Incentives to Learn

Michael Kremer[*] Edward Miguel[**] Rebecca Thornton[***]

January 2007

Abstract: In many education systems, those who perform well on exams at one level of education receive free or subsidized access to the next level of education. Merit scholarships could potentially increase effort, but skeptics argue they are inequitable, weaken intrinsic motivation, and are subject to gaming. 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, showed test score gains in program schools and we cannot reject the hypothesis that test gains were the same for these girls with low pre-test scores as for higher scoring girls. 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 was evidence of positive spillovers to boys, and student school 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. 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 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 could potentially induce greater student effort and that effort is an important input in educational production, potentially with positive externalities for other students.

However, 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 can 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] Some economists argue that the impact of incentives depends on the particular context and framing of the incentive (Akerlof and Kranton 2005, Fehr and Gächter 2002, Fehr and List 2004).

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 credit-hours students completed decreased, suggesting that students took fewer courses in order to keep up their grades. Similarly, after the HOPE college scholarship program was introduced in Georgia, the average SAT score for high school seniors rose almost 40 points

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

(Cornwell et al. 2002), but there was a 2% average reduction in completed college credits, a 12% decrease in full course-load completion, and 22% increase in summer school enrollment (Cornwell et al 2003). Beyond course-load selection, merit award incentives could potentially produce test cramming and even cheating rather than real learning.

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 and little evidence of the hypothesized negative effects of merit scholarship systems. Girls who were eligible for the scholarship in the 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 gains.

Our results also shed light on the empirical relevance of several leading objections to merit awards described above. First, equity concerns are alleviated by our finding of 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.

Second, surveys of students provide no evidence that program incentives weakened the intrinsic motivation to learn, nor that they led to gaming or cheating. The student merit scholarship program we study 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 remain large in the year following the competition, after incentives were removed. This, together with the evidence of broad test score improvements even among those program school girls with little chance of winning an award, suggests that the results reflect real learning and not just gaming.

Disaggregation of the results by district suggests that incentive effects may be context-specific. The merit scholarship program 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 district), test scores gains were large among both girls and boys and student attendance also increased significantly, suggesting another possible underlying 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.

The evidence from Busia district – where we have strong evidence the program worked – that boys experienced significant test score gains even though they were ineligible for the scholarship, together with the finding of test gains among low-scoring girls, suggests there are substantial 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 to other settings, the evidence for positive externalities within the classroom creates a new rationale for merit scholarships as well as for public education subsidies more broadly.

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, but their small sample size limits statistical precision, complicating inference. 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 suggestive evidence of gains for some girls (Angrist, Lang and

Oreopoulos 2006). As noted above, in the U.S. scholarships lead students to get better grades but to take less ambitious course loads (Binder, 2002, Cornwell et al 2002, Cornwell et al 2003)

Ashworth et al. (2001) use propensity score matching to estimate that weekly allowances given to 16-19 year old students from low-income U.K. households based on school enrollment and academic achievement raised school enrollment among eligible youth by 6 percentage points and by 4 percentage points among the ineligible, suggesting externalities. 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 their maintaining a satisfactory level of academic performance led to academic gains. They note that the impact of these vouchers may not have been due only to expanding the range of school choice available to participants, 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 for good matriculation exam performance in 20 Israeli secondary schools. Students offered the merit award were 6-8 percentage points more likely to pass exams than comparison students in a pilot program that randomized awards among schools. 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 Israel study in several important ways. First, by examining a program in which scholarships were randomized at the school level, we are able to estimate externality impacts. Second, due to political and logistical issues, the program in Israel and its evaluation, which was meant to run for three years, were discontinued after the first year, so post-competition program impact estimates are unavailable. Third, our sample includes more than three times as many schools, and our program and comparison groups are more balanced on observable characteristics. Finally, in addition to test scores, we collected data on student school attendance,

teacher attendance, and student attitudes and time use which allow us to explore potential mechanisms and evaluate objections to merit scholarships.

The paper proceeds as follows: section 2 describes schooling in Kenya and the Girls Scholarship Program, section 3 discusses the data and sample, section 4 presents the empirical results and evaluates the main objections to merit scholarships, and the final section concludes.

2. The Girls Scholarship Program

2.1 Schooling in Kenya

Schooling in Kenya consists of eight years of primary school followed by four years of secondary school. While approximately 85% of children of primary school age in western Kenya are enrolled in school (Central Bureau of Statistics 1999), there are high dropout rates in grades 5, 6, and 7 and only about one-third finish primary school. Dropout rates are especially high for girls in these grades.[2]

Secondary school admission depends on performance on the grade 8 government Kenya Certificate of Primary Education (KCPE) exam. To prepare, students in grades 4-8 take exams at the end of each school year. These exams are standardized for each district and test students in English, geography/history, mathematics, science, and Swahili. Students must pay a fee to take the exam, US$1-2 depending on the year. 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 either supervision or grading. Exam monitors document and punish any instances of cheating, and report these cases to the district office.

When the scholarship program we study was introduced, primary schools charged school fees to cover their non-teacher costs, including textbooks for teachers, chalk, and classroom maintenance.

[2] 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.

These fees averaged approximately US$6.40 (KSh 500)[3] per family each year. In practice, while these fees set a benchmark for bargaining between parents and headmasters, most parents did not pay the full fee. In addition to this fee, there were also fees for activities, such as taking exams, as well as costs of school supplies, certain textbooks, and uniforms. The scholarship project we study was introduced in part to assist families of high-achieving girls to cover these costs.

In late 2001, then-president Daniel Arap Moi announced a national ban on primary school fees, but the central government did not provide alternative sources of school funding and other policymakers made unclear statements on whether schools could impose "voluntary" fees. Schools varied in the extent to which they continued collecting fees in 2002, but this is difficult to quantitatively assess. 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 using a World Bank loan. Our study focuses on program impacts in 2001 and 2002 before primary school fees were eliminated by the Kibaki reform.

2.2 Project Description and Timeline

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

There were 127 sample primary schools, half of which were invited to participate in the scholarship program in March 2001. The randomization first stratified schools by Busia and Teso districts, and by administrative divisions within the districts[4], and also stratified them by participation

[3] One US dollar was worth 78.5 Kenyan shillings (KSh) in January 2002 (Central Bank of Kenya 2002).
[4] Divisions are subsets of districts, with eight divisions in all within our sample.

in a past NGO assistance program which provided classroom flip charts.[5] Randomization into program and comparison groups was then carried out within each stratum using a computer random number generator.

The NGO awarded scholarships to the highest scoring 15% of grade 6 girls in the program schools within each district (amounting to 110 girls in Busia and 90 in Teso). Each district (Busia and Teso) had separate tests as well as separate competitions for the merit award. Scholarship winners from grade 6 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, and 57% of program schools (36 of 63 schools) had at least one 2001 winner; among schools with at least one winner, there was an average of 5.6 winners per school.

The scholarship program provided winning grade 6 girls with an award for the next two academic years, grades 7 and 8 (through the end of primary school). In each year, the award consisted of: (1) a grant of US$6.40 (KSh 500) 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. Although the program did not include explicit monitoring to make sure that parents spent the grant on school supplies, the public presentation in a school assembly likely generated some community pressure to use the money in ways that benefited their daughter's education.[6] Since many parents would not otherwise have fully paid school fees, schools with winners benefited to some degree from the award money paid directly to the school. Some of these

[5] All GSP schools had previously participated in an evaluation of a flip chart program, and are a subset of that sample. These schools are representative of local primary schools along most dimensions but exclude some of the most advantaged as well as some of the worst off – see Glewwe et al. (2004) for details on the sample. The flip chart program did not affect any measures of educational performance (not shown). Stratification means there are balanced numbers of flipchart and non-flipchart schools across the GSP program and comparison groups.

[6] 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 purchases were made from the scholarship money on school supplies such as math kits, notebooks, and pencils.

funds may have also benefited teachers if they were used to improve the staff room, for instance, although the amounts dedicated to this were likely small.

Two cohorts of grade 6 girls competed for scholarships. Girls registered for grade 6 in January 2001 in program schools were the first eligible cohort (cohort 1) and those registered for grade 5 in January 2001 made up the second cohort (cohort 2), competing in 2002. Cohort 1 students took the usual end-of-year grade 5 exams in November 2000, and these are used as baseline test scores in the analysis.[7] 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 (not shown).

In March 2001, NGO staff met with school headmasters to give each school community the choice of whether or not to participate. Headmasters were asked to relay information about the program to parents via a school assembly and in September and October the NGO held additional community meetings to reinforce knowledge about program rules in advance of the November 2001 district exams. After these meetings, NGO enumerators began collecting school attendance data during unannounced visits.

The baseline 2000 test score is a very strong predictor of being a top 15% performer on the 2001 test in both program and comparison schools, as expected. Students below the median baseline test score having almost no chance of winning the scholarship (Figure 1). In particular, the odds of winning are only 1% for the bottom quartile of girls in the baseline test distribution and 3% for the second quartile, compared to 23% and 33% in the top quartiles in Busia and Teso, respectively.

Children whose parents have more schooling were also more likely to be in the top 15% of test performers: average years of parent education are nearly three years greater for scholarship

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

winners than losers (7.7 years for winners versus 4.8 years for non-winners), and this large effect is statistically significant at 99% confidence. Note that this link between parent education and the child test score is no stronger in the program schools than in the comparison schools. However, there is no statistically significant difference between winners and non-winners in terms of household ownership of iron roofs or latrines (regressions not shown), suggesting a weaker correlation between household wealth and winning a scholarship.

Official exams were again held in late 2002 in Busia district. The government cancelled the 2002 exams in Teso district because of concern about possible disruptions in the run-up to the December 2002 national elections, so the NGO instead administered its own standardized exams in February 2003, after the election. Thus the second cohort of winners were chosen in Busia based on the official 2002 district exam, while Teso winners were chosen based on the NGO exam. In this second round, 70% of program schools (44 of 63 schools) had at least one winner, an increase over 2001, and in all 78% of program schools had at least one winner in either 2001 or 2002.

The NGO again visited all schools during 2002 to conduct unannounced attendance checks and administer questionnaires to grade 5-7 students, collecting information on study effort, habits, and attitudes toward school. This student survey indicates that most girls understood program rules, with roughly ninety 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 15 percentage points more likely than boys to know that "only girls are eligible for the scholarship" (91% for girls versus 76% for boys), although the proportion among boys is still high, indicating that the vast majority of boys knew that they were ineligible.[8] Girls were very likely (69%)

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

to report that their parents had mentioned the program to them, suggesting some parental encouragement.

3. Data and Sampling Issues

In this section we first provide information about the dataset used in this paper. We next discuss program implementation in Busia and Teso districts, and in particular examine the implications of sample attrition related to the scholarship program in Teso district. Finally, we compare characteristics of the 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 Busia district and Teso 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.

School participation data are based on four unannounced checks collected by NGO enumerators, one conducted in September or October 2001 and one in each of the three terms of the 2002 academic year. School participation rates are approximately 85% in the main sample (Table 1, Panel C). We use the unannounced check data rather than official school attendance registers, since registers are often unreliable. Finally, 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.

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 were educationally disadvantaged relative to Luhyas: Teso district parents in the data set have 0.4 years less schooling than Busia

parents on average. There is also a tradition of suspicion of outsiders in Teso district, and this has at times led to misunderstandings between NGOs and some people 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 a Teso district primary school, severely damaging the school, killing seven students, and injuring 27 others. Although the school struck by lightning was not in the sample for the scholarship program, the NGO had been involved with another assistance program in that school, and there were other strange coincidences – for instance, the names of several victims were the same as NGO staff members who had recently visited the school. Some members of the community associated the lightning strike deaths with the NGO, and the incident led some schools to pull out of the girl's scholarship program: of the original 58 sample schools in Teso district, five pulled out immediately following the lightning strike, and one school located in Busia, but near the Teso district border (and with a substantial ethnic Teso population) also pulled out shortly thereafter.[9] Figure 2 presents the location of the lightning strike and the schools that pulled out, four of which are located near the strike. Three of the six schools that pulled out were treatment schools and three were comparison schools. The schools pulled out prior to administration of the follow-up test, and we decided at that time that we would supplement the analysis of the sample as a whole with an analysis disaggregated by district to deal with this attrition issue and obtain a sample with minimal school-level attrition (namely, the Busia district schools).

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, possibly due to the lightning strike, pre-existing differences, or some combination of

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

these and other factors. When teachers were asked to rate local parental support for the program, 90% of the Busia teachers claimed that parents were either "very positive" or "somewhat positive" but the analogous rate in Teso was only 58%, and this difference across the districts is statistically significant at 99% confidence.

Thus, although the monetary value of the award was identical everywhere, local social prestige associated with winning may have differed sharply between Busia and Teso. It remains possible that attitudes towards the program in Teso district would have become more positive over time in the context of a permanent program, because people would learn that their fears associated with program were incorrect, but that remains speculative.

3.3 Sample Attrition

Recall that six schools pulled out of the program after the lightning strike. In addition, there were five other schools, three in Teso district and two in Busia, with incomplete exam scores for 2000, 2001 or 2002. Not surprisingly, given the reported differences in the response to the scholarship program, we also find differences in sample attrition patterns across Busia and Teso districts. In particular, we find large differences in attrition across program and comparison schools in Teso district, but not in Busia district. In Busia differences between program and comparison schools are small and not statistically significant: for cohort 1, 79% of girls (76% of boys) in Busia program schools and 78% of girls (77% of boys) in Busia comparison schools took the 2001 exam (Table 2). Among cohort 2 students in Busia, there is again almost no difference between the program school students and comparison school students in the proportion who take the 2002 exam (50% versus 48% for girls, and 50% versus 52% for boys). For both program and comparison students, there is more attrition by 2002 as students drop out of school, transfer to other schools, or decide not to take the exam.

Attrition patterns in Teso district schools are strikingly different: for cohort 1, 53% of girls in program schools (54% of boys) took the 2001 exam, but the rate for comparison school girls is much

higher, at 65% (and similarly high for boys, at 66%, Table 2).[10] There are also attrition gaps across program and comparison schools among cohort 2, although these are smaller than for cohort 1.[11]

Non-parametric Fan locally weighted regressions, with bootstrapped standard errors clustered by school, display the proportion of cohort 1 students taking the 2001 exam as a function of their baseline 2000 test score. These indicate that Busia district students across all levels of initial academic ability have a similar likelihood of taking the 2001 exam and remaining in the sample (Figure 3, Panels A and B). Although, theoretically, the introduction of a scholarship could have induced poor but high-achieving students to take the exam in program schools, leading to an upward bias in estimated program impacts, we do not find evidence for such a change in exam participation patterns in either Busia or Teso district. Rather, students with low initial achievement are somewhat more likely to take the 2001 exam in Busia program schools relative to comparison schools, and this difference is statistically significant in the 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 these figures suggest any bias is likely to be small.

In contrast, not only were attrition rates high and unbalanced across treatment groups for cohort 1 in Teso, but significantly more high-achieving students took the 2001 exam in comparison schools relative to program schools, and this is likely to strongly bias estimated program impacts toward zero in Teso (Figure 4, Panels A and B). Among high ability girls in Teso with a score of at least +1 standard deviation on the baseline 2000 exam, comparison school students were over 20 percentage points more likely to take the 2001 exam than program school students, and this difference is statistically significant at 95% confidence. The comparable gap among high ability

[10] Table 2 excludes the six schools that pulled out of the program completely. The differential attrition patterns across program and comparison school are even more pronounced when they are included (not shown).
[11] 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).

Busia girls is near zero and not statistically significant. There are similar though less pronounced gaps between comparison and program schools for Teso district boys (Panel B). Pooling boys and girls, in Teso program schools students who did not take the 2001 exam scored 0.05 standard deviations lower on average at baseline (on the 2000 test) than those who took the 2001 exams, but the difference is 0.57 standard deviations in the Teso comparison schools, and the estimated difference-in-differences is significant at 95% confidence (regression not shown). These attrition patterns in Teso may be due to some high-achieving pupils in program schools feeling especially "vulnerable" to the program in communities where there was mistrust of the NGO and fear of the program. Several schools that initially pulled out of the program also had relatively high average baseline 2000 test scores.

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 (Figure 1). The proportion of cohort 1 program school girls taking the 2001 exam as a function of the baseline score (Figure 3 Panel A and Figure 4 Panel A) does not correspond closely to the likelihood of winning an award in either district. 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.

To summarize, Teso district primary schools had higher rates of sample attrition than Busia schools in 2001 and the gap in attrition between program and comparison schools was large in Teso district but zero in Busia. In addition, a much higher proportion of high ability students (according to baseline exam scores) took the exam in Teso district comparison schools than in Teso program schools, likely biasing program impact estimates toward zero. In the analysis below, we first show estimated pooled effects in Busia and Teso districts together. We then examine the effects of the two scholarship programs separately in each of the two districts because of the pronounced differences in attrition rates and in the community response to the scholarship program.

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, number of siblings, proportion of ethnic Luhyas, or the ownership of a latrine, iron roof, or mosquito net across Busia district program and comparison schools (Table 3, Panel A). Most of the characteristics we examine in the 2002 pupil survey data were stable between 2001 and 2002 (e.g. parent education). Household characteristics are also broadly similar across program and comparison schools in the Teso main sample, but there are certain differences, including larger household size and lower likelihood of owning an iron roof among program students (Table 3, Panel B). This may in part be due to the different attrition across Teso program and comparison schools discussed above.

The 2000 baseline test score distributions provide further evidence on the comparability of the program and comparison groups. Formally, in Busia district we cannot reject the hypothesis that mean baseline test scores are the same across program and comparison schools for either girls or boys. The two distributions are similar graphically (Figure 5), and we cannot reject equality of the distributions using the Kolmogorov-Smirnov Test (p-value = 0.35 for cohort 1 Busia girls). In Teso, where several schools dropped out of the sample, we do statistically reject the hypothesis of equality between program and comparison baseline test scores distributions (p-value = 0.08 for cohort 1 Teso girls). We discuss the implications of this difference in Teso in our empirical results below.

4. Empirical Strategy and Results

In this section we first describe our estimation strategy, then present the main test score results. We next evaluate the leading objections to merit scholarships made by educators, and finally discuss the cost-effectiveness of the program in boosting academic test scores.

4.1 Estimation Strategy

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:

$$(1) \qquad TEST_{ist} = \alpha + \beta_1 TREAT_s + \beta_2 (TREAT_s * MALE_{is}) + \gamma_1 MALE_{is} + X_{ist}' \gamma_2 + \mu_s + \varepsilon_{ist}$$

$TEST_{ist}$ is the normalized test score for student i in school s in year t.[12] $TREAT_s$ is the program school indicator, and $MALE_{is}$ is an indicator variable for male students. The coefficient β_1 captures the average program impact on females, the population targeted for program incentives, and the β_2 term any treatment effect differences by gender. X_{ist} is a vector that includes the average school baseline (2000) test score when we use the main 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 school random effect perhaps capturing common local or headmaster characteristics, and ε_{ist}, which captures unobserved student ability or idiosyncratic shocks.

4.2 Academic Test Score Impacts

In the analysis, we focus on the *main sample* which consists of students in schools that did not pull out of the program and for which we have mean school baseline 2000 test scores (Table 1, Panel C), consisting of 116 schools and 7,258 students. The main sample contains data for 91% of the schools in the *baseline sample*, of whom 51% are program school students.[13] The *longitudinal sample* contains the main sample cohort 1 students who also have individual 2000 baseline test scores. We

[12] Test scores were normalized separately by district, as the exams offered were different in Busia and Teso.

[13] Average test scores are slightly higher in the main sample than in the baseline sample, since the students dropped from the sample are typically somewhat below average in academic achievement (discussed above). Also note 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 for cohort 2 we focus on the 2002 exam alone.

first present estimated program effects among the longitudinal sample in the first year of the program and then move on to the main sample results and robustness checks.

Cohort 1 Longitudinal Sample: Pooled results for Busia and Teso districts

The program raised test scores by 0.12 standard deviations on average among girls and boys in 2001 and 2002, among all Busia and Teso district longitudinal sample students (Table 4, Panel A, regression 1). The average impact rises slightly to 0.13 standard deviations (standard error 0.06, regression 2) and becomes statistically significant at 95% confidence when the individual baseline 2000 test score is included as an explanatory variable, as this baseline control reduces residual variation. The 2000 test score is strongly related to the 2001 test score as expected (point estimate 0.80, standard error 0.02). Estimated program impacts are nearly identical for girls and boys (regression 3), a first indication of positive program externalities for boys, who were ineligible for awards. The overall program impact on test scores, pooling Busia and Teso districts, is thus positive and statistically significant, and moderate in magnitude.

These are relatively large impacts: to illustrate with previous findings from Kenya, the average test score for grade 7 students who take a grade 6 exam is approximately one standard deviation higher than the average score for grade 6 students (Glewwe et al. 1997). Thus the estimated average program effect corresponds roughly to an additional 0.13 grades worth of average primary school learning. These effects are similar to the estimated effect of reducing class size by ten students in Israeli schools (Angrist and Lavy 1999).

We next test whether test score effects differ for students with different baseline academic performance, using the cohort 1 longitudinal sample of girls (who have pre-program 2000 test data). In no case can we reject the hypothesis that treatment effects are the same throughout the baseline distribution. There is no evidence of larger effects at the top of the distribution in a regression specification interacting the baseline test score with the treatment indicator (coefficient estimate

0.034, standard error 0.063, regression not shown). When indicators for quartiles of the baseline exam distribution are interacted with the treatment indicator variable, the average treatment effects in the baseline test quartiles (from bottom to top) are all positive at 0.11, 0.07, 0.23, and 0.09 standard deviations, respectively (regression not shown), and we cannot reject the hypothesis that treatment effects in all quartiles are equal (F-test p-value = 0.23). Although estimating the program effect separately for each quartile reduces statistical power somewhat, the positive and large estimated test score gains for girls with little to no chance of winning the award is further evidence of positive externalities of the merit award program within the classroom.

Cohort 1 Longitudinal Sample: Results by district

Disaggregation by district 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 0.08 (Table 4, Panel A, regression 4). While the baseline 2000 test score distributions are similar across program and comparison schools for both Busia girls and boys (Figure 5), the test score distribution in program schools shifts markedly to the right for cohort 1 girls in the first year of the program (Figure 6, Panel A).[14] The vertical lines in each figure indicate the minimum score necessary to win an award each year. The same cohort 1 longitudinal sample – namely, those main sample students who have 2000 individual test scores – is used in both Figures 5 and 6. Test gains are not as large for boys, but there are perceptible shifts in both the left and right tails of the program school distributions (Figure 6, Panel B). The evidence on program gains throughout the baseline test score distribution is presented using an alternative, and perhaps more transparent, non-parametric approach in Figure 7, including 95% confidence bands on the treatment effects.

The estimated program impact for Teso district is near zero at -0.02 standard deviations (Table 4, Panel A, regression 5), but it is difficult to reject a wide variety of hypotheses regarding

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

effects in Teso due to attrition. To get a handle on the possible bias due to sample attrition in Teso (discussed above), we construct non-parametric bounds on program effects using the trimming method in Lee (2002). While the bounds for Busia schools are tight since there was essentially no differential attrition across program and comparison groups there (Table 2), the bounds for cohort 1 girls in Teso district are wide, ranging from -0.24 standard deviations as a lower bound up to 0.22 standard deviations as an upper bound in a specification analogous to regression 5 (not shown). Using this conservative bounding method, it is not possible to draw firm conclusions about program treatment effects in Teso district in the presence of observed sample attrition.

In addition to constructing Lee (2002) bounds, we also carried out two additional tests to characterize program impacts in Teso in the presence of sample attrition. We first imputed missing 2001 test scores among longitudinal sample students as a linear function of their 2000 score. This suggests that, in the absence of attrition, program impacts for cohort 1 girls in Teso would have been positive and reasonably large: the estimated impact for Teso district girls using this imputation becomes 0.12 standard deviations (standard error 0.14 – regression not shown). However, this estimate is of course only suggestive given possible omitted variables correlated with attrition. A final approach for addressing sample attrition bias in Teso district is to focus on impacts for cohort 2 girls alone, since attrition rates are similar for them in both program and comparison schools (Table 2), although we are unable to determine the exact nature of attrition due to their lack of baseline 2000 data. The estimated impact is near zero (estimate 0.00 standard deviations, standard error 0.11 – regression not shown). Whatever interpretation is given to the Teso district results – either unreliable estimates due to attrition or simply no program 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.

Lastly, program impact point estimates increase slightly in both districts when the individual baseline test control is not included as an explanatory variable (in a specification analogous to Table 4, Panel A, regression 1, but run separately by district) but standard errors increase: the estimate for

Busia is 0.22 standard deviations (standard error 0.19) and for Teso becomes positive at 0.08 standard deviations (standard error 0.15 – regressions not shown).

Cohorts 1 and 2 Main Sample

We find similar results for cohort 1 and cohort 2 girls and boys in the main sample: there is an overall impact of 0.10 standard deviations (Table 4, Panel B, regression 1) which becomes statistically significant at 90% confidence when the mean school 2000 test score (computed among students in the main sample) is included as an explanatory variable. The average program effect for girls is large and statistically significant in the main sample at 0.14 standard deviations (standard error 0.06, regression 3), while the average effect for boys falls to 0.07 – the most noteworthy difference between the results for the longitudinal (Table 4, Panel A) versus the main sample (Table 4, Panel B), although we still cannot reject equal effects for girls and boys in the main sample.

The average program impact for Busia district girls in the main sample is 0.25 standard deviations (standard error 0.07, statistically significant at 99% confidence – Table 4, Panel B, regression 4)[15], again much larger than the estimated effect for Teso girls, at -0.02 standard deviations (regression 5). The estimated effect for Busia boys is reasonably large though only marginally statistically significant at 0.13 standard deviations (standard error 0.07, significant at 90% confidence), while the analogous effect for Teso boys is near zero.

The program effect among Busia and Teso cohort 2 students in the year of the competition was similar to that of Busia and Teso cohort 1 students when they competed. The program effect for cohort 1 girls in 2001 is 0.18 standard deviations (standard error 0.08, statistically significant at 95% confidence, Table 5, regression 1), and the effect for cohort 2 in 2002, when they competed for the

[15] Among Busia main sample girls, impacts are somewhat larger for mathematics, science, and geography / history than English and Swahili, but differences by subject are not statistically significant (regression not shown).

award, is 0.13 (standard error 0.07, significant at 90% confidence). Figure 8 presents test scores for Busia cohort 2 main sample students when they were competing for the scholarship.

Robustness checks

Point estimates are broadly unchanged in an intention to treat (ITT) analysis using the full baseline sample (using data for all 127 of the original sample schools). The point estimate for the pooled Busia and Teso sample is once again an average test score gain of 0.12 standard deviations. The average program impact for Busia girls is even larger at 0.27 standard deviations (standard error 0.17, not shown), and is 0.06 standard deviations (standard error 0.15, not shown) for Teso girls. These regressions do not include the mean school 2000 test control as an explanatory variable, however, since that data is missing for several schools, and thus standard errors are considerably larger in these specifications. Thus the ITT analysis leads to somewhat more positive average estimated program impacts in both Busia and Teso districts, consistent with the hypothesized downward sample attrition bias discussed above (in Section 3.3), but standard errors are larger.

Estimates are unchanged 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.[16] 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 socioeconomic status households.[17]

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

[16] 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).
[17] Note that although the program had similar test score impacts across socioeconomic backgrounds, students with more educated parents nonetheless were disproportionately likely to win because they have higher baseline scores.

various measure of baseline school quality, including the mean 2000 test score as well as the proportion of 6th grade girls in 2000 that were among the top 15% in their district. Neither of these interaction effects are statistically significant at traditional confidence levels (not shown), so we cannot reject the hypothesis that average program effects were the same across schools at various academic quality levels.

4.3 Evaluating Common Objections to Merit Scholarships

The results in section 4.2 indicate there were large test score gains in scholarship program schools in the year students competed for the award. In this subsection, we also provide evidence evaluating three leading objections to merit awards. First, we examine whether providing an external incentive resulted in any detectable reduction in intrinsic motivation. Second we examine whether the merit award resulted in "gaming" the test instead of actual school effort and learning. Lastly, we discuss some equity concerns.

Objection 1: Do external incentives reduce intrinsic motivation?

One critique of merit-based awards is that providing external incentives reduces intrinsic motivation, and that these adverse effects could persist after the incentive is removed. We first examine test score impacts after incentives are removed to assess the magnitude of any such effects, and instead find evidence of positive medium-run program impacts on human capital. Using self-reported attitudes and behaviors from surveys, we then find no evidence of detrimental effects on several measures of intrinsic motivation.

Post-competition test score effects

In the main sample, the program not only raised test scores for cohort 1 girls when it was first introduced in 2001 but also appears to continue boosting their scores in 2002: the estimated program

impact for cohort 1 girls in 2002 is 0.125 standard deviations (standard error 0.078, p-value = 0.11, Table 5, regression 1). This is evidence the program had lasting effects on learning, rather than simply being due to cramming for the competition exam or cheating. When we focus on Busia district alone, there is even stronger evidence of persistent test score impacts for cohort 1 girls in 2002, with a coefficient estimate of 0.25 standard deviations (standard error 0.09, significant at 99% confidence, regression 2).

The NGO exams administered in February 2003 provide further evidence on post-competition impacts. Although originally administered because 2002 exams were cancelled in Teso district, they were also offered in the Busia sample schools. In the standard specification (as in Table 5) the average program impact for cohort 1 Busia girls in early 2003 was 0.19 standard deviations (standard error 0.07, statistically significant at 99% confidence), and the gain for cohort 2 Busia girls is positive and marginally statistically significant at 0.15 standard deviations (standard error 0.08 – regression not shown). Though program effects fall somewhat for Busia girls in cohort 1 over time – from 0.28 standard deviations in the year of the competition (2001), to 0.25 standard deviations the following year (Table 5, regression 2), and 0.19 at the start of the second year after the competition – program impacts are quite persistent, and we cannot reject the hypothesis that effects in 2001, the competition year, are equal to the 2002 and 2003 post-competition effects (p-values 0.96 and 0.38, respectively).[18]

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

[18] Estimated treatment effects for boys are positive in all years and cohorts, for both the pooled Busia and Teso district sample (Table 5, regression 3) and Busia alone (regression 4), but only for Busia boys in 2001 are effects statistically significant.

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 72% of the time. There are no statistically significant differences in this index across the program and comparison schools for either girls or boys (Table 6, Panel A), and thus there is no evidence that external incentives dampened intrinsic motivation to learn in the pooled sample (or in Busia and Teso separately, not shown) as captured by this measure.

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. Although some Teso schools pulled out of the program, there were apparently no adverse effects on pupil attitudes in remaining schools as reported in surveys.

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 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 6, 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 any increased study effort may have come out of children's leisure time or through intensified effort during school hours, as discussed below.

Objection 2: Do external incentives lead to "gaming" of the exam?

Merit scholarship opponents fear that program test score gains could just reflect gaming, cramming or cheating, and not actual study effort and learning. It is useful to understand the channels underlying test score gains since some mechanisms, such as increased test coaching or cramming, might temporarily raise test scores without really improving retained learning, while others – such as

increased teacher and pupil effort – have more positive interpretations. Below we find evidence for positive program impacts on several objective measures of teacher and student effort, most importantly in terms of school attendance, and no evidence of gaming the system.

Teacher school attendance

The estimated program impact on teacher school attendance overall, in the pooled Busia and Teso sample, is large and statistically significant at 5.1 percentage points (standard error 2.1 percentage points, Table 7, panel A, regression 1).[19] Together with the test score impacts above, teacher attendance is the second educational outcome for which there are large, positive, and robustly statistically significant impacts in the pooled Busia and Teso district sample.

In our data, it is difficult to distinguish between teacher attendance in 6th grade classes versus other grades, since the same teacher often teaches a subject (i.e., mathematics) in several different grades and the data were recorded on a teacher by teacher basis rather than by grade and subject. Data from another sample of primary schools in Busia and Teso, though, reveals that 62.9% of all teachers teach at least one 6th grade class. If all attendance gains were concentrated among this subset of teachers, the implied program effect for teachers that teach at least one 6th grade class would be a 5.1 / 0.629 = 8.1 percentage point increase in attendance.

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 6.7 percentage points (standard error 2.5, significant at 99% confidence, Table 7, Panel A, regression 2), reducing overall teacher absenteeism by approximately one half. The implied effect among those teaching 6th grade if attendance gains were concentrated in this group is 10.7 percentage points. Note that the mean school baseline 2000 test score is positively but only moderately correlated with teacher attendance and all

[19] 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 3.5 percentage points (standard error 1.6, not shown).

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 (3.2 percentage points, regression 3).

There are a number of different mechanisms that might increase teacher effort in response to the merit scholarship program, including ego rents, social prestige in the community, and even gifts from winners' parents. While we cannot rule out those mechanisms, we have anecdotal evidence that increased parental monitoring may have also played a role. The June 2003 teacher interviews suggest greater parental monitoring occurred in Busia but not 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 district primary 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 (Table 6, Panel B). 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 (regression not shown), indicating that teachers probably did not substantially divert classroom effort towards girls in program schools. This finding, together with the increased teacher attendance, provides a plausible explanation for any positive educational spillovers for boys, namely, greater teaching effort directed to the class as a whole.

Student school attendance

In addition to teacher attendance gains, we find suggestive evidence of student attendance gains. The estimated scholarship program impact on school participation in the pooled Busia and Teso sample is positive at 2.1 percentage points (standard error 1.7, Table 7, Panel B, regression 1). There are again sharp differences by district. The program significantly increased participation as measured during unannounced enumerator visits in 2001 and 2002 in Busia district: for cohort 1 and cohort 2 in the main sample, the program increased school participation by 4.7 percentage points (standard error 2.5, regression 2), corresponding to an approximately 30% reduction in absenteeism. Average gains are slightly larger among Busia girls, at 5.0 percentage points (standard error 2.4, significant at 95% confidence, regression 3). Since school participation information was collected for all students, even those who did not take the 2001 or 2002 exams, these estimates are potentially less subject to sample attrition bias than test scores. School participation impacts are near zero and not statistically significant in Teso district (regressions 6 and 7).[20]

The scholarship program increased average school participation by 6.2 percentage points (standard error 4.2, Table 7, Panel B, regression 4) among Busia cohort 1 girls in 2001, and by an even larger 9.4 percentage points (standard error 4.9) among cohort 2 in 2001 in a pre-competition effect. School participation gains for Busia girls competing in 2002 are also positive but surprisingly small (regression 4), though we cannot reject that effects are the same in both years.[21] 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 immediate run-up to

[20] In the Busia comparison sample, girls with higher baseline test scores have significantly higher average school participation: students above the mean 2000 score were 10 percentage points more likely to be present in school during 2001 than those with below mean scores (standard error 0.01, regression not shown). This 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.

[21] Although the point estimate goes in the expected direction, there is no significant program effect on reducing school drop-outs in 2001 and 2002 (regression not shown).

exams (term 3) due to cramming, for instance. School participation gains are similar for Busia girls and boys (regressions 4 and 5). We again cannot reject the hypothesis that school participation gains among cohort 1 girls are equal across baseline 2000 test score quartiles (regression not shown).[22]

Parental investments in education

We also find some 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 district 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 6, 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).[23]

Cheating, cramming, and Hawthorne effects

Another concern related to the interpretation of our findings is the possibility of cheating on the exams, but this appears unlikely. First, exams in Kenya are always administered by outside monitors,

[22] The observed increase in school participation for low baseline test score girls also allows us to place an upper bound on their expected returns to increased study effort. This bound is low, indicating that any increase in their effort (as proxied by school participation) is also unlikely to be due to an attempt to win the award. The probability a program school girl obtained a test score high enough to win an award is a function of her baseline 2000 score. For a girl with a given baseline score, an upper bound on the effectiveness of greater effort in increasing the odds of winning is the probability a girl with that baseline score wins (since the probability of winning cannot fall below zero even at zero effort). Empirically, this upper bound is approximately one percentage point for Busia girls with baseline scores less than the mean of zero (Figure 1 Panel A). Since the scholarship is worth US$38, this means that their expected gain from effort is at most US$0.38 (ignoring any non-monetary utility benefits for winners). Girls show a 5.0 percentage point average gain in school participation (Table 7), and this translates into roughly 5.0% x 180 school days per year = 9 additional school days. Thus girls competing for the scholarship would only choose to attend school these additional nine days if their productivity at non-school activities was less than US$0.38 / 9 days ≈ US$0.04 per day. This is an implausibly low wage even for teenage girls in rural Kenya, providing further suggestive evidence in favor of externality benefits for low performing girls in program schools.

[23] 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 6 percentage points (significant at 95% 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.

not teachers from the school, and district records from the outside exam monitors indicate there were no documented instances of cheating in any sample school during either 2001 or 2002. Several findings reported above also argue against cheating: test score gains among cohort 1 students in scholarship schools persisted a full year (and more) after the exam competition when there was no longer any direct incentive to cheat, and there were substantial gains among program school boys ineligible for the scholarship who would experience no direct program benefits from cheating (although cheating by teachers could still potentially explain that pattern).

In terms of test "cramming", there is no evidence that extra test preparation coaching increased in the program schools for either girls or boys (Table 6, Panel B).[24] A separate teacher incentive project 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.

A final 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 section 4.4 below) but in no other case did a program generate such substantial test score gains.

[24] 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 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.

Objection 3: Do external incentives help only those with a chance of winning?

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 the scholarship awards flow mainly to 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 and we cannot reject the hypothesis of equal gains across all baseline test quartiles. Similarly, there are no significant program interaction effects with household socioeconomic measures, including parent education, indicating that even girls with poorly educated parents also gained significantly from the program in terms of test performance (section 4.2).

The focus so far has been program impacts on the mean of the test score distribution, but program impacts on inequality per se are also important in both theoretical and policy debates over merit scholarships. Perhaps not surprisingly, given the observed gains throughout the baseline test score distribution, there was only a small overall increase in test score variance for program school girls relative to comparison girls in the main 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 district program school girls, while the analogous variances for Busia comparison girls are 0.92 in 2000, 0.90 in 2001 and 0.92 in 2002; however the difference across the two groups is not statistically significant at traditional confidence levels in any year.[25] 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

[25] 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.

statistically significant (not shown)[26]. The bottom line is that there is no detectable increase in academic test score inequality as a result of the program, and the program improved disadvantaged students' academic performance.

4.4 Program Cost-Effectiveness

Appendix A compares the cost-effectiveness of six programs that have recently been conducted in the Kenyan study area: the girls' merit scholarship program that is the focus of this paper, a teacher incentive program (Glewwe et al. 2003), a textbook provision program (Glewwe et al. 1997), flip chart program (Glewwe et al. 2004), a deworming program (Miguel and Kremer 2004), and a child sponsorship program that provided a range of inputs, (Kremer et al. 2003). We conclude that providing merit scholarship incentives for students is the most cost-effective way to improve test scores among these six programs, and this is true even if one only values benefits for girls with low baseline test scores. The second most cost-effective program in terms of boosting test scores is the teacher incentive program, followed by textbook provision. Girls merit scholarships may also be a reasonably cost-effective way to boost school participation in some contexts (in particular, in Busia district schools), although even there it appears far less cost-effective than deworming.

5. Conclusion

Merit-based scholarships are an important part of the educational system in many countries. We present evidence that such programs can raise test scores and boost classroom effort as captured in teacher attendance. We find little evidence for the often hypothesized negative impacts of merit awards, and instead find considerable evidence for positive program spillovers. In particular, we

[26] One potential concern with these figures is the changing sample size, as different pupils took the 2000, 2001, and 2002 exams. But even if we consider the Busia girls cohort 1 longitudinal sample, where the sample is identical across 2000 and 2001, there are again no significant differences in test variance across program and comparison schools in either 2000 (program girls variance 0.89, comparison 0.92) or in 2001 (0.97 versus 0.89, respectively).

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 completely ineligible for the award – show higher test scores.

This evidence on within classroom learning externalities has several implications for research and for 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 the 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, where spillovers among students and teachers are likely.[27]

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 enough from merit scholarship program externalities to make it potentially cost effective for them, even neglecting the benefits to higher scoring students.

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 alternatively to conduct multiple competitions, each restricted to a restricted geographic area. For instance, if each Kenyan location – a small administrative unit – awarded merit scholarships to its residents independently of other locations, children would only compete against others who live in the same area, where many households share comparable socioeconomic conditions, thus effectively targeting a certain share of merit scholarships to disadvantaged households. To the extent that such a policy would put more

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

students near the margin of winning a scholarship, it could potentially generate even greater incentive effects and spillover benefits.

More speculatively, the finding in Busia, where the program clearly succeeded, that student attendance increased even for students with little or no chance of winning the award suggests that there was strategic complementarity between the effort levels of girls eligible for the award, the effort of teachers, and of other students. If such complemenatarity 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).

The classroom externality findings also have broader implications for education policy and the design of education systems. Currently, much thinking and debate about admissions and financing policies in education is static, focusing on what is optimal for the stage of education under consideration. Thus, for example, the debate about merit scholarships at the tertiary level in the United States often focuses on *assignment* and *equity* objectives within that level of education. However, to the extent that individual incentives for study effort are suboptimal, because effort creates positive externalities for classmates, in addition to thinking about educational admissions and financing policy from the standpoint of *assignment* and *equity* within a given level of education, it is important to also think about these policies from a third perspective, that of *incentives* at earlier levels of the educational system. This makes the problem inherently dynamic, since admissions and financial aid policies at any stage of the education system affect incentives at all earlier stages. Our externality findings suggest that merit scholarships at one stage of education may not only be useful as a way of helping talented students continue their education, but could also help increase performance across the board in the previous stage of the education system. In some cases, of course,

incentive objectives may align perfectly with assignment or equity objectives, but this will not be true generically. We believe that thinking about incentive effects in education provides a possible rationale for some common features of educational systems, can help explain some stylized facts – including differences between the U.S. and other educational systems – and suggests some policy implications, and we plan to pursue these issues in future research.

We find especially large average program effects on girls' test scores in Busia district, 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.

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. 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 (Leuven et al. 2003) while a second largely failed (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. Yet one pattern does emerge clearly from this growing literature: there is no evidence from any study that merit scholarships generate the adverse impacts on academic performance sometimes feared by educators and psychologists, nor is there evidence in our study that other leading objections to merit awards are empirically salient.

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.

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

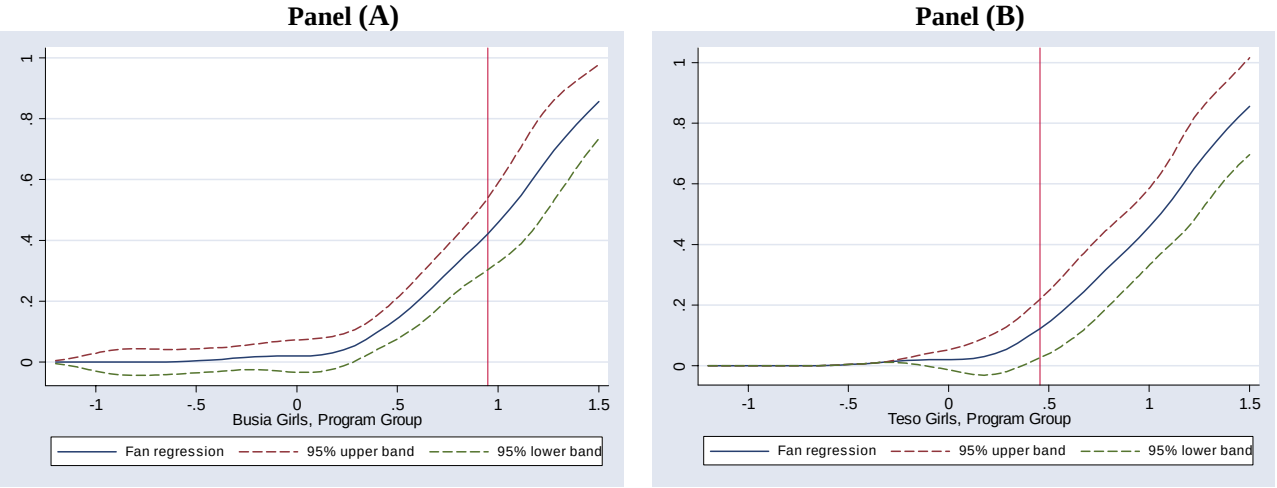

Figure 2: Map of Busia District and Teso District, Kenya,
with location of Girls Scholarship Program Schools (legend below)

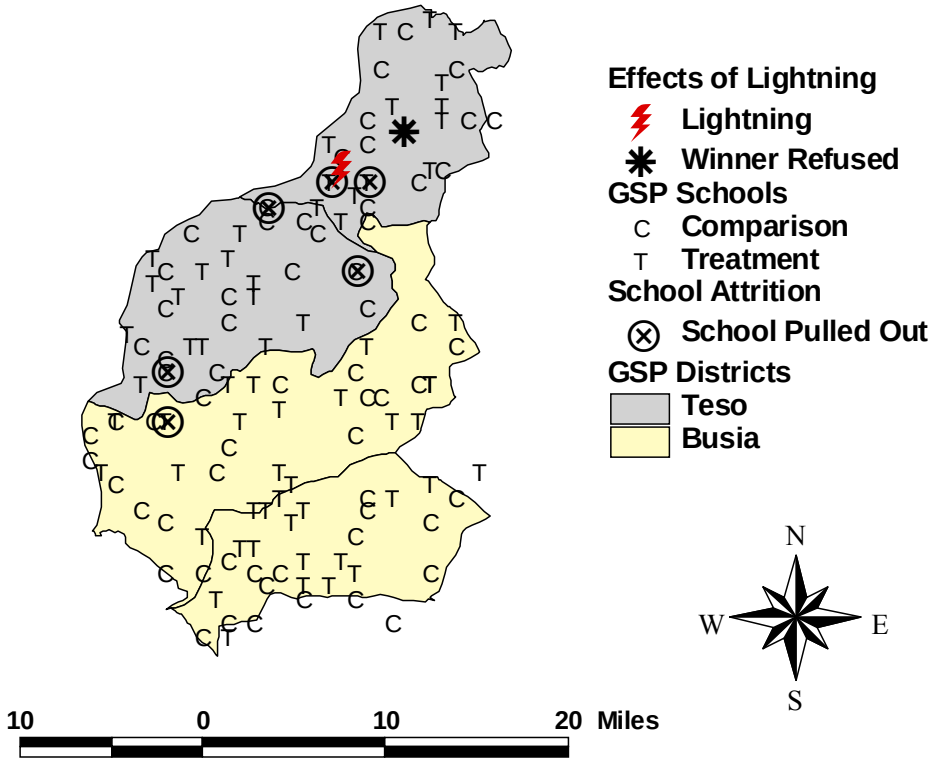

Figure 3: Proportion of Baseline Students in the 2001 Main sample by Baseline (2000) Test Score
Cohort 1 Busia Girls (Panel A) and Busia Boys (Panel B)
(Non-parametric Fan locally weighted regressions)

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

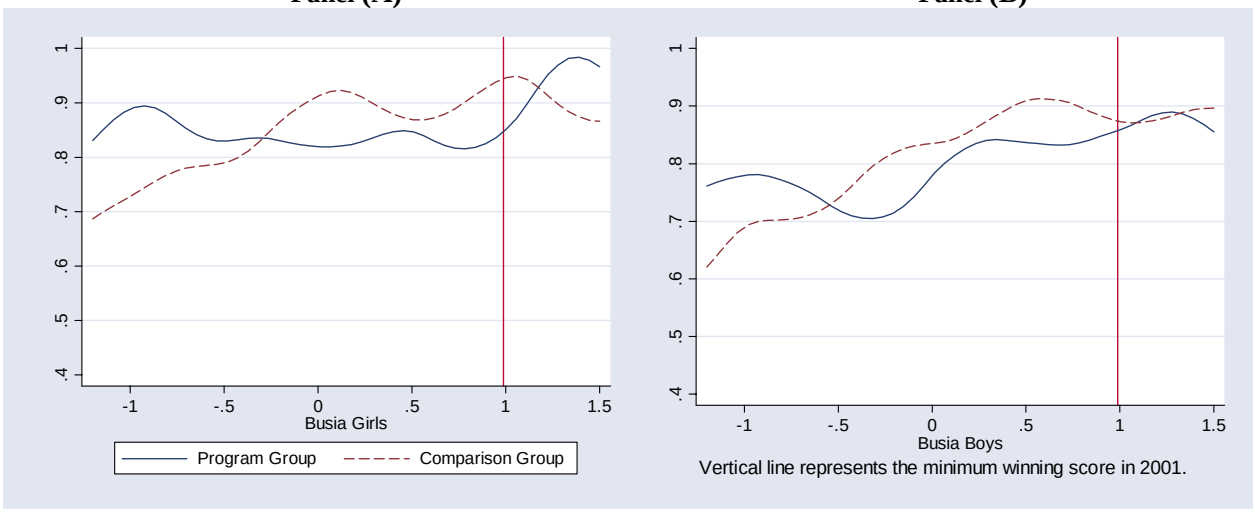

Figure 4: Proportion of Baseline Students in the 2001 Main sample by Baseline (2000) Test Score
Cohort 1 Teso Girls (Panel A) and Teso Boys (Panel B)
(Non-parametric Fan locally weighted regressions)

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

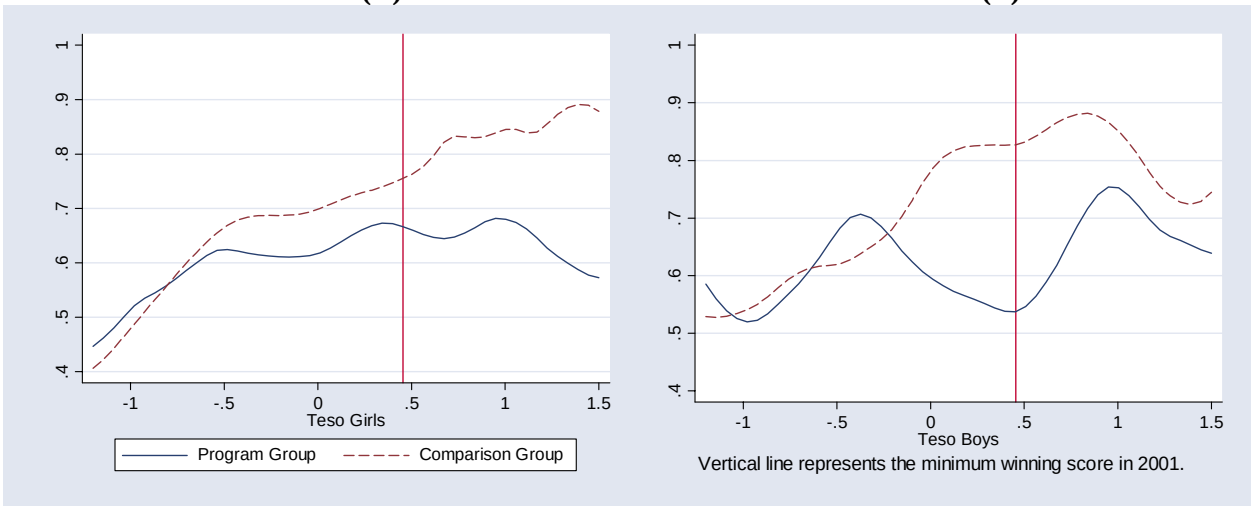

Figure 5: Baseline (2000) Test Score Distribution
Cohort 1 Busia Girls (Panel A) and Busia Boys (Panel B)
(Non-parametric kernel densities)

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

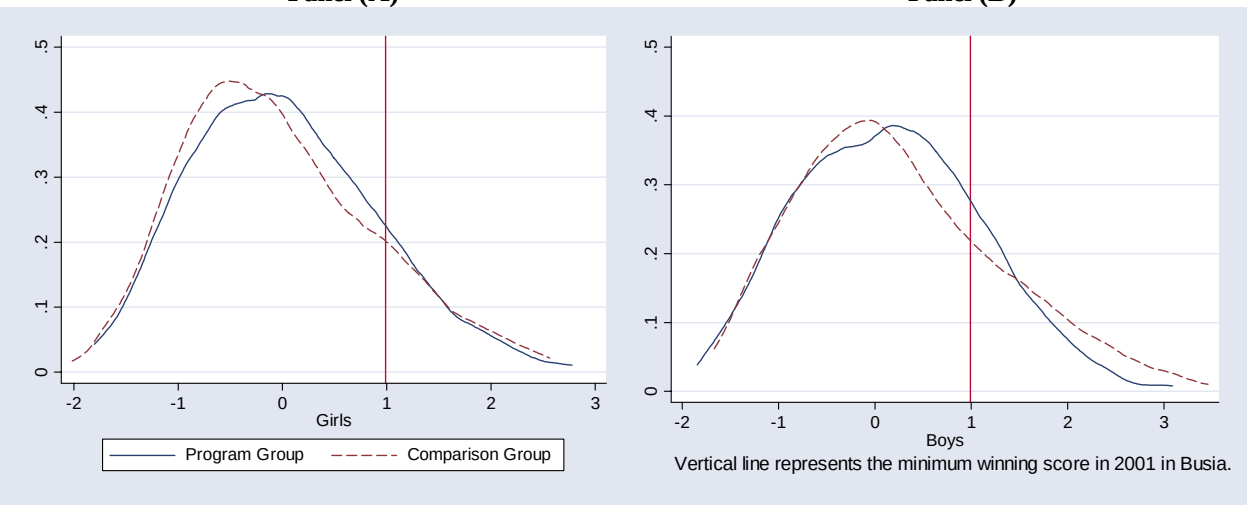

Figure 6: Year 1 (2001) Test Score Distribution
Cohort 1 Busia Girls (Panel A) and Busia Boys (Panel B)
(Non-parametric kernel densities)

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

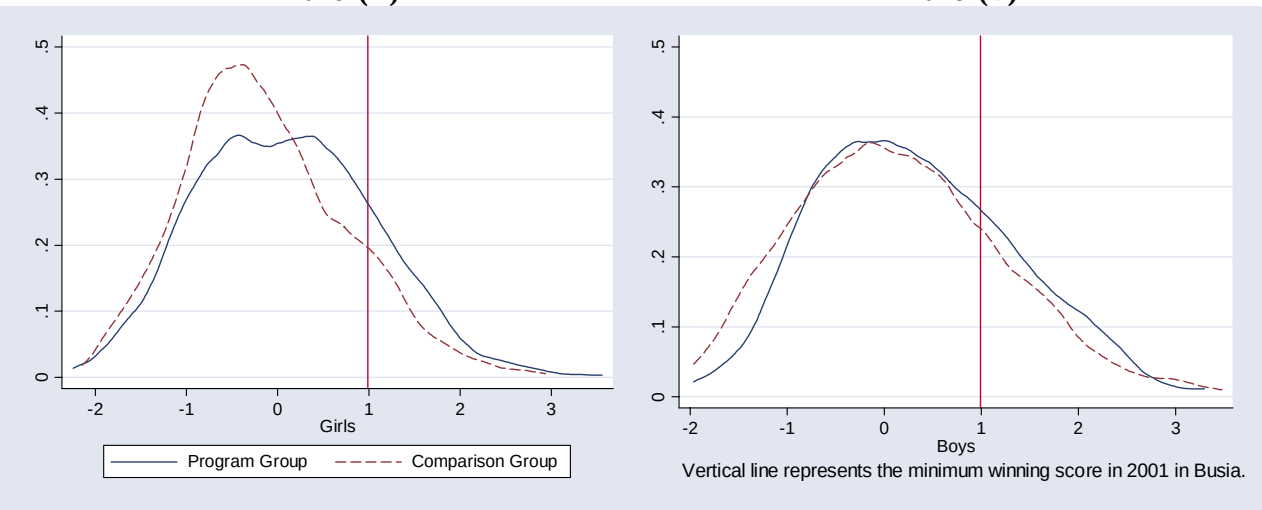

Figure 7: Year 1 (2001) Test Score Impacts by Baseline (2000) Test Score
Difference between Program Schools and Comparison Schools
Cohort 1 Busia Girls (Panel A) and Busia Boys (Panel B)
(Non-parametric Fan locally weighted regression)

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

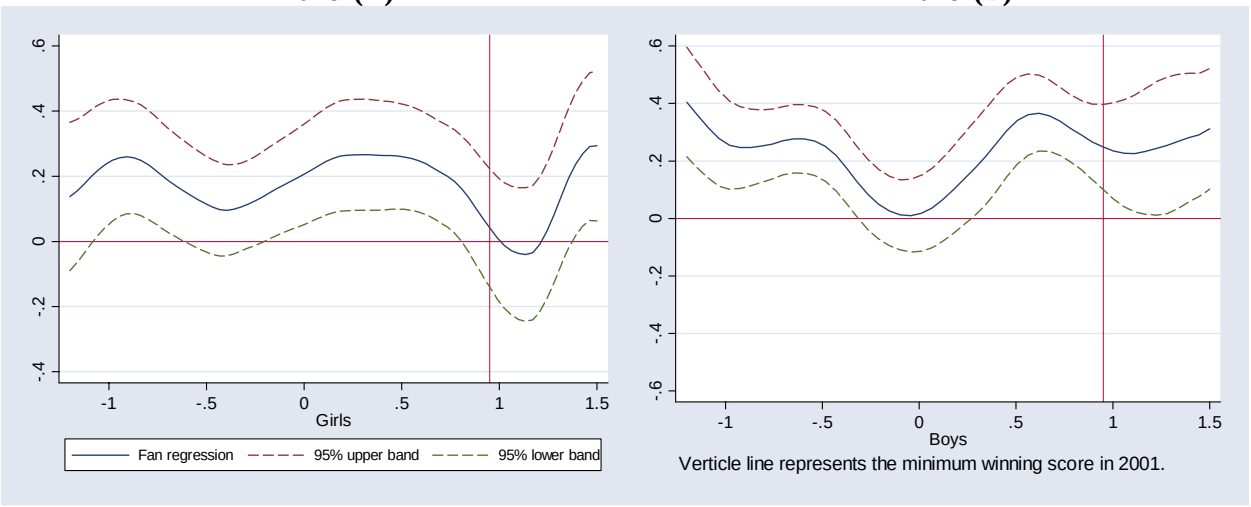

Figure 8: Year 2 (2002) Test Score Distribution
Cohort 2 Busia Girls (Panel A) and Busia Boys (Panel B)
(Non-parametric kernel densities)

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

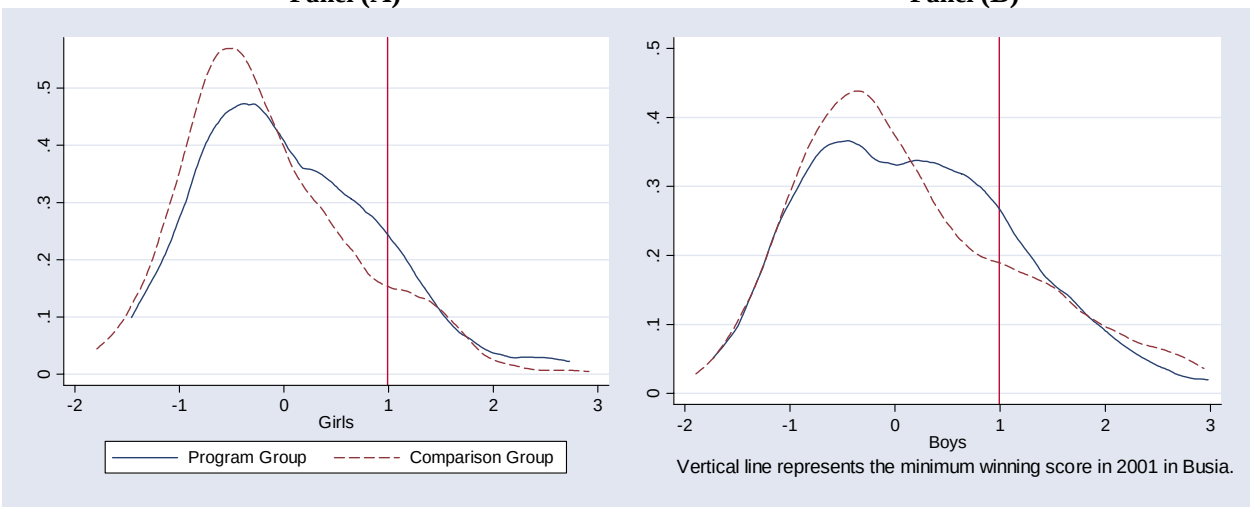

Table 1: Summary Statistics

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

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

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

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

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

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

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

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

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

Table 3: Demographic and Socio-Economic Characteristics Across Program and Comparison schools Cohort 1 and Cohort 2, Busia 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	13.9	13.7	0.2
			(0.1)			(0.2)
Father's education (years)	5.2	5.2	0.2	4.9	4.9	0.00
			(0.5)			(0.5)
Mother's education (years)	4.6	4.6	0.1	4.0	4.2	-0.2
			(0.4)			(0.4)
Total children in household	7.0	6.5	0.5	6.3	6.2	0.1
			(0.5)			(0.5)
Proportion ethnic Luhya	0.49	0.47	0.03	0.48	0.44	0.03
			(0.05)			(0.05)
Latrine ownership	0.96	0.94	0.02	0.95	0.93	0.02
			(0.01)			(0.02)
Iron roof ownership	0.77	0.77	0.00	0.72	0.75	-0.02
			(0.03)			(0.03)
Mosquito net ownership	0.33	0.33	0.00	0.27	0.26	0.01
			(0.03)			(0.04)
Test Score 2000–Baseline sample (cohort 1 only)	-0.05	-0.12	0.07	0.04	0.10	-0.07
			(0.18)			(0.19)
Test Score 2000–Main sample (cohort 1 only)	0.07	0.03	0.04	0.15	0.28	-0.13
			(0.19)			(0.19)
Panel B: Teso District	Program	Comparison	Difference (s.e.)	Program	Comparison	Difference (s.e.)
Age in 2001	14.1	13.8	0.21	14.1	14.1	0.05
			(0.19)			(0.18)
Father's education (years)	4.9	5.1	-0.18	4.7	4.6	0.12
			(0.87)			(0.73)
Mother's education (years)	4.1	4.2	-0.19	3.6	3.9	-0.28
			(0.78)			(0.4)
Total children in household	5.8	4.7	1.06**	5.4	4.8	0.62
			(0.47)			(0.50)
Proportion ethnic Luhya	0.07	0.08	-0.01	0.07	0.08	-0.01
			(0.02)			(0.02)
Latrine ownership	0.97	0.97	0.00	0.93	0.93	0.00
			(0.01)			(0.02)
Iron roof ownership	0.58	0.67	-0.09	0.49	0.59	-0.09**
			(0.04)			(0.04)
Mosquito net ownership	0.35	0.40	-0.05	0.28	0.29	0.01
			(0.04)			(0.03)
Test Scores 2000–Baseline sample (cohort 1 only)	0.04	-0.11	0.15	0.19	0.10	0.09
			(0.18)			(0.17)
Test Scores 2000–Main sample (cohort 1 only)	0.06	0.06	0.01	0.20	0.25	-0.05
			(0.19)			(0.17)

Notes: Standard errors in parenthesis. Significantly different than zero at 90% (*), 95% (**), 99% (***) confidence. Sample includes baseline students in cohort 1 and cohort 2 in 2001 in program and comparison schools in Busia district. Data is from 2002 Student Questionnaire, and from Busia District 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 Impact on Test Scores, Girls and Boys

	Dependent variable: Normalized test scores from 2001 and 2002				
Panel A: Longitudinal Sample	Busia and Teso Districts			Busia District	Teso District
	(1)	(2)	(3)	(4)	(5)
Program school	0.12	0.13**	0.12*	0.19**	-0.02
	(0.13)	(0.06)	(0.07)	(0.08)	(0.09)
Male * Program School			0.01	0.01	0.01
			(0.05)	(0.05)	(0.09)
Male			0.16***	0.09**	0.28***
			(0.04)	(0.04)	(0.07)
Individual test score, 2000		0.80***	0.79***	0.85***	0.69***
		(0.02)	(0.02)	(0.03)	(0.02)
Sample Size	4294	4294	4294	2858	1436
R^2	0.00	0.61	0.61	0.67	0.53
Mean of dependent variable	0.13	0.13	0.13	0.13	0.12
Panel B: Main sample	Busia and Teso Districts			Busia District	Teso District
	(1)	(2)	(3)	(4)	(5)
Program school	0.10	0.10*	0.14**	0.25***	-0.02
	(0.13)	(0.05)	(0.06)	(0.07)	(0.08)
Male * Program School			-0.07	-0.12*	0.02
			(0.05)	(0.07)	(0.07)
Male			0.31***	0.30***	0.32***
			(0.04)	(0.06)	(0.05)
Mean school test score, 2000		0.77***	0.77***	0.85***	0.66***
		(0.05)	(0.05)	(0.05)	(0.06)
Sample Size	10068	10068	10068	6123	3945
R^2	0.00	0.25	0.27	0.33	0.19
Mean of dependent variable	0.08	0.08	0.08	0.10	0.05

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

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

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

	Dependent variable: Normalized test scores from 2001 and 2002			
	----------Girls----------		----------Boys----------	
	Busia and Teso	Busia District	Busia and Teso	Busia District
	(1)	(2)	(3)	(4)
Program year, Cohort 1 (2001)	0.18**	0.28***	0.10	0.18**
	(0.08)	(0.10)	(0.07)	(0.09)
Program year, Cohort 2 (2002)	0.13*	0.21**	0.04	0.11
	(0.07)	(0.10)	(0.10)	(0.13)
Post-competition year, Cohort 1 (2002)	0.12	0.25***	0.05	0.07
	(0.08)	(0.09)	(0.07)	(0.09)
Mean school test score, 2000	0.75***	0.83***	0.78***	0.87***
	(0.05)	(0.05)	(0.06)	(0.06)
Sample Size	4736	2917	5332	3206
R^2	0.29	0.36	0.26	0.32
Mean of dependent variable	-0.06	-0.03	0.21	0.21

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

Table 6: Program Impact on Education Habits, Inputs, and Attitudes in 2002, Main sample, Cohort 2 Girls and Boys, Busia and Teso Districts

	Busia and Teso Districts ----------Girls----------		Busia and Teso Districts ----------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 parenthesis. Disturbance terms are allowed to be correlated across observations in the same school, but not across schools. Each coefficient estimate is the product of a separate regression, where the explanatory variables are a program school indicator, as well as mean school test score in 2000. Main sample includes students who were registered in grade 6 in January 2001, in schools that did not pull out of the program, and for whom we have school average test score data in 2000. The sample size varies from 700-850 observations, depending on the extent of missing data in the dependent variable.

[a] 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).

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

Panel A: Teacher attendance	Dependent variable: Teacher attendance in 2002		
	Busia and Teso Districts (1)	Busia District (2)	Teso District (3)
Program school	0.051***	0.067***	0.032
	(0.020)	(0.025)	(0.034)
Mean school test score, 2000	0.034***	0.034**	0.044**
	(0.012)	(0.016)	(0.019)
Sample Size	1065	652	413
R^2	0.03	0.04	0.03
Mean of dependent variable	0.84	0.86	0.83

Panel B: Student school participation	Dependent variable: Average Student School Participation (2001, 2002)						
	Busia and Teso Districts	Busia District		Busia Girls	Busia Boys	Teso District	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Program school	0.021	0.047*	0.050**			-0.019	-0.023
	(0.017)	(0.025)	(0.024)			(0.020)	(0.022)
Male * Program School			-0.007				0.007
			(0.017)				(0.017)
Male			-0.021				-0.025
			(0.012)				(0.009)
Program year, Cohort 1 (2001)				0.062	0.084*		
				(0.042)	(0.051)		
Program year, Cohort 2 (2002)				0.019	-0.024		
				(0.022)	(0.033)		
Post-competition year, Cohort 1 (2002)				0.029	0.016		
				(0.030)	(0.027)		
Pre-competition year, Cohort 2 (2001)				0.094*	0.096*		
				(0.049)	(0.060)		
Mean school test score, 2000	0.029**	0.015	0.015	0.014	0.096	0.054***	0.054***
	(0.013)	(0.016)	(0.016)	(0.015)	(0.060)	(0.016)	(0.016)
Sample Size	14034	8422	8422	4021	4401	5612	5612
R^2	0.01	0.01	0.01	0.90	0.88	0.01	0.01
Mean of dependent variable	0.85	0.85	0.85	0.86	0.84	0.85	0.85

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

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.

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

Appendix A: Girls Scholarship Program Cost-effectiveness

Appendix Table1 reports the test score gains associated with various ICS programs in Western Kenya, their costs, and several different measures of cost effectiveness. The average test score gain in girls' scholarship program schools, among girls and boys in both years of the program, in Busia and Teso districts pooled together is 0.12 standard deviations (Table 4). The deworming (Miguel and Kremer 2004), flip chart (Glewwe et al. 2004), and child sponsorship programs (Kremer et al.2003) did not produce statistically significant test score impacts. The comparable gain for textbook program schools was only 0.04 standard deviations (Glewwe et al. 1997), and for teacher incentive program schools over two years was 0.07 standard deviations (Glewwe et al. 2003). The GSP program is thus the most effective in raising test scores of any of these programs. The test gains in the teacher incentive program were concentrated in the year of the competition, and then fell in subsequent years. Glewwe et al. (2003) interpret this, together with evidence of increased test preparation sessions, as indicating that the program did not increase long-term learning.

One important issue in evaluating cost-effectiveness is whether to treat all payments under the program as social costs or whether to consider some as transfers. Appendix Table 1, column 4 reports "education budget cost effectiveness", which shows the test score gain per pupil divided by program costs per pupil. This is the relevant calculation for an education policymaker seeking to maximize test gains with a given budget. From the standpoint of a social planner, however, payments to families in the scholarship program, and to teachers in the teacher incentive program, could be considered transfers. If seen as pure transfers, the social cost is simply the deadweight loss involved in raising funds. In calculating "social cost effectiveness" (column 5) we follow a rule of thumb often used in developed countries and treat the marginal cost of raising one dollar as 1.4 dollars (Ballard et al. 1985). To make the education budget and social cost effectiveness figures comparable, we also multiply costs in the education budget calculations by 1.4 to reflect likely tax distortions.

It is worth noting that the effective transfer is the net benefit to recipients after making allowances for any disutility of their increased effort. Assuming that students and teachers are rational and that the key incentive for students is the cash award, the total additional effort exerted should not be greater than the value of the award. Thus the education budget cost effectiveness calculation yields an upper bound on the true social cost of the program (Appendix Table 1, column 4), and a lower bound is generated by treating the entire payment as a transfer (as in column 5).

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

The education budget approach also provides a valid measure of cost-effectiveness from the perspective of a social planner who values only the welfare gains for girls with low baseline 2000 test scores, since these girls have little chance of winning an award (Figure 1) and thus expected transfers to them are essentially zero. Low performing girls' parents tend to have less education than average, as discussed above. For this relatively disadvantaged group, the merit award program is a particularly cost-effective way to boost test scores relative to textbook provision, since textbooks only raised scores for students in the top quartile and not elsewhere (Glewwe et al 1997), while merit awards lead to test score gains throughout the baseline distribution.

The estimates for both the girls' scholarship program and the teacher incentive program do not include costs associated with administering academic exams in the schools, which are substantial. Including testing costs, the social cost per 0.1 standard deviation average test score gain nearly doubles

for the girls scholarship schools (Busia and Teso together), from US$1.41 to US$2.78, and more than doubles from US$1.36 to US$3.70 for the teacher incentive program. Many countries, like Kenya during the study period, already carry out regular standardized testing in primary schools, in which case the additional exam costs are unnecessary and the previous lower estimates are relevant.

Although test score cost effectiveness figures are similar for the girls scholarship and teacher incentive programs overall, the scholarship program is more attractive in other dimensions. The girls scholarship program also generated large impacts on teacher attendance, and any benefits of this to pupils in other grades are not considered in the above calculations. If winners have high returns to additional education, then to the extent that winners obtain more education than they would have otherwise, this yields additional benefits. Finally, the distributional impact of the teacher incentive program is likely to be less equitable since it provides cash awards to teachers, who tend to be well-off in rural Kenya.

School participation is a second educational outcome measure important to policymakers. Deworming provision is far more cost effective in this dimension than the other interventions, including merit awards, at an average cost of only US$3.50 per additional year of school participation. There are no significant school participation gains from teacher incentives, textbook provision, flipcharts, or for the merit awards in the pooled Busia and Teso sample. However, for Busia district alone the cost per additional year of school participation (Table 7 Panel B) is US$4.24 / 0.047 = US$90, making merit awards the second most cost-effective of these six programs in that case. The cost per additional year of school participation for the child sponsorship program was US$99, making it slightly less cost-effective than merit scholarships in Busia.

Appendix Table 1: Test Score Cost-effectiveness of Various Kenyan Primary School Interventions

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

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