupil per 0.1 s.d. gain, adjustment for deadweight loss and transfers (5)
Girls scholarship program					
Busia and Teso Districts	0.12 s.d.	$4.24	$3.53	$4.94	$1.41
Busia District	0.19 s.d.	$3.55	$1.77	$2.48	$0.71
Teacher incentives (Glewwe et al. 2003)	0.07 s.d.	$2.39	$3.41	$4.77	$1.36
Textbook provision (Glewwe et al. 1997)	0.04 s.d.	$1.50	$4.01	$5.61	$5.61
Deworming project (Miguel and Kremer 2004)	≈ 0	$1.46	+∞	+∞	+∞
Flip chart provision (Glewwe et al. 2004)	≈ 0	$1.25	+∞	+∞	+∞
Child sponsorship program (Kremer et al. 2003)	≈ 0	$7.94	+∞	+∞	+∞

Notes: All costs are in nominal US$ at the time the particular program was carried out (all programs were conducted between 1996 and 2002). The deadweight loss and transfers adjustments are described in section 6 of the text. Column 4 is referred to as "education budget cost effectiveness" in the text and column 5 is referred to as "social cost effectiveness". School participation cost-effectiveness figures are presented in the text. Costs for the child sponsorship program exclude classroom construction.

Appendix Table A: Timeline of the Girls Scholarship Program, 2000-2003

Time	Activity
2000	
November	Grade 5 students in cohort 1 take district exams (these are baseline scores in the econometric analysis).
2001	
March	Announced Girls' Scholarship Program to Head Teachers in all treatment schools. Head Teachers disseminate information to parents and students.
June	Lightning strikes school in Teso (Korisai P.S.)
September – October	NGO holds Parent-Teacher meetings in all schools to remind parents and students of the program and upcoming tests.
September – October	Field officers perform unannounced school visits to collect attendance data.
November	District exams are administered in Busia district and Teso. For grade 6 students in cohort 1, these exams are used to determine first cohort of scholarship winners.
2002	
January	NGO holds school assemblies to announce the first round of winners and give scholarships.
January – October	Field officers perform unannounced visits to schools to collect attendance data.
February – June	Field officers administer the student survey to all grade 5, 6 and 7 students.
November	District exams are administered in Busia district. For grade 6 Busia students in cohort 2, these exams are used to determine second cohort of scholarship winners. Teso district exams are canceled due to Kenyan national elections.
2003	
January	NGO holds school assemblies to announce the second round of winners and give scholarships (only in Busia).
February	NGO administers standardized exams in both Busia and Teso districts. These exams are used to determine the second round of scholarship winners, among Teso students in cohort 2 who were in grade 6 in 2002.