Incentives to Learn

Michael Kremer[*] Harvard University, Brookings Institution, and NBER

Edward Miguel[**] University of California, Berkeley, and NBER

Rebecca Thornton[***] University of Michigan

January 2008

Abstract: We report results from a randomized evaluation of a merit scholarship program in Kenya in which girls who scored well on academic exams at the end of 6th grade had their school fees paid and received a cash grant for school supplies over the next two years. In the sample as a whole, girls eligible for the scholarship showed substantial gains in academic exam scores, and teacher attendance also improved significantly in program schools. There was also evidence of positive externalities: girls with low pre-test scores, who were unlikely to win scholarships, showed test score gains in program schools. We cannot reject the hypothesis that test score gains were the same for girls with low and high pre-test scores. We see no evidence for weakened intrinsic motivation or gaming, and effects persist after incentives were removed. There is also evidence of heterogeneity in program effects, suggesting the impact of incentives is context dependent. In one of the two study districts, test score effects were large, there were positive spillovers to boys, and student attendance increased in program schools. In the other district, attrition complicates estimation, but we cannot reject the hypothesis that there was no program effect.

[*] 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, University of Michigan, 611 Tappan, 300 Lorch Hall, Ann Arbor, MI 48109-1220, USA; rebeccal@umich.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. Kehinde Ajayi and Garret Christensen provided valuable research assistance. 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

In many education systems, those who perform well on exams covering the material of one level of education receive free or subsidized access to the next level of education. Independent of their role in allocating access to higher levels of education, such merit scholarships are attractive to the extent that they can potentially induce greater student effort and that effort is an important input in educational production, potentially with positive externalities for other students.

This paper estimates the impact of a merit scholarship program for girls in Kenyan primary schools. The scholarship schools were randomly selected from among a group of candidate schools, allowing differences in educational outcomes between the program and comparison schools to be attributed to the scholarship. We find evidence for positive program impacts on academic performance: girls who were eligible for scholarships in program schools had significantly higher test scores than comparison school girls. Teacher attendance also improved significantly in program schools, establishing a plausible behavioral mechanism for the test score gains.

The merit scholarship program we study was conducted in two neighboring Kenyan districts. Separate randomizations into program and comparison groups were conducted in each district, allowing for separate analysis by district. In the larger and somewhat more prosperous district (Busia), test scores gains were large among both girls and boys, and teacher attendance also increased, suggesting a possible mechanism for the test score gains. In the smaller district (Teso), the analysis is complicated by attrition of scholarship program schools and students, so bounds on estimated treatment effects are wide, but we cannot reject the hypothesis that there was no program effect there.

We find positive program externalities among girls with low pre-test scores, who were unlikely to win; in fact, we cannot reject the hypothesis that test score gains were the same for girls with low versus high pre-test scores. Evidence from Busia district – where there were positive test score gains overall – that boys also experienced significant test score gains even though they were

ineligible for the scholarship, together with the gains among low-scoring girls, suggests there may be positive externalities to student effort, either directly among students or through the program's impact on teacher effort. Such externalities within the classroom would have important policy implications. Human capital externalities in production are often cited as a justification for government education subsidies (Lucas 1988). However, 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). To the extent that the results from this program generalize, the evidence for positive classroom externalities creates a new rationale for merit scholarships as well as for public education subsidies more broadly.

Many educators remain skeptical about merit scholarships. First, some argue that their benefits flow disproportionately to well-off pupils, exacerbating inequality (Orfield, 2002). Second, while standard economic models suggest incentives should increase individual study effort, some educators note that alternative theories from psychology argue that extrinsic rewards interfere with intrinsic motivation and could thus reduce effort in some circumstances (for a discussion in economics, see Benabou and Tirole, 2003). 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.[1] A third set of concerns relates to multi-tasking and the potential for gaming the incentive system. Binder et al. (2002) argue that while scholarship eligibility in New Mexico increased student grades, the number of completed credit-hours fell, suggesting that students

[1] Early experimental psychology research supported the idea that reward-based incentives increase student effort (Skinner, 1961). However, laboratory research conducted in the 1970s studied behavior before and after pupils received "extrinsic" motivational rewards and found that 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 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 some settings – such as 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. Some economists also argue that incentives' impact depends on context and framing (Akerlof and Kranton, 2005; Fehr and Gächter, 2002; Fehr and List, 2004).

took fewer courses to keep their grades up. Beyond course-load selection, merit award incentives could potentially produce test cramming and even cheating rather than real learning.[2]

Surveys of students in our Kenyan data provide no evidence that program incentives weakened intrinsic motivation to learn or led to gaming or cheating. The program did not lead to adverse changes in student attitudes towards school, nor did it increase extra test preparation tutoring, and program school test score gains remained large in the year following the competition, after incentives were removed. This suggests that the test score improvements reflect real learning.

This paper is related to a number of recent papers on merit awards in education. In the context of tertiary education, Leuven et al. (2003) use an experimental design to estimate the effect of a financial incentive on the performance of Dutch university students. They estimate large positive effects concentrated among academically strong students. Initial results from a large experimental study among Canadian university freshmen suggests no overall exam score gains during the first year of a merit award program, although there is evidence of gains for some girls (Angrist, Lang, and Oreopoulos, 2006). As noted above, U.S. scholarships have stimulated students to get better grades but to take less ambitious course loads (Binder, 2002; Cornwell et al., 2002; Cornwell et al., 2003).

Angrist et al. (2002) and Angrist et al. (2006) show that a Colombian program that provided vouchers for private secondary school to students conditional on maintaining satisfactory academic performance led to academic gains. They note that the impact of these vouchers may have been due not only to expanding school choice, but also to the incentives associated with conditional renewal of scholarships, but they are unable to disentangle these two channels.

The work closest to ours is that of Angrist and Lavy (2002), who examine a scholarship program that provided cash grants for performance on matriculation exams in 20 Israeli secondary schools. In a pilot program that randomized awards among schools, students offered the merit award

[2] Similarly, after the Georgia HOPE college scholarship was introduced, average SAT scores for high school seniors rose almost 40 points (Cornwell et al., 2002), but there was a 2% reduction in completed college credits, a 12% decrease in full course-load completion, and a 22% increase in summer school enrollment (Cornwell et al., 2003).

were 6 to 8 percentage points more likely to pass exams than comparison students. A second pilot that randomized awards at the individual level within a different set of Israeli schools did not produce significant impacts. This could be because program impact varies with context, or possibly because positive within-school spillovers made any program effects in the second pilot difficult to pick up. Our study differs from the Israeli study in several ways, including our estimation of externality impacts, larger school sample size, and richer data on school attendance and student attitudes and time use, which allow us to better illuminate potential mechanisms for the test score results.

2. The Girls Scholarship Program

2.1 Primary and Secondary Education in Kenya

Schooling in Kenya consists of eight years of primary school followed by four years of secondary school. While approximately 85% of primary school age children in western Kenya are enrolled in school (Central Bureau of Statistics, 1999), there are high dropout rates in grades 5, 6, and 7 and only about one-third of students finish primary school. Dropout rates are especially high for girls.[3]

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

The Kenyan central government pays the salaries of almost all teachers, but when the scholarship program we study was introduced, primary schools charged school fees to cover their

[3] For instance, girls in our baseline sample of pupils in grade 6 (in comparison schools) had a dropout rate of 9.9% from early 2001 through early 2002, versus 7.3% for boys.

non-teacher costs, including textbooks for teachers, chalk, and classroom maintenance. These fees averaged approximately US$6.40 (KSh 500)[4] 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 school supplies, certain textbooks, uniforms, and some activities such as taking exams. The project we study was introduced in part to assist the families of high-achieving girls to cover these costs.[5]

2.2 Project Description and Timeline

The Girls' Scholarship Program (GSP) was carried out by Dutch NGO ICS Africa in two rural Kenyan districts, Busia and Teso. Busia is mainly populated by a Bantu-speaking ethnic group (Luhyas) with agricultural traditions while Teso is populated primarily by a Nilotic-speaking group (Tesos) with pastoralist traditions.

There were 127 sample primary schools, 64 of which were invited to participate in the program in March 2001 (Table 1, Panel A). The randomization first stratified schools by district, and by administrative divisions within districts[6], and also stratified them by participation in a past program which provided classroom flip charts.[7] Randomization into program and comparison groups was then carried out within each stratum using a computer random number generator. In line with the initial stratification, we often present results separately by district.

[4] One US dollar was worth 78.5 Kenyan shillings (KSh) in January 2002 (Central Bank of Kenya, 2002).
[5] 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. Moi's successor Mwai Kibaki eliminated primary school fees in early 2003. This time the policy was implemented consistently, in part because the 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 2003 reform.
[6] Divisions are subsets of districts, with eight divisions in all within our sample.
[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 and results. The flip chart program did not affect any measures of educational performance (not shown). Stratification means there are balanced numbers of flipchart and non-flipchart schools across the GSP program and comparison groups.

The NGO awarded scholarships to the highest scoring 15% of grade 6 girls in the program schools within each district (110 girls in Busia and 90 in Teso). Each district (Busia and Teso) had separate tests and competitions for the merit award.[8] Scholarship winners were chosen based on their total test score on district-wide exams administered by the Ministry of Education across five subjects. Schools varied considerably in the number of winners: 56% of program schools (36 of 64 schools) had at least one 2001 winner, and among schools with at least one winner, there was an average of 5.5 winners per school.

The scholarship program provided winning grade 6 girls with an award for the next two academic years. In each year, the award consisted of: (1) a grant of US$6.40 (KSh 500) to cover the winner's school fees, paid to her school; (2) a grant of US$12.80 (KSh 1000) for school supplies paid directly to the girl's family; and (3) public recognition at a school awards assembly held for students, parents, teachers, and local government officials. These scholarships were "full rides," and were substantial considering that Kenyan GDP per capita is only around US$400 and most households in the two districts have incomes below the Kenyan average. Although the program did not include explicit monitoring to make sure that parents purchased school supplies for their daughter, the public presentation in a school assembly likely generated some community pressure to do so.[9] Since many parents would not otherwise have fully paid fees, schools with winners benefited to some degree from the award money paid directly to the school.

Two cohorts of grade 6 girls competed for the 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 were the second cohort (cohort 2), competing in 2002. There were 11,728 students in grade 5 and grade 6 registered in January 2001; these students make up the *baseline*

[8] Note that student incentive impacts could potentially differ in programs where the top students within each school (rather than district-wide) win awards.

[9] It is impossible to determine exactly how the award was spent without detailed household expenditure data, which we lack. However, qualitative interviews conducted by the authors revealed that some winning girls reported that purchases were made from the scholarship money on school supplies such as math kits, notebooks, and pencils.

sample (Table 1, Panel B). Most cohort 1 students had taken the usual end-of-year grade 5 exams in November 2000, and these are used as baseline test scores in the analysis.[10] Because the NGO restricted award eligibility to girls already enrolled in program schools in January 2001 before the program was announced, there was no incentive for students to transfer schools, and incoming transfer rates were in fact low and nearly identical in program and comparison schools (4.4% into program schools and 4.8% into comparison schools).

In March 2001, after random assignment of schools into program and comparison groups, NGO staff met with school headmasters to invite schools to participate; each of the schools chose 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, enumerators began collecting school attendance data during unannounced visits. District exams were given in Busia and Teso in November 2001. Those baseline sample students who took the 2001 test make up the *intention to treat (ITT) sample* (Table 1, Panel C).

As expected, the baseline 2000 test score is a very strong predictor of being a top 15% performer on the 2001 test. Students below the median baseline test score had almost no chance of winning the scholarship. In particular, the odds of winning were only 3% for the bottom quartile of girls in the baseline test distribution and 5% for the second quartile, compared to 13% and 55% in the top baseline quartiles.

Children whose parents had more schooling were also more likely to be in the top 15% of test performers: average years of parent education are approximately one year greater for scholarship winners (10.7 years) than losers (9.6 years), and this difference is significant at 99% confidence. Note, however, that the link between parent education and child test scores is no stronger in program

[10] Unfortunately, there is incomplete 2000 baseline exam data for cohort 2 (when they were in grade 4), especially in Teso district where many schools did not offer an exam, and thus baseline comparisons focus on cohort 1.

schools than comparison schools. There is no statistically significant difference between winners and non-winners in terms of household ownership of iron roofs or latrines (regressions not shown) however, suggesting a weaker link with household wealth.

Official exams were again held in late 2002 in Busia. The government cancelled the 2002 exams in Teso district because of concerns about possible disruptions in the run-up to the December 2002 national elections, so the NGO instead administered its own standardized exams modeled on government tests in February 2003 after the election. Thus the second cohort of winners was chosen in Busia based on the official 2002 district exam, while Teso winners were chosen based on the NGO exam. In this second round, 67% of program schools (43 of 64) had at least one winner, an increase over 2001, and in all, 75% of program schools had at least one winner in either 2001 or 2002.

Enumerators again visited all schools during 2002 to conduct unannounced attendance checks and to administer questionnaires to students, collecting information on their study effort, habits, and attitudes toward school. This student survey indicates that most girls understood program rules, with 88 percent of cohort 1 and 2 girls claiming to have heard of the program. Girls had somewhat better knowledge about program rules governing eligibility and winning than boys: girls were 9.4 percentage points more likely than boys to know that "only girls are eligible for the scholarship" (84% for girls versus 74% for boys), although the vast majority of boys knew they were ineligible.[11] Girls were very likely (70%) to report that their parents had mentioned the program to them, suggesting some parental encouragement.

3. Data and Sample Construction

[11] Note that some measurement error is likely for these survey responses, since rather than being filled in by an enumerator who individually interviewed students, the surveys were filled in by students themselves 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.

In this section we provide information about the dataset used in this paper and discuss program implementation, in particular examining the implications of sample attrition. We then compare characteristics of program and comparison group schools.

3.1 Test Score Data and Student Surveys

Test score data were obtained from the District Education Offices (DEO) in each program district. Test scores were normalized in each district such that scores in the comparison sample (girls and boys together) are distributed with mean zero and standard deviation one.

The 2002 surveys collected information on household characteristics and study habits and attitudes from all cohort 1 and cohort 2 students present in school on the day of the survey. This means that, unfortunately, pupil survey information is missing for those absent from school on the day of the survey. The collection of the survey in 2002, after one year of the program, is unlikely to be a severe problem for many important predetermined household characteristics (e.g. parent schooling, ethnic identity, children in the household), which are not affected by the program. When examining impacts of the scholarship program on school-related behaviors that could have been affected by the scholarship, we examine the effects on cohort 2, who were administered the survey in the year that they were competing for the scholarship.

Finally, school participation data are based on four unannounced checks collected by NGO enumerators, one in September or October 2001 and one in each of the three terms of the 2002 academic year. We use the unannounced check data rather than official school attendance registers, since registers are often unreliable.

3.2 Community Reaction to the Program in Busia and Teso Districts

Community reaction to the program and school-level attrition varied substantially between the two districts where the program was carried out. Historically, Tesos are educationally disadvantaged

relative to Luhyas: in our data, Teso district parents have 0.2 years less schooling than Busia parents on average. There is also a tradition of suspicion of outsiders in Teso, and this has at times led to misunderstandings with NGOs there. A government report noted that indigenous religious beliefs, traditional taboos, and witchcraft practices remain stronger in Teso than in Busia (Government of Kenya, 1986).

Events that occurred during the study period appear to have interacted in an adverse way with these pre-existing factors in Teso district. In June 2001 lightning struck and severely damaged a Teso primary school, killing seven students and injuring 27 others. Although that school was not in the scholarship program, the NGO had been involved with another assistance program there. Some community members associated the lightning strike with the NGO, and this appears to have led some schools to pull out of the girl's scholarship program. Of 58 Teso sample schools, five pulled out immediately following the lightning strike, as did one school located in Busia with a substantial ethnic Teso population.[12] Three of the six schools that pulled out were treatment schools and three were comparison. The intention to treat (ITT) sample students whose schools did not pull out, and whose schools had baseline average school test scores for 2000, comprise the *restricted sample* (Table 1, Panel D).

Structured interviews conducted during June 2003 with a representative sample of 64 teachers in 18 program schools confirm the stark differences in program reception across Busia and Teso districts. When teachers were asked to rate local parental support for the program, 90% of Busia teachers claimed that parents were either "very" or "somewhat positive" but the analogous rate in Teso was only 58%, and this difference across districts is significant at 99% confidence. Thus, although the monetary value of the award was identical everywhere, local social prestige associated with winning may have differed between Busia and Teso.

[12] Moreover, one girl in Teso who won the ICS scholarship in 2001 later refused the scholarship award, reportedly because of negative views toward the NGO.

3.3 Sample Attrition

Approximately 65% of the baseline sample students took 2001 exams. These students are the main sample for the intention to treat analysis. Not surprisingly, given the reported differences in the response to the scholarship program, we find differences in sample attrition patterns across Busia and Teso districts. In Busia, differences between program and comparison schools are small and not statistically significant: for cohort 1, 83% of girls (81% of boys) in program schools and 78% of girls (77% of boys) in comparison schools took the 2001 exam (Table 2, Panel A). Among cohort 2 students in Busia, there is again almost no difference between program and comparison school students in the proportion who take the 2002 exam (52% versus 48% for girls, and 52% versus 53% for boys; Table 2, Panel C). There is more attrition by 2002 as students drop out, transfer schools, or decide not to take the exam.

Attrition patterns in Teso schools are strikingly different: for cohort 1, 62% of girls in program schools (64% of boys) took the 2001 exam, but the rate for comparison school girls is much higher, at 76% (and for boys 77%; Table 2, Panel A). There are also attrition gaps across program and comparison schools in cohort 2, although these are smaller than for cohort 1.[13]

In addition to the six schools that pulled out of the program after the lightning strike, five other schools (three in Teso and two in Busia) had incomplete exam scores for 2000, 2001, or 2002; the remaining schools make up the restricted sample. There was similarly differential attrition between program and comparison students to this *restricted sample* (Table 2, Panel B). Cohort 1 students in the restricted sample who also had both 2000 and 2001 individual test scores comprise the *longitudinal sample* (Table 1, Panel E).

[13] There was lower 2002 attrition in Teso in part 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 2001 government test (see main text above).

To better understand attrition patterns, we use the baseline test scores from 2000 to examine which students were more likely to attrit. Non-parametric Fan locally weighted regressions display the proportion of cohort 1 students taking the 2001 exam as a function of their baseline 2000 test score in Busia and Teso (Figure 1). These plots indicate that Busia students across all levels of initial academic ability had a similar likelihood of taking the 2001 exam. Although, theoretically, the introduction of a scholarship could have induced poor but high-achieving students to take the exam in program schools, we do not find strong evidence of such a pattern in either Busia or Teso. 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 significant in the extreme left tail of the baseline 2000 distribution. This slightly lower attrition rate among low achieving Busia program school students most likely leads to a downward bias (toward zero) in estimated treatment effects, but any bias in Busia appears likely to be small.[14]

In contrast, not only were attrition rates high and unbalanced across treatment groups in Teso, but significantly more high-achieving students took the 2001 exam in comparison schools relative to program schools, and this is likely to bias estimated program impacts toward zero in Teso (Figure 1, Panels C and D). Among high ability cohort one girls in Teso with a score of at least +0.1 standard deviations on the baseline 2000 exam, comparison school students were almost 14 percentage points more likely to take the 2001 exam than program school students, and this difference is statistically significant at 99% confidence; the comparable gap among high ability Busia girls is near zero (not shown). There are similar gaps between comparison and program schools for boys. Pooling boys and girls in Teso, program school students who did not take the 2001 exam scored 0.50 standard

[14] As mentioned above, pupils with high baseline 2000 test scores were much more likely to win an award in 2001, as expected, with the likelihood of winning rising monotonically and rapidly with the baseline score. However, the proportion of cohort 1 program school girls taking the 2001 exam as a function of the baseline score does not correspond closely to the likelihood of winning an award in either district (not shown). This pattern, together with the very high rate of 2001 test taking for boys, and for comparison school girls, indicates that competing for the NGO award was not the main reason most students took the test.

deviations lower on average at baseline (on the 2000 test) than those who took the 2001 exams, but the difference is far less at 0.37 standard deviations in the Teso comparison schools. These attrition patterns in Teso are in part due to the fact that several of the Teso schools that pulled out had relatively high baseline 2000 test scores. The average baseline score of students in schools that pulled out of the program was 0.20 standard deviations in contrast to an average baseline score of 0.01 standard deviations for students in schools that did not pull out of the program, and the estimated difference-in-differences is statistically significant at 99% confidence (regression not shown).

3.4 Characteristics of the Program and Comparison Groups

We utilize 2002 pupil survey data to compare program and comparison students, and find that the randomization was largely successful in creating groups comparable along observable dimensions. We find no significant differences in parent education, proportion of ethnic Tesos, or the ownership of an iron roof across Busia program and comparison schools (Table 3, Panel A). Household characteristics are also broadly similar across program and comparison schools in the Teso main sample, but there are certain differences, including lower likelihood of owning an iron roof among program students (Table 3, Panel B). This may in part be due to the differential attrition across Teso program and comparison schools discussed above.

Baseline test score distributions provide further evidence on the comparability of the program and comparison groups. Formally, in the Busia longitudinal sample we cannot reject the hypothesis that mean 2000 test scores are the same across program and comparison schools for either girls or boys, nor equality of the distributions using the Kolmogorov-Smirnov Test (p-value = 0.32 for cohort 1 Busia girls). In Teso, where several schools dropped out, the hypothesis of equality between program and comparison baseline test scores distributions is rejected at moderate confidence levels (p-value = 0.07 for cohort 1 Teso girls). We discuss the implications of this difference in Teso below.

4. Empirical Strategy and Results

We focus on reduced form estimation of the program impact on test scores. To better understand possible mechanisms underlying test score impacts, we also estimate program impacts on several channels including measures of teacher and student effort. The main estimation equation is:

$$(1) \qquad TEST_{ist} = \alpha + \beta_1 TREAT_s + X_{ist}' \gamma_1 + \mu_s + \varepsilon_{ist}$$

$TEST_{ist}$ is the normalized test score for student *i* in school *s* in the year of the competition (i.e., 2001 for cohort 1 students and 2002 for cohort 2 students) .[15] $TREAT_s$ is the program school indicator and the coefficient β_1 captures the average program impact on the population targeted for program incentives. X_{ist} is a vector that includes the average school baseline (2000) test score when we use the restricted sample and denotes the individual baseline score for the longitudinal sample, as well as any other controls. The error term consists of μ_s, a common school level error component perhaps capturing common local or headmaster characteristics, and ε_{ist}, which captures unobserved student ability or idiosyncratic shocks. In practice we cluster the error term at the school level and include cohort fixed effects, as well as district fixed effects in the regressions pooling Busia and Teso.

4.1 Test Score Impacts

In the analysis, we focus on the *intention to treat (ITT) sample, restricted sample,* and *longitudinal sample.* The ITT sample includes all students who were in the program and comparison schools in 2000, and who had test scores in 2001 (for cohort 1) or in 2002 (cohort 2). The restricted sample consists of students in schools that did not pull out of the program and also had average baseline 2000 test scores, and it contains data for 91% of the schools in the ITT sample. The longitudinal sample contains the restricted sample cohort 1 students who also have individual baseline test scores.

[15] Test scores were normalized separately by district and cohort; different exams were offered each year by district.

[16] We first present estimated program effects among girls in the ITT sample and then move on to the restricted and longitudinal samples. We then turn to results among boys and robustness checks.

ITT sample

The program raised test scores by 0.19 standard deviations for girls in Busia and Teso districts (Table 4, Panel A, Column 1). These effects were strongest among students in Busia where the program increased scores by 0.27 standard deviations, significant at the 90% level. In Teso, the effects were positive, an increase in 0.09 standard deviations, but not statistically significant. 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 large in these specifications. [17]

To limit possible bias due to differential sample attrition across program groups, especially in Teso, we construct non-parametric bounds on program effects using Lee's (2002) trimming method. In the pooled Busia and Teso sample, bounds range from 0.16 to 0.22 standard deviations – relatively tightly bounded effects. In Busia, the bounds are exactly the non-trimmed program estimate of 0.27, due to the lack of differential attrition across groups. The upper and lower bounds of the program effect in Teso are very wide, ranging from -0.17 to 0.23. Under the bounding assumptions in Lee (2002), we thus cannot reach definitive conclusions about the program effect in Teso district.

In Teso, we can also focus on impacts for cohort 2 girls alone, since attrition rates are similar across program and comparison schools for this group (Table 2). Yet the estimated impact remains small in magnitude (estimate 0.04 standard deviations, standard error 0.16, regression not shown). Whichever way one interprets the Teso results – unreliable estimates due to attrition, no program

[16] Recall that test scores in 2000 are missing for most cohort 2 students in Teso district because many schools there did not offer grade 4 exams, so the longitudinal sample only contains cohort 1 students.

[17] Program effects in the ITT sample were similar for both cohorts in the year they competed: the program effect for cohort 1 girls in 2001 is 0.22 standard deviations (standard error 0.13) and the effect for cohort 2 in 2002 is 0.16 (standard error 0.12 – regressions not shown).

impacts, or a combination of both – the program was clearly less successful in Teso at a minimum in the sense that fewer schools chose to take part.

Restricted sample

Among restricted sample girls, there is an overall impact of 0.18 standard deviations (standard error 0.12, Table 4, Panel B, regression 1) which decreases slightly to 0.15 standard deviations but becomes statistically significant at 99% confidence when the mean school 2000 test score is included as an explanatory variable. The average program impact for Busia district girls in the restricted sample is 0.25 standard deviations (standard error 0.07, significant at 99% confidence – regression 3)[18], much larger than the estimated Teso effect, at only 0.01 standard deviations (regression 4).

In the pooled Busia and Teso sample, the Lee bounds range from 0.09 to 0.21 standard deviations, indicating an overall positive effect of the program. In Busia alone, there was very little differential attrition between the treatment and the comparison groups to the restricted sample, thus the upper and lower bounds are still exactly 0.25 standard deviations. The upper and lower bounds in Teso, however, are very wide, ranging from -0.17 to 0.17.

Cohort 1 longitudinal sample

The program raised test scores by 0.19 standard deviations on average among longitudinal sample girls in Busia and Teso district (Table 4, Panel C, regression 1). The average impact falls to 0.12 standard deviations (standard error 0.09, regression 2) when the individual baseline 2000 test score is included as an explanatory variable. The 2000 test score is strongly related to the 2001 test score as expected (point estimate 0.80, standard error 0.02).

[18] For Busia restricted sample girls, impacts are somewhat larger for mathematics, science, and geography/history than for English and Swahili, but differences across subjects are not statistically significant (regression not shown).

Disaggregation by district again yields a large estimated impact for Busia and a much smaller one for Teso. The estimated impact for Busia district is 0.19 standard deviations, standard error 0.12 (Table 4, Panel C, regression 3), while the estimated program impact for Teso district is near zero at -0.01 standard deviations (regression 4), but it is again difficult to reject a wide variety of hypotheses regarding effects in Teso due to attrition: the bounds for girls in Teso district range from -0.19 to 0.16 standard deviations. The Lee bounds for Busia and Teso taken together range from -0.03 to 0.25 standard deviations, while in Busia, the bounds are again relatively tight due to minimal differential attrition across groups.

The test score distribution in program schools shifts markedly to the right for cohort 1 Busia girls (Figure 2, Panel A) while there is a much smaller visible shift in Teso (Panel C).[19] The vertical lines in each figure indicate the minimum score necessary to win an award in each district.

Note that the intention to treat (ITT) analysis leads to larger estimated average program impacts in Busia and Teso districts (0.19 standard deviations, Table 4, Panel A, regression 1) than in the main and longitudinal samples (0.15 standard deviations and 0.12 standard deviations, respectively). This is consistent with the hypothesized downward sample attrition bias noted above.

In sum, the academic performance effects of competing for the scholarship are large among girls. To illustrate the magnitude 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 estimated average program effect for girls roughly corresponds to an additional 0.2 grades worth of primary school learning.

Test score effects for boys

There is some evidence that the program raised test scores among boys, though by less than among girls. Being in a scholarship program school is associated with a 0.08 standard deviation gain in test

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

scores on average among boys in 2001, among the Busia and Teso ITT sample (Table 5, Panel A, regression 1). The gain in Busia, 0.10 standard deviations (regression 2), is larger than in Teso, at 0.04 standard deviations (regression 3), though neither of these effects are significant at traditional confidence levels. The Lee bounds for boys reveal familiar patterns: in the pooled Busia and Teso sample, bounds range from 0.02 to 0.12 standard deviations, but among Busia boys, the bounds are tight, equal to 0.10, while among Teso boys, bounds are wide, from -0.25 to 0.19 standard deviations.

Among restricted sample boys, there is an overall impact of 0.05 standard deviations (Table 5, Panel B, regression 1). In Busia, the program increased test scores among boys by 0.15 standard deviations, statistically significant at 90% confidence (regression 3) – roughly 60% of the size of the analogous effect for Busia girls, at 0.25 standard deviations – while the results for Teso remain close to zero. In the pooled Busia and Teso sample, the Lee bounds are wide (ranging from -0.06 to 0.17 standard deviations), but among Busia boys, the bounds range from 0.09 to 0.18 standard deviations. Among Teso boys, the bounds are again very wide, from -0.25 to 0.18 standard deviations.

In the cohort 1 longitudinal sample, the overall impact is 0.09 standard deviations (Table 5, Panel C, regression 1), and this rises to 0.14 standard deviations (standard error 0.06, regression 2) and becomes statistically significant at 99% confidence when the individual baseline test score is included as an explanatory variable. Effects are again concentrated in Busia (regression 3) with smaller, non-significant effects among Teso boys (regression 4). Longitudinal sample Busia boys show some visible gains (Figure 2, Panel B).

Although average program effects among boys, who were not eligible for the scholarship, are much smaller than among girls in the ITT and restricted samples, we cannot reject equal treatment effects for girls and boys in the longitudinal sample (regression not shown). Below in section 4.2, we discuss possible mechanisms for effects among boys, including our leading explanations of higher teacher attendanceand within-classroom externalities among students.

Heterogeneous impacts by academic ability

We next test whether test score effects differ as a function of baseline academic performance, focusing the analysis on the cohort 1 longitudinal sample (who have pre-program 2000 test data). The average treatment effects for girls across the four baseline test quartiles (from top to bottom) are 0.00, 0.23, 0.13, and 0.12 standard deviations, respectively (Table 6, Panel A, regression 1), and we cannot reject the hypothesis that treatment effects are equal in all quartiles (F-test p-value = 0.31). Although estimating the program effect separately for each quartile reduces statistical power somewhat, the positive and large estimated test score gains among girls with little to no chance of winning the award is suggestive evidence for positive externalities . As expected, effects are larger among Busia girls at 0.08, 0.29, 0.19, and 0.23 standard deviations, with the largest gains in the second quartile, namely those students striving for the top 15% winning threshold (regression 2). Effects for Teso students are again close to zero.

Evidence on program gains throughout the baseline test score distribution is presented using a non-parametric approach in Figure 3, including bootstrapped 95% confidence bands on the treatment effects. Once again, treatment effects are visibly larger among Busia students.

Robustness checks

Estimates are similar when individual characteristics collected in the 2002 student survey (i.e. student age, parent education, and household asset ownership) are included as additional explanatory variables.[20] Interactions of the program indicator with these characteristics are not statistically significant at traditional confidence levels for any characteristic (regressions not shown), implying that test scores did not increase significantly more on average for students from higher

[20] 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).

socioeconomic status households.[21] Theoretically, spillover benefits could also be larger in schools with more high achieving girls striving for the award. We estimate these effects by interacting the program indicator with measures of baseline school quality, including the mean 2000 test score as well as the proportion of grade 6 girls that were among the top 15% in their district on the 2000 test. Neither of these terms are significant at traditional confidence levels (not shown), so we cannot reject the hypothesis that average effects were the same across schools at various academic quality levels.

4.2 Channels for Merit Scholarship Impacts

Teacher attendance

The estimated program impact on overall teacher school attendance in the pooled Busia and Teso sample is large and statistically significant at 4.8 percentage points (standard error 2.0 percentage points, Table 7, panel A, regression 1).[22] Together with the test score impacts above, teacher attendance is the second educational outcome for which there are large, positive, and statistically significant impacts in the pooled Busia and Teso district sample.

In our data, it is difficult to distinguish between teacher attendance in grade 6 classes versus other grades. The same teacher often teaches a subject (e.g. mathematics) in several different grades and the data set does not allow us to isolate particular teacher attendance observations by the grade they were teaching at the time of the attendance check. However, data from another sample of primary schools in Busia and Teso reveals that 62.9% of all teachers teach at least one grade 6 class. If all attendance gains were concentrated among this subset of teachers, the implied program effect for teachers that teach at least one grade 6 class would be an even larger 4.8 / 0.629 = 7.6 percentage point increase in attendance.

[21] Note that although the program had similar test score impacts across socioeconomic backgrounds, students with more educated parents nonetheless were more likely to win because they have higher baseline scores.

[22] These results are for all regular (senior and assistant) classroom teachers. A regression that also includes nursery teachers, administrators (head teachers and deputy head teachers), and classroom volunteers yields a somewhat smaller but still statistically significant point estimate of 3.6 percentage points (standard error 1.6, not shown).

Although teacher attendance gains are significant in the pooled sample, the strongest effects are once again in Busia district: the impact on teacher attendance there was 7.0 percentage points (standard error 2.4, significant at 99% confidence, Table 7, Panel A, regression 2), reducing overall teacher absenteeism by approximately one half. The implied effect among those teaching grade 6 if attendance gains were concentrated in this group is 11.1 percentage points. Note that the mean school baseline 2000 test score is positively but only moderately correlated with teacher attendance and all results are robust to excluding this term. Estimated program impacts in Busia are not statistically significantly different by teacher's gender or experience (not shown). Program impacts on teacher attendance are positive but smaller and not significant in Teso (1.6 percentage points, regression 3).

Recall the ITT sample gains are 0.27 standard deviations for Busia girls (Table 4, Panel A) and 0.10 standard deviations for Busia boys (Table 5, Panel A). A study in a rural Indian setting finds that a 10 percentage point increase in teacher attendance increased average primary school test scores by 0.10 standard deviations there (Duflo and Hanna, 2006). If a similar relationship holds in rural Kenya, the estimated teacher attendance gain of 11.1 percentage points would explain a bit less than half of the overall test score gain among girls, and almost exactly the entire effect for boys. The remaining gains for girls are likely to be due to increased student effort and, more speculatively, within classroom spillovers.

Several mechanisms could potentially have increased teacher effort in response to the merit scholarship program, including ego rents, social prestige, and even gifts from winners' parents. While we cannot rule out those mechanisms, we have anecdotal evidence that increased parental monitoring played a role. The June 2003 teacher interviews suggest greater parental monitoring occurred in Busia but not in Teso. 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 another 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." There were no comparable accounts from teachers in Teso schools.

Yet there is little quantitative evidence the program changed teacher behavior beyond increasing attendance. Program school students were no more likely than comparison students to report being called on by a teacher in class during the last two days, nor to have done more homework (as we discuss in Table 8 below). Similarly, program impacts on classroom inputs, including the number of flipcharts and desks (using data gathered during 2002 classroom observations) are similarly near zero and not statistically significant (regressions not shown).

One way teachers could potentially game the system is by diverting their effort towards students eligible for the program, but there is no statistically significant difference in how often girls are called on in class relative to boys in the program versus comparison schools based on student survey data (not shown), indicating that program school teachers probably did not substantially divert attention to girls. This finding, together with the increased teacher attendance, provides a concrete explanation for spillovers for boys, namely, greater teaching effort directed to the class as a whole.

Student attendance

We find suggestive evidence of student attendance gains. The dependent variable is school participation during the competition year. 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, although attrition concerns are not entirely eliminated since school participation data was not collected at schools that dropped out of the program.[23] For cohort 1

[23] In the Busia comparison sample, girls with higher average school participation have significantly higher baseline test scores: cohort one girls who were present in school on the first visit during the competition year (2001), had baseline 2000 scores 0.14 standard deviations than those who were not present (standard error 0.08, regression not shown). This cross-sectional correlation is consistent with the view that improved attendance may be an important channel through which the program generated test score gains, although by itself is not decisive due to potential omitted variable bias.

students, there is one observation in 2001, while for cohort 2 there were three unannounced attendance checks in 2002.

While the estimated program impact on school participation among girls in the pooled Busia and Teso sample is near zero, the impact in Busia is positive at 3.2 percentage points (significant at 90%, Table 7, Panel B, regression 2). This corresponds to a reduction of roughly one quarter in mean school absenteeism.

The largest student attendance effects occurred in 2001, corresponding to the competition year for cohort one students: for cohort 1 Busia students, there was a 8 percentage point increase in attendance. There is also some evidence of pre-program effects in 2001 among cohort 2 students in both Busia and Teso sample (regressions not shown). 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 attendance spiked in the run-up to term 3 exams, due to cramming, for instance. We cannot reject the hypothesis that school participation gains among cohort 1 girls are equal across baseline 2000 test score quartiles (not shown). School participation gains are much smaller for boys, both overall and in Busia district (Table 7, Panel C).

The scholarship program had no statistically significant effect on dropping out of school in the competition year in either Busia or Teso, among boys or girls (not shown).

Post-competition test score effects

In the restricted sample, the program not only raised test scores for cohort 1 girls when it was first introduced in 2001 but appears to have continued boosting their scores in 2002: the estimated program impact for cohort 1 girls in 2002 is 0.12 standard deviations (standard error 0.08, p-value = 0.12, not shown). This is suggestive evidence that the program had lasting effects on learning, rather than simply encouraging cramming or cheating. When we focus on Busia district alone, there is even stronger evidence, with a coefficient estimate of 0.24 standard deviations (standard error 0.09,

significant at 95% confidence, not shown)[24]. These persistent gains can be seen in Figure 4 (especially in Panel A, for Busia girls), which presents the distribution of test scores for longitudinal sample students. Once again there are no detectable gains in Teso district (Panels C and D).

February 2003 exams provide further evidence. Although originally administered because 2002 exams were cancelled in Teso district, they were also offered in our Busia sample schools. In the restricted sample the average program impact for cohort 1 Busia girls was 0.19 standard deviations (standard error 0.07, statistically significant at 99% confidence – regression not shown).

Student attitudes and behaviors

We also attempted to measure "intrinsic motivation" for education directly using eight survey 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 fetching water or playing sports. When the 2002 survey was administered, cohort 2 girls were competing for the award (cohort 1 girls had already competed in 2001), so in what follows we focus on cohort 2. Overall, students report preferring the school activity in 72% of the questions. There are no statistically significant differences in this index across the program and comparison schools for either girls or boys (Table 8, Panel A), and thus no evidence that external incentives dampened intrinsic motivation to learn as captured by this measure.[25] 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.

There is also no evidence that study habits changed adversely in other dimensions measured by the 2002 student survey. Program school students were no more or less likely than comparison

[24] The significant effect of the scholarship program on second year test scores among cohort 1 students is not merely due to the winners in those schools: we find no significant impacts of winning the award on 2002 test scores. In addition, the post-competition results remain significant when excluding the winners from the sample (not shown).
[25] In an SUR framework including all attitude measures in Table 8 Panel A, we cannot reject the hypothesis that the joint effect is zero for girls (p-value=0.92) and boys (p-value=0.36).

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 in the pooled Busia and Teso sample (Table 8, Panel B) as well as in each district separately (not shown). In the case of chores, the estimated zero impact indicates the program did not lead to lost home production, suggesting that any increased study effort came out of children's leisure or through intensified effort during school hours.

We also find weak evidence of increased investments in girls' school supplies by households, suggesting another possible mechanism for test score gains. In the pooled Busia and Teso sample, the estimated program impact on the number of textbooks girls have at home and the number of new books (the sum of new textbooks and exercise books) their household recently purchased for them are positive though not statistically significant (Table 8, Panel C). Point estimates for Busia girls alone are similarly positive and somewhat larger, and in the case of textbooks at home, marginally statistically significant (0.27 additional textbooks, standard error 0.17, not shown).[26]

One concern related to the interpretation of our findings is the possibility of cheating on the exams, but this appears unlikely. Exams in Kenya are administered by outside monitors, and district records from those monitors indicate no documentation of cheating in any sample school in either 2001 or 2002. Several above findings 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 direct incentive to cheat, and there were substantial gains among program school boys ineligible for the scholarship (although cheating by teachers could still potentially explain that latter result).

Regarding "cramming," there is no evidence that extra test preparation coaching increased in the program schools for either girls or boys (Table 8, Panel B).[27] A separate teacher incentive project

[26] There is a significant increase in textbook use among Busia program girls in cohort 1 in 2002: girls in program schools report using textbooks at home 5 percentage points (significant at 90% confidence) more than comparison school girls, further suggestive evidence of greater parental investment. However, there are no such gains among the cohort 2 students competing for the award in 2002.

[27] Similarly, recent work on high-stakes tests suggests that individuals may increase their effort only during the actual test-taking, potentially making test scores a good measure of effort that day but an unreliable measure of actual learning or ability (Segal, 2006). While the tests in Kenya were high-stakes, the fact that we also see similar

run earlier in the same region led to increased test preparation sessions and boosted short-run test scores, but had no measurable effect on either student or teacher attendance or long-run learning, consistent with the hypothesis that teachers responded to that program by seeking to manipulate short run scores (Glewwe et al., 2003). There is no evidence for similar effects in the program we study, although a definitive explanation for the differences across these two programs remains elusive.

Another issue is the Hawthorne effect, namely, an effect driven by students knowing 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 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 Kremer, Miguel, and Thornton, 2005) but no other program generated such substantial test score gains.

Merit scholarships and inequality

The equity critiques of merit scholarships resonate with our results in one sense: the scholarship award winners do tend to come from families where parents have significantly more years of educational attainment, and thus from relatively advantaged households (see section 2.2). But in terms of student test score performance, we find that program impacts are not just concentrated among the best students: there are positive estimated treatment effects for girls throughout the baseline test score distribution (Table 6). There are also no significant program interaction effects with household socioeconomic measures, including parent education, and even girls with poorly educated parents gained from the program.

Program impacts on inequality per se are important in both theoretical and policy debates over merit scholarships. Perhaps not surprisingly, given the observed gains throughout the test score

---

test score gains for cohort 1 in 2002 when there was no longer a scholarship at stake indicates that the effects we estimate are likely due to real learning rather than solely to increased motivation on the competition testing day.

distribution, there was only a small overall increase in test score variance for cohort 1 program school girls relative to cohort 1 comparison girls in the ITT sample: the overall variance of test scores rises from 0.88 in 2000 at baseline to 0.94 in 2001 and 0.97 in 2002 for Busia 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.[28] The changes in test variance over time for boys in Busia program versus comparison schools, as well as for Teso girls and boys, are similarly small and never statistically significant (not shown). [29]

5. Conclusion

Merit-based scholarships are an important part of the educational system in many countries, but are often debated on the grounds of effectiveness and equity. We present evidence that such programs can raise test scores and boost classroom effort as captured in teacher attendance. We also find suggestive evidence for program spillovers. In particular, we estimate positive program effects among girls with low pre-test scores who had little realistic chance of winning the scholarship. In the district where the program had larger positive effects, even boys – who were ineligible for awards – show somewhat higher test scores. These positive externalities are likely to be due to higher teacher attendance or positive peer effects among students – or a combination of these reasons. Our data are not able to distinguish which is the greater cause of the estimated test score impacts.

In addition to the girls' merit scholarship program, a number of other school programs have recently been conducted in the study area: a teacher incentive program (Glewwe et al., 2003),

[28] 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 reduce inequality in program relative to comparison schools. We thank Joel Sobel for this point.
[29] One potential concern with these figures is the changing sample sizes in 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 no significant differences in test variance across program and comparison schools in either year.

textbook provision program (Glewwe et al., 1997), flip chart program (Glewwe et al., 2004), deworming program (Miguel and Kremer, 2004), and a child sponsorship program that provided a range of inputs (Kremer et al., 2003). By comparing the cost-effectiveness of each program, we conclude that providing merit scholarship incentives is arguably the most cost-effective way to improve test scores among these six programs. Considering Busia and Teso districts together, the girls scholarship program is almost exactly as cost-effective in boosting test scores as the teacher incentive program, followed by textbook provision (see Kremer, Miguel, and Thornton, 2005 for details). Considering Busia alone, girls' scholarships are more cost-effective than the other programs.

Our evidence on within classroom learning externalities has several implications for research and public policy. Methodologically, these externality effects suggest that other merit award program evaluations that randomize eligibility among individuals within schools may understate program impacts, due to contamination across treatment and comparison groups. This issue may be important for the interpretation of results from the other recent merit award studies described in the introduction, and, more broadly, for any education program evaluation that assigns treatment to a subset of students within a classroom.[30]

Substantively, a key reservation about merit awards for educators has been the possibility of adverse equity impacts. It is likely that relatively advantaged students gained the most from the program we study: scholarship winners do come from the most educated households. However, groups with little chance at winning an award, including girls with low baseline test scores and poorly educated parents, also gained considerably in merit scholarship program schools.

One way to spread the benefits of a merit scholarship program even more widely could be to restrict the scholarship competition to poorer pupils, schools, or regions, or to conduct multiple competitions, each in a restricted geographic area. For instance, if each Kenyan location – a small administrative unit – awarded merit scholarships to its residents independently of other locations,

[30] Miguel and Kremer (2004) also discuss treatment effect estimation in the presence of externalities.

children would only compete against others in the same area, where many have comparable socioeconomic conditions. To the extent that such a policy would put more students near the margin of winning a scholarship, it could potentially generate even greater incentive effects and spillovers.

More speculatively, the spillover benefits among students with little chance of winning the award are consistent with a model of strategic complementarity between the effort levels of girls eligible for the award, the effort of teachers, and of other students. If such complementarity is sufficiently strong, there could be multiple equilibria in the classroom learning culture. Educators often stress the importance of classroom culture. Multiple equilibria could help explain why 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).

Our finding that merit scholarships motivate students to increase effort, and that this may generate positive externalities for other students, provides a potential public policy rationale for the widespread practice of structuring education systems so that those who perform well in one level of education are entitled to free or subsidized access to the next level. It also suggests that centralized education systems that are responsible for not just higher education but also for lower levels of schooling may prefer different higher education admissions procedures than individual institutions of higher education would choose in a decentralized system. Individual institutions might choose to admit students based on a mix of aptitude and achievement tests that optimally predicts achievement in higher education. Relative to this benchmark, a centralized education authority might prefer to place higher weight on achievement tests because this creates incentives for students to exert higher effort in lower levels of education, creating positive spillovers for other students. This could potentially help explain why many European countries, with their more centralized education systems, place more weight on achievement relative to aptitude testing in determining admission to

higher education. It is also consistent with the view that student effort in secondary school is low in the United States relative to Europe. (See Harbaugh (2003) for an argument along these lines.)

We find especially large average program effects on girls' test scores in Busia, on the order of 0.2 to 0.3 standard deviations, but do not find significant effects in neighboring Teso district. Our inability to find these effects may in part be due to differential sample attrition across Teso program and comparison schools, which complicates the econometric analysis. However, it may also simply reflect the lower value placed on winning the merit award there, or a lack of local political support among some parents and community opinion leaders.

Establishing where, how, and why student incentive programs succeed or fail thus remains an important priority for future research. The sharply different effects of the program impacts we estimate – measured either by test scores or program participation – across two neighboring districts raises important questions about how local responses to merit awards vary across time and space, and thus how successfully student incentive programs of this kind will scale up. The recent literature (surveyed in the introduction) has not yet yielded consistent findings about merit scholarship impacts. Thus for example, Angrist and Lavy (2002) find that one of their two Israeli pilot programs generated positive impacts on learning and the other did not. One of the two experimental university merit award programs in OECD countries has produced positive test score impacts among high-achieving students and negative impacts among low-achieving students (Leuven et al., 2003) while a second largely found no effects (Angrist, Lang, and Oreopoulos, 2006). It may be impossible for any single study to establish why these types of programs generally succeed or fail, but accumulating evidence across studies may be more promising. However, our study provides no evidence that merit scholarships generate the adverse impacts on academic performance sometimes feared by educators and psychologists or that other leading objections to merit awards are empirically important.

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.

Akerlof, George, and Rachel Kranton. (2005). “Identity and the Economics of Organizations,” *Journal of Economic Perspectives*, 19(1), 9-32.

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.

Angrist,. Joshua, Eric Bettinger and Michael Kremer. (2006) “Long-Term Consequences of Secondary School Vouchers: Evidence from Administrative Records in Colombia”, *American Economic Review*, 847-862.

Angrist, Joshua, Daniel Lang, and Philip Oreopoulos. (2006). "Lead Them to Water and Pay Them to Drink: An Experiment with Services and Incentives for College Achievement", NBER WP#12790.

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 paper.

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).

Duflo, E, and R. Hanna (2006). "Monitoring Works: Getting Teachers to Come to School", unpublished manuscript MIT and NYU.

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.

Fehr, E. and John List. (2004). "The Hidden Costs And Returns Of Incentives—Trust and Trustworthiness Among CEOs". *Journal of the European Economic Association*, 2(5).

Fehr, E. and S. Gächter. (2002). "Do Incentive Contracts Crowd Out Voluntary Cooperation?", Institute for Empirical Research in Economics, University of Zürich, Working Paper No. 34.

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.

Harbaugh, Rick. (2003). "Achievement vs. Aptitude", Claremont Working Papers in Economics Series.

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, Miguel, Edward, and Rebecca Thornton. (2005). "Incentives to Learn", *National Bureau of Economic Research Working Paper* #10971.

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).

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.

Segal, C. (2006). "Incentives, Test Scores, and Economic Success", Harvard Business School Mimeo.

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.

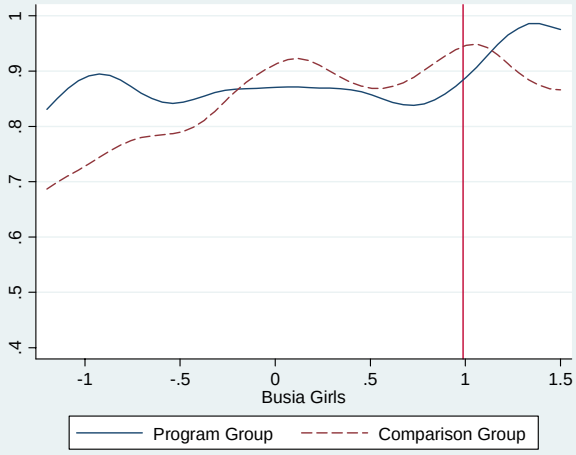
United Nations. (2003). *The Right to Education*, Economic and Social Council Special Rapporteur Katarina Tomasevski. Available online at: (http://www.right-to-education.org/content/unreports/unreport12prt1.html#tabel1).

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

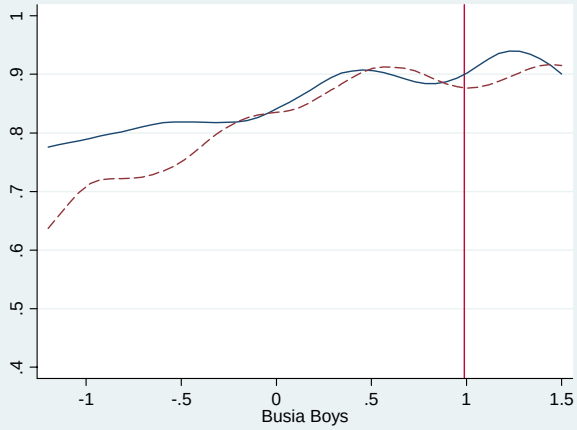
World Bank. (2004). *Strengthening the Foundation of Education and Training in Kenya: Opportunities and Challenges in Primary and General Secondary Education*. Nairobi.

Figure 1: Proportion of Baseline Students with 2001 Test Scores by Baseline (2000) Test Score
Cohort 1 (non-parametric Fan locally weighted regressions)

Panel (A) – Busia Girls

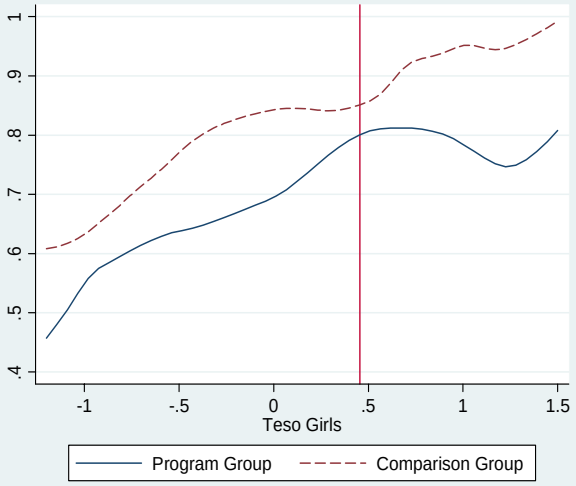


Panel (B) – Busia Boys

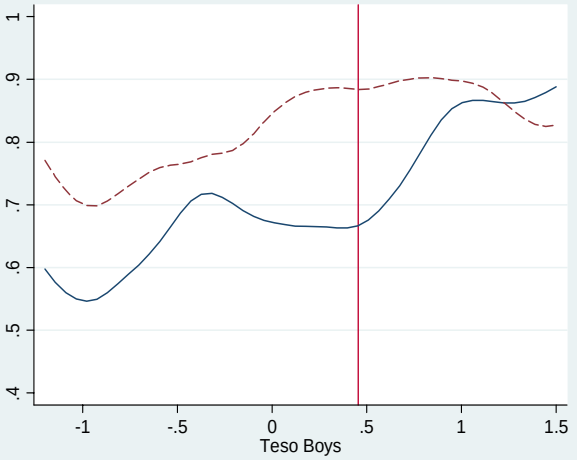


Vertical line represents the minimum winning score in 2001.

Panel (C) – Teso Girls



Panel (D) – Teso Boys

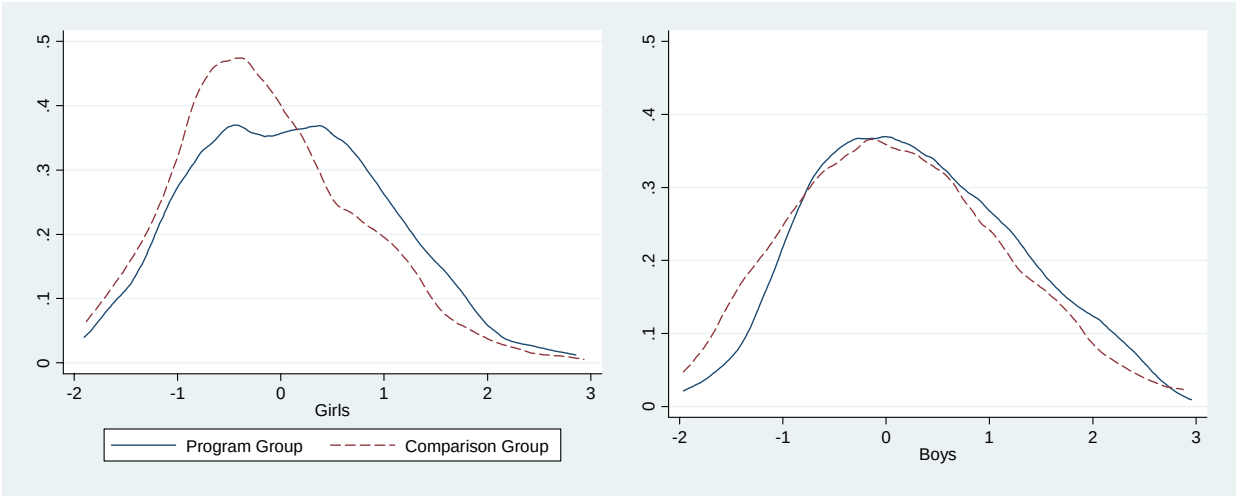


Vertical line represents the minimum winning score in 2001.

Notes: These figures present non-parametric Fan locally weighted regressions using an epanechnikov kernel and a bandwidth of 0.7. The sample used in these figures includes students in the baseline sample who have 2000 test scores.

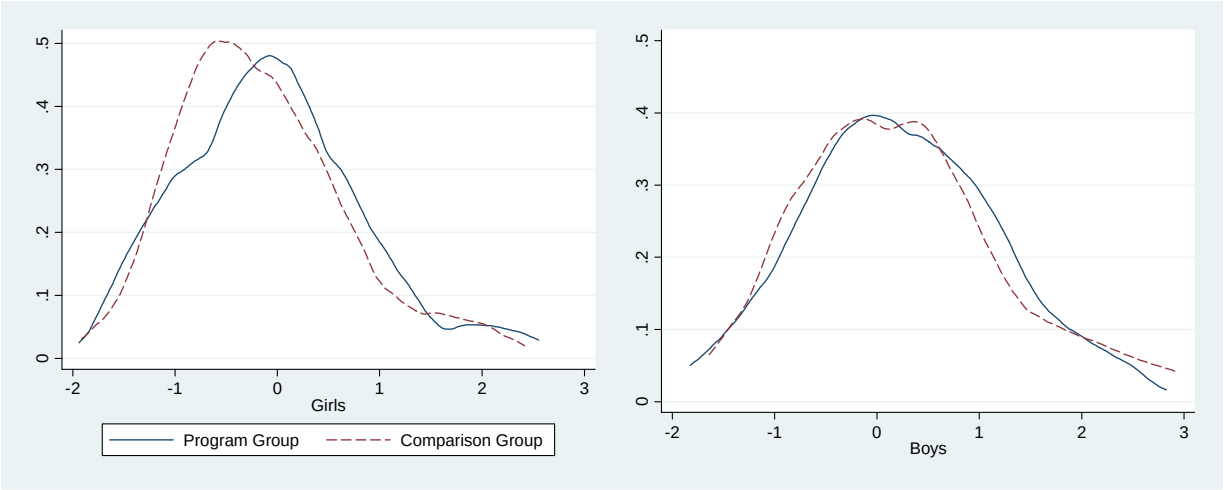
Figure 2: Competition Year Test Score Distribution (Cohort 1 in 2001, Cohort 2 in 2002)
ITT Sample (Non-parametric kernel densities)

Panel (A) – Busia Girls



Panel (B) – Busia Boys

Panel (C) – Teso Girls

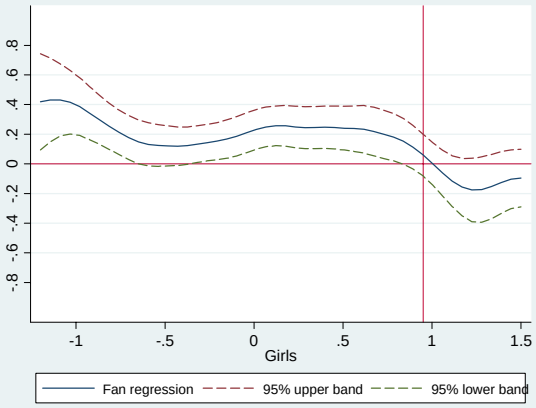


Panel (D) – Teso Boys

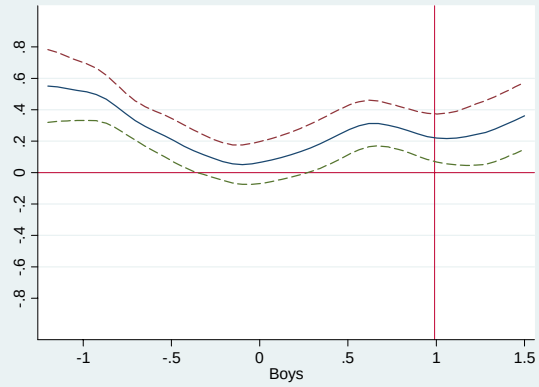
Notes: These figures present non-parametric kernel densities using an epanechnikov kernel.

Figure 3: Year 1 (2001) Test Score Impacts by Baseline (2000) Test Score Difference between Program and Comparison Schools, Longitudinal Sample (non-parametric Fan locally weighted regression)

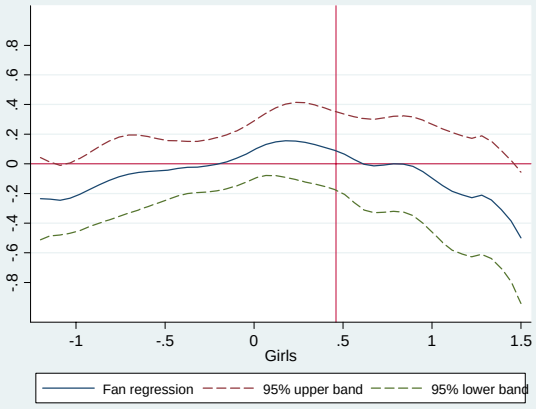
Panel (A) – Busia Girls



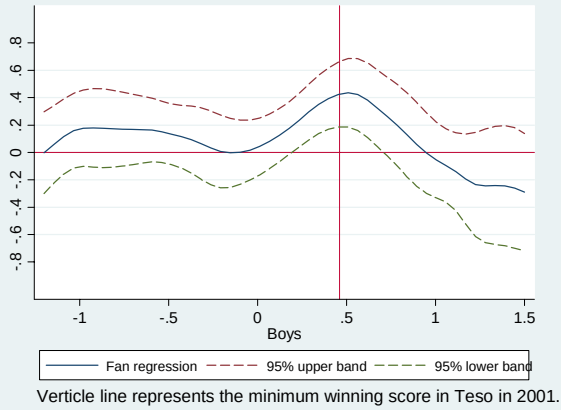
Panel (B) – Busia Boys



Panel (C) – Teso Girls



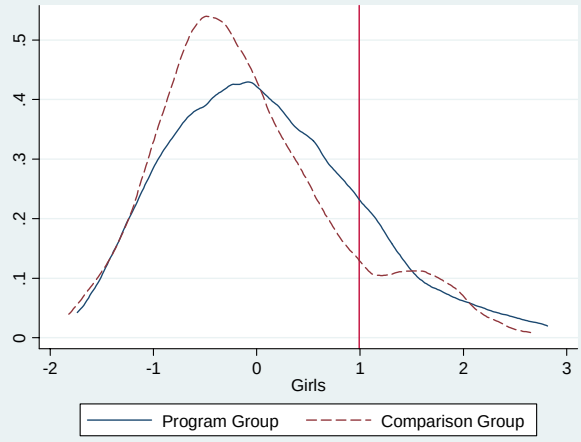
Panel (D) – Teso Boys



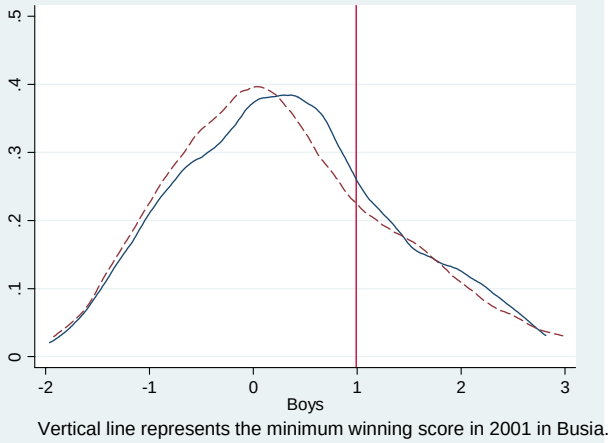
Notes: These figures present non-parametric Fan locally weighted regressions using an epanechnikov kernel and a bandwidth of 0.7. Confidence intervals were constructed by drawing 50 bootstrap replications.

Figure 4: Year 2 (2002) Test Scores, Cohort 1, Longitudinal Sample
(non-parametric kernel densities)

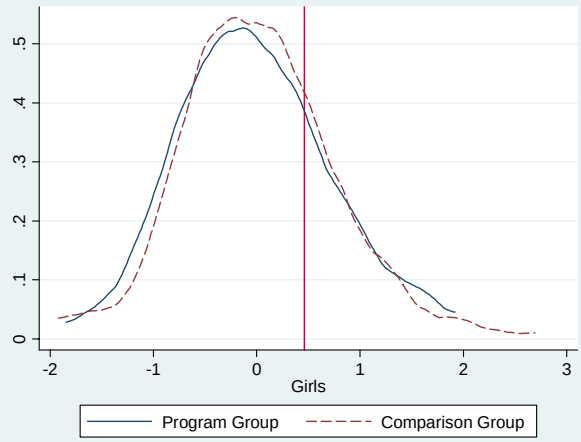
Panel (A) – Busia Girls



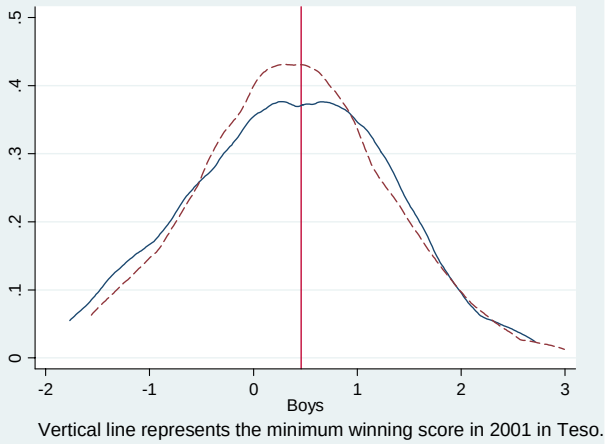
Panel (B) – Busia Boys



Panel (C) – Teso Girls



Panel (D) – Teso Boys



Notes: These figures present non-parametric kernel densities using an epanechnikov kernel.

Table 1: Summary Sample Sizes

	Busia District				Teso District			
Panel A:	Program Schools		Comparison Schools		Program Schools		Comparison Schools	
Number of schools	34		35		30		28	
	Cohort 1		Cohort 2		Cohort 1		Cohort 2	
Panel B: Baseline sample	Program	Comparison	Program	Comparison	Program	Comparison	Program	Comparison
Number of girls	744	767	898	889	571	523	672	572
Number of boys	803	845	945	1024	602	503	739	631
Panel C: Intention to treat (ITT) sample								
Number of girls	614	599	463	430	356	397	399	344
Number of boys	652	648	492	539	385	389	508	445
Panel D: Restricted sample								
Number of girls	588	597	449	427	304	342	380	333
Number of boys	607	648	470	531	328	334	484	436
Panel E: Longitudinal sample								
Number of girls	360	408	--	--	182	203	--	--
Number of boys	398	453	--	--	205	219	--	--

Notes: The baseline sample refers to all students that were registered in grade 6 (cohort 1) or grade 5 (cohort 2) in January 2001. The ITT sample consists of all baseline sample students with either 2001 (cohort 1) or 2002 (cohort 2) test scores. The restricted sample consists of ITT sample students in schools that did not pull out of the program, with average school test scores in 2000. The longitudinal sample contains those cohort 1 restricted sample students who took the 2000 test. A dash (-) indicates that the data are unavailable (for instance, cohort 2 in not included in the longitudinal sample).

Table 2: Proportion of Baseline Sample Students in Other Samples

	--------Busia District--------			--------Teso District--------		
	Program	Comparison	Difference (s.e.)	Program	Comparison	Difference (s.e.)
Panel A: Cohort 1 in ITT sample						
Girls	0.83	0.78	0.04	0.62	0.76	-0.14***
			(0.03)			(0.04)
Boys	0.81	0.77	0.05	0.64	0.77	-0.13***
			(0.04)			(0.04)
Panel B: Cohort 1 in restricted sample						
Girls	0.79	0.78	0.01	0.53	0.65	-0.12
			(0.04)			(0.09)
Boys	0.76	0.77	-0.01	0.54	0.66	-0.12
			(0.06)			(0.09)
Panel C: Cohort 2 in ITT sample						
Girls	0.52	0.48	0.03	0.59	0.60	-0.01
			(0.04)			(0.08)
Boys	0.52	0.53	-0.01	0.69	0.71	-0.02
			(0.04)			(0.07)
Panel D: Cohort 2 in restricted sample						
Girls	0.50	0.48	0.02	0.57	0.58	-0.02
			(0.04)			(0.09)
Boys	0.50	0.52	-0.02	0.65	0.69	-0.04
			(0.04)			(0.08)

Notes: Standard errors in parentheses. Significantly different than zero at 90% (*), 95% (**), 99% (***) confidence. The denominator for these proportions consists of the baseline sample, all grade 6 (cohort 1) or grade 5 (cohort 2) students who were registered in school in January 2001. Cohort 2 data for Busia district students is based on the 2002 Busia district exams, which were administered as scheduled in late 2002. Cohort 2 data for Teso district students is based on the February 2003 NGO exam.

Table 3: Demographic and Socio-Economic Characteristics Across Program and Comparison Schools Cohort 1 and Cohort 2, Busia and Teso Districts

	----------Girls----------			----------Boys----------		
Panel A: Busia District	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)	10.8	10.4	0.4 (0.4)	10.2	9.9	0.3 (0.3)
Mother's education (years)	9.2	8.8	0.4 (0.3)	8.3	8.1	0.2 (0.4)
Proportion ethnic Teso	0.07	0.06	0.01 (0.03)	0.07	0.07	0.01 (0.03)
Iron roof ownership	0.77	0.77	0.00 (0.03)	0.72	0.75	-0.03 (0.03)
Test score 2000–baseline sample (cohort 1 only)	-0.05	-0.12	0.07 (0.18)	0.04	0.10	-0.07 (0.19)
Test score 2000–main sample (cohort 1 only)	0.07	0.03	0.04 (0.19)	0.15	0.28	-0.13 (0.19)
Panel B: Teso District						
Age in 2001	14.0	13.8	0.20 (0.18)	14.1	14.1	-0.05 (0.18)
Father's education (years)	11.0	10.8	0.2 (0.4)	10.0	10.0	0.0 (0.4)
Mother's education (years)	8.5	8.4	0.1 (0.5)	7.5	8.2	-0.7 (0.5)
Proportion ethnic Teso	0.84	0.80	0.05 (0.05)	0.85	0.80	0.05 (0.04)
Iron roof ownership	0.58	0.67	-0.09** (0.04)	0.49	0.59	-0.09** (0.04)
Test scores 2000–baseline sample (cohort 1 only)	0.04	-0.11	0.15 (0.18)	0.19	0.10	0.09 (0.17)
Test scores 2000–main sample (cohort 1 only)	0.06	0.06	0.01 (0.19)	0.20	0.25	-0.05 (0.17)

Notes: Standard errors in parentheses. Significantly different than zero at 90% (*), 95% (**), 99% (***) confidence. Sample includes all baseline sample students with the relevant data. Data is from 2002 Student Questionnaire, and from Busia District and Teso District Education Office records. The sample size is 7,401 questionnaires, 65% of the baseline sample in Busia and 60% in Teso (the remainder either had left school by the 2002 survey or were not present in school on the survey day).

Table 4: Program Test Score Impacts, Cohorts 1 and 2 Girls

	Dependent variable: Normalized test scores from 2001 and 2002			
Panel A: ITT sample	Busia and Teso		Busia	Teso
	(1)		(2)	(3)
Program school	0.19*		0.27*	0.09
	(0.11)		(0.16)	(0.14)
Sample size	3602		2106	1496
R^2	0.01		0.02	0.00
Mean of dependent variable	-0.06		-0.03	-0.12
Lee lower bound	0.16		0.27*	-0.17
	(0.11)		(0.16)	(0.14)
Lee upper bound	0.22**		0.27*	0.23*
	(0.11)		(0.16)	(0.13)
Panel B: Restricted sample	Busia and Teso		Busia	Teso
	(1)	(2)	(3)	(4)
Program school	0.18	0.15***	0.25***	0.01
	(0.12)	(0.06)	(0.08)	(0.08)
Mean school test score, 2000		0.76***	0.80***	0.69***
		(0.04)	(0.06)	(0.05)
Sample size	3420	3420	2061	1359
R^2	0.01	0.29	0.34	0.22
Mean of dependent variable	-0.06	-0.06	-0.03	-0.11
Lee lower bound	0.09	0.09	0.25***	-0.17**
	(0.11)	(0.05)	(0.08)	(0.07)
Lee upper bound	0.25**	0.21***	0.25***	0.17***
	(0.11)	(0.05)	(0.08)	(0.07)
Panel C: Longitudinal sample	Busia and Teso		Busia	Teso
	(1)	(2)	(3)	(4)
Program school	0.19	0.12	0.19	-0.01
	(0.14)	(0.09)	(0.12)	(0.10)
Individual test score, 2000		0.80***	0.83***	0.74***
		(0.04)	(0.05)	(0.05)
Sample size	1153	1153	768	385
R^2	0.01	0.62	0.65	0.58
Mean of dependent variable	-0.05	-0.05	-0.03	-0.09
Lee lower bound	-0.13	-0.03	0.08	-0.19**
	(0.11)	(0.07)	(0.10)	(0.09)
Lee upper bound	0.47***	0.25***	0.29***	0.16
	(0.12)	(0.10)	(0.12)	(0.11)

Notes: Significantly different than zero at 90% (*), 95% (**), 99% (***) confidence. OLS regressions, Huber robust standard errors in parentheses. Disturbance terms are allowed to be correlated across observations in the same school, but not across schools. District fixed effects are included in Panel A regression 1, and Panels B and C in regressions 1 and 2, and cohort fixed effects are included in all specifications. Test scores were normalized such that comparison group test scores had mean zero and standard deviation one.

Table 5: Program Test Score Impacts, Cohorts 1 and 2 Boys

	Dependent variable: Normalized test scores from 2001 and 2002		
Panel A: ITT sample	Busia and Teso	Busia	Teso
	(1)	(2)	(3)
Program school	0.08	0.10	0.04
	(0.13)	(0.20)	(0.14)
Sample size	4058	2331	1727
R^2	0.00	0.00	0.00
Mean of dependent variable	0.18	0.19	0.16
Lee lower bound	0.02	0.10	-0.25*
	(0.13)	(0.20)	(0.13)
Lee upper bound	0.12	0.10	0.19
	(0.13)	(0.20)	(0.13)

Panel B: Restricted sample	Busia and Teso		Busia	Teso
	(1)	(2)	(3)	(4)
Program school	0.05	0.07	0.15*	-0.03
	(0.14)	(0.07)	(0.09)	(0.09)
Mean school test score, 2000		0.77***	0.86***	0.65***
		(0.06)	(0.07)	(0.08)
Sample size	3838	3838	2256	1582
R^2	0.00	0.23	0.29	0.16
Mean of dependent variable	0.19	0.19	0.20	0.18
Lee lower bound	-0.09	-0.05	0.09	-0.25***
	(0.13)	(0.06)	(0.08)	(0.07)
Lee upper bound	0.17	0.17***	0.18**	0.18***
	(0.13)	(0.07)	(0.09)	(0.07)

Panel C: Longitudinal sample	Busia and Teso		Busia	Teso
	(1)	(2)	(3)	(4)
Program school	0.09	0.14**	0.24***	-0.03
	(0.14)	(0.06)	(0.08)	(0.09)
Individual test score, 2000		0.86***	0.91***	0.77***
		(0.02)	(0.02)	(0.03)
Sample size	1275	1275	851	424
R^2	0.00	0.71	0.75	0.63
Mean of dependent variable	0.24	0.24	0.23	0.27
Lee lower bound	-0.20	0.02	0.18**	-0.22**
	(0.13)	(0.06)	(0.07)	(0.09)
Lee upper bound	0.34***	0.23***	0.28***	0.13
	(0.13)	(0.07)	(0.08)	(0.10)

Notes: Significantly different than zero at 90% (*), 95% (**), 99% (***) confidence. OLS regressions, Huber robust standard errors in parentheses. Disturbance terms are allowed to be correlated across observations in the same school, but not across schools. District fixed effects are included in Panel A regression 1, and Panels B and C in regressions 1 and 2, and cohort fixed effects are included in all specifications. Test scores were normalized such that comparison group test scores had mean zero and standard deviation one.

Table 6: Program Test Score Quartile Effects,
Longitudinal Sample Cohort 1

	Dependent variable: Normalized test scores from 2001		
Panel A: Girls	Busia and Teso	Busia	Teso
	(1)	(2)	(3)
Top quartile * treatment	0.00	0.08	-0.15
	(0.13)	(0.16)	(0.27)
Second quartile * treatment	0.23***	0.29***	0.12
	(0.10)	(0.11)	(0.17)
Third quartile * treatment	0.13	0.19	0.01
	(0.09)	(0.13)	(0.10)
Bottom quartile * treatment	0.12	0.23	-0.10
	(0.20)	(0.30)	(0.14)
Sample size	1153	768	385
R^2	0.54	0.58	0.50
Mean of dependent variable	-0.05	-0.03	-0.09
Panel B: Boys	Busia and Teso	Busia	Teso
	(1)	(2)	(3)
Top quartile * treatment	-0.11	0.03	-0.38*
	(0.12)	(0.15)	(0.19)
Second quartile * treatment	0.18**	0.24**	0.06
	(0.09)	(0.11)	(0.15)
Third quartile * treatment	0.11	0.10	0.04
	(0.09)	(0.11)	(0.15)
Bottom quartile * treatment	0.18*	0.33***	-0.10
	(0.10)	(0.13)	(0.16)
Sample size	1275	851	424
R^2	0.63	0.68	0.56
Mean of dependent variable	0.24	0.23	0.27

Notes: Significantly different than zero at 90% (*), 95% (**), 99% (***) confidence. OLS regressions, Huber robust standard errors in parentheses. Disturbance terms are allowed to be correlated across observations in the same school, but not across schools. District fixed effects are included in Panel A and Panel B regression 1, and cohort fixed effects and quartile fixed effects are included in all specifications. Test scores were normalized such that comparison group test scores had mean zero and standard deviation one. Quartiles refer to scores in the pre-program 2000 test score distribution.

Table 7: Program Impacts on Teacher attendance in 2002 (Panel A) and School Participation Impacts in 2001 and 2002, Cohorts 1 and 2 (Panels B and C)

Panel A: Teacher attendance	Dependent variable: Teacher attendance in 2002		
	Busia and Teso	Busia	Teso
	(1)	(2)	(3)
Program school	0.048***	0.070***	0.016
	(0.020)	(0.024)	(0.035)
Mean school test score, 2000	0.040***	0.034**	0.033*
	(0.012)	(0.016)	(0.020)
Sample size	1065	652	413
R^2	0.02	0.04	0.01
Mean of dependent variable	0.84	0.86	0.83
Panel B: Girls' school participation	Dependent variable: Average Student School Participation		
	Busia and Teso	Busia	Teso
	(1)	(2)	(3)
Program school	0.006	0.032*	-0.029
	(0.015)	(0.018)	(0.023)
Mean school test score, 2000	0.028**	0.010	0.054***
	(0.013)	(0.015)	(0.016)
Sample size	3343	2033	1310
R^2	0.01	0.01	0.02
Mean of dependent variable	0.88	0.87	0.88
Panel C: Boys' school participation	Dependent variable: Average Student School Participation		
	Busia and Teso	Busia	Teso
	(1)	(2)	(3)
Program school	-0.009	0.006	-0.030
	(0.018)	(0.027)	(0.021)
Mean school test score, 2000	0.021	-0.002	0.050***
	(0.018)	(0.024)	(0.014)
Sample size	3757	2221	1536
R^2	0.00	0.00	0.02
Mean of dependent variable	0.85	0.85	0.85

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

The teacher attendance visits were unannounced, and actual teacher presence at school recorded during three unannounced school visits in 2002. The teacher attendance sample includes all senior and assistant classroom teachers, and excludes nursery school teachers and administrators, in all schools participating in the program.

The sample in Panels B and C includes students in schools that did not pull out of the program. 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 averaged in the Panels B and C regressions, so that each school year receives equal weight.

Table 8: Program Impact on Education Habits, Inputs, and Attitudes for Cohort 2 Restricted Sample in 2002

	---------- Busia and Teso Districts ----------			
	----------Girls----------		----------Boys----------	
Dependent Variables:	Estimated impact (s.e.)	Mean (s.d.) of dep. var.	Estimated impact (s.e.)	Mean (s.d.) of dep. Var.
Panel A: Attitudes towards education				
Student prefers school to other activities (index) [a]	0.02	0.72	0.01	0.72
	(0.01)	(0.18)	(0.01)	(0.18)
Student thinks s/he is a "good student"	0.02	0.73	0.03	0.73
	(0.04)	(0.44)	(0.03)	(0.44)
Student thinks being a "good student" means "working hard"	-0.02	0.69	0.03	0.63
	(0.03)	(0.46)	(0.03)	(0.48)
Student thinks can be in top three in the class	0.00	0.33	-0.03	0.40
	(0.04)	(0.47)	(0.03)	(0.49)
Panel B: Study/work habits				
Student went for extra coaching in last two days	-0.04	0.40	-0.02	0.42
	(0.04)	(0.49)	(0.05)	(0.49)
Student used a textbook at home in last week	0.01	0.85	0.04	0.80
	(0.03)	(0.36)	(0.03)	(0.40)
Student did homework in last two days	0.03	0.78	-0.01	0.73
	(0.04)	(0.41)	(0.04)	(0.45)
Teacher asked the student a question in class in last two days	0.03	0.81	0.02	0.82
	(0.04)	(0.39)	(0.03)	(0.38)
Amount of time did chores at home [b]	0.02	2.63	0.01	2.41
	(0.05)	(0.82)	(0.05)	(0.81)
Panel C: Educational inputs				
Number of textbooks at home	0.09	3.83	-0.15	3.61
	(0.19)	(2.15)	(0.15)	(2.19)
Number of new books bought in last term	0.15	1.54	-0.03	1.37
	(0.14)	(1.48)	(0.12)	(1.42)

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 parentheses. 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. Surveys were not collected in schools that dropped out of the program. The sample size varies from 700-850 observations, depending on the extent of missing data in the dependent variable.

[a] 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.

[b] 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).