

**Incentivizing schooling for learning:
Evidence on the impact of alternative targeting approaches¹**

Felipe Barrera-Osorio

Harvard Graduate School of Education

and

Deon Filmer

World Bank

Author's Manuscript: Final post peer-reviewed accepted version

Forthcoming in *Journal of Human Resources* (<http://jhr.uwpress.org/>)

Abstract. We evaluate a primary school scholarship program in Cambodia with two different targeting mechanisms, one based on poverty level and the other on baseline test scores (“merit”). Both targeting mechanisms increased enrollment and attendance. However, only the merit-based targeting induced positive effects on test scores. We show that the asymmetry of response is unlikely to have been driven by differences between recipients’ characteristics. We explore evidence for the channels by which the framing of the scholarships might have led to different impacts. The results suggest that in order to balance equity and efficiency a two-step targeting approach might be preferable: first, target low-income individuals, and then, among them, target based on merit.

JEL classification codes: I21; I24; I28; O10

Keywords: education; Cambodia; randomization; scholarships; merit-based targeting; poverty-based targeting.

¹ We thank Luis Benveniste, Norbert Schady, Beng Simeth, Tsuyoshi Fukao, and the members of the Primary School Scholarship Team of the Royal Government of Cambodia’s Ministry of Education for valuable input and assistance in carrying out this work. Adela Soliz provided able research assistance. The paper has also benefitted from comments by David Deming, Kevin Lang, Leigh Linden, Muna Meky, Richard Murnane, Halsey Rogers, Shwetlena Sabarwal, and Katja Vinha, as well as from comments by seminar participants at University of Texas at Austin, Harvard Graduate School of Education and Boston University. The authors are responsible for any errors. This work benefited from funding from the World Bank as well as from the EPDF Trust Fund (TF095245). Barrera-Osorio received from the Harvard Graduate School of Education a Faculty Grant to develop the evaluation. The findings, interpretations, and conclusions expressed in this paper are those of the authors and do not necessarily represent the views of the World Bank, its Executive Directors, or the governments they represent.

1. Introduction

Studies of the impacts of programs that incentivize schooling through cash transfers typically have found that direct and indirect costs are important determinants of school participation and that programs reducing such costs are quite effective in inducing higher enrollment and attendance rates. However, the programs do not show consistent positive impacts on learning outcomes.

One hypothesis that emerges from this literature (particularly the literature from developing countries) is that monetary incentives may increase the participation of low-achieving students in school, and that school systems may be ill-prepared to teach them well (Behrman, Parker, and Todd 2011; Filmer and Schady 2009, 2014). This gives rise to a potential equity-efficiency tradeoff: targeting high-achieving students through an incentive based on academic performance such as a merit scholarship may yield higher learning outcomes—but if academic performance is correlated with economic circumstance, then those outcomes would come at the cost of reaching the poor. Poverty-targeted incentives, on the other hand, do reach the poor and do induce greater schooling—but if there is little learning to show for it, one might question the usefulness of that approach. The potential tradeoff between equity and efficiency has been at the center of discussions in many contexts: discussions on college scholarships in the United States (Orfield 2002), social programs in developing countries (Coady et al., 2003), and more generally on poverty reduction strategies in the presence of tight budget constraints (Bardhan, 1996).

This paper addresses the equity-efficiency tradeoff directly in the context of Cambodia, a low-income country, by evaluating the impacts of a government scholarship program that started

in 2008.² The main objective of the government with this program was to increase school enrollment and progression in Grades 4 to 6. The program included two approaches to targeting that were run in parallel: schools in one group offered scholarships based on the economic status of the student's household (which we refer to as poverty-based targeting), and schools in another group offered scholarships based on the student's performance on a baseline test (which we refer to as merit-based targeting). Other than the mechanism for selecting the scholarship recipients, the conditions of the program under both targeting mechanisms were identical.

We find that both targeting approaches increase enrollment and attendance indicators. For example, the probability of reaching Grade 6 increased by 19 percentage points for poverty scholarship recipients, and by 14 percentage points for merit scholarship recipients (over a counterfactual rate of 60 percent). However, only the merit-based scholarship produces a positive impact on achievement as measured by test scores: performance on a math test was 0.17 standard deviations higher, and on the digit span test was 0.15 standard deviations higher, for the merit scholarship recipients—with both of these estimates statistically significantly different from zero. On the other hand estimates of these impacts among poverty scholarship recipients were close to zero (-.05 and -.06 standard deviations respectively for the math and digit span tests) and insignificantly different from zero.

This asymmetry in results does not appear to be driven by the characteristics of the students themselves. High baseline achievers who received the poverty-based scholarship performed no better in follow-up tests than the corresponding control group. On the other hand, similarly poor

² This study narrowly defines efficiency as the effects of a program on educational outcomes per dollar cost of the program. We do not investigate the effects on efficiency of raising the money for the program. The program transfers cash to poor households, which is potentially welfare-enhancing in itself, but we do not address this aspect of the program directly. Another issue that is not addressed is whether either of the interventions influences other outcomes such as health.

individuals who received a merit-based scholarship did perform better on the follow-up test (as did non-poor merit-based scholarship recipients). In summary, “identical” children who received differently labelled scholarships experienced different impacts on learning outcomes. We provide suggestive evidence that merit-based scholarship recipients and their families exerted additional effort as a result of the scholarship—as measured by homework and by expenditures on education—whereas poverty-based scholarship recipients did not.

Our findings suggest that the impact of the program was dependent on the way that the intervention was framed. We document pathways that are consistent with motivation being an important determinant of the program’s success in incentivizing students, and that the framing of the scholarship itself affected motivation. In order to maximize both equity and efficiency objectives, these results suggest that demand-side programs should ensure that student (and family) motivation is enhanced through the program, rather than potentially undermined. An intervention that targets students with an incentive that recognizes their high academic potential—while ensuring that the poorest students are among that set—would be one potential way to achieve this.

We cannot investigate all the pathways for how framing might matter—for example, the data do not allow us to assess the extent to which teachers might respond differently to being presented with a poverty-based scholarship recipient as opposed to a merit-based scholarship recipient. In addition, as with other evaluations of this type, our results should be understood as conditional on the environment in which this intervention is taking place. Based on this intervention we do not know whether wholesale simultaneous improvement in the quality of schooling provided would have resulted in better learning outcomes among recipients of the

poverty-based scholarships. Further research to understand how additional channels might mediate impacts, or what alternative approaches might work for improving learning outcomes among the poor, would be important to provide further guidance to policy. But our results suggest that even a modest change in how conditional cash transfers are implemented—that is altering their framing so that they motivate greater effort—can positively affect impacts.

This paper is organized as follows. After motivating the analysis and discussing the pertinent literature in section 2, we describe the setting, the program, and the evaluation design in section 3. In section 4, we present the empirical strategy, the data, and the validation of the identification strategy. The main results appear in section 5. In section 6, we explore pathways for the impacts and the results on the equity-efficiency tradeoff. A conclusion follows, which reviews the implications of the findings for designing future incentive programs and points to potential future research.

2. Motivation: The efficiency of incentive programs

Monetary incentive programs (scholarships included) are thought to induce greater schooling for three main reasons. First, the direct and indirect costs of attending school, along with the lack of financing to cover those costs, may deter families from making optimal decisions on education; reducing those costs may induce greater investment in education. Second, students and families may discount future returns to education very heavily and, as a result, they may not invest the optimal effort in education. Monetary incentives will increase the short-run benefits of such investments. Third, families and students may not have complete information on the returns to education. Monetary incentives may serve as a means of signaling that education is important.

Scholarship programs, such as the one evaluated here, can be thought of as a particular type of Conditional Cash Transfer (CCT) program—one that has an individual rather than a family as the recipient, and one that focuses on a single sector (education) rather than being conditional on actions related to two sectors (health and education).³ Extensively studied and increasingly popular in much of the developing world, CCT programs have become the largest form of social assistance provided in several countries (see the reviews in Fiszbein and Schady, 2009; Independent Evaluation Group, 2011; Baird et al. 2013; Saavedra and Garcia 2013). Much of the rigorous evidence on the impact of CCTs on student enrollments has come from middle-income countries, mainly in Latin America, where baseline enrollment rates are high and impacts relatively small (in part because there is little room for an increase). In contrast, this paper presents evidence from a scholarship program in Cambodia, where enrollment rates are low. It adds to the much smaller evidence base on CCTs from low-income countries, such as Pakistan (Chaudhury and Parajuli, 2008) and Malawi (Bair, McIntosh, and Ozler, 2009), as well as evidence from other programs in Cambodia (Filmer and Schady, 2008, 2011, 2014).

CCT programs designed to increase enrollment and attendance have indeed raised these measures of school participation, but in some settings they have resulted in negative or insignificant changes in learning outcomes (Behrman, Parker, and Todd, 2005; Behrman, Sengupta, and Todd, 2000; Filmer and Schady, 2014; Krishnaratne, White and Carpenter 2013; McEwan forthcoming; Murnane and Ganimian 2014), whereas in others they have had positive results (Baird, McIntosh and Ozler, 2011; Barham et al, 2013). Similarly, the effects of cash incentives linked to changes in achievement itself have yielded mixed results. These programs

³ Conditional Cash Transfer (CCT) programs transfer cash to families that comply with a set of conditions, such as the enrollment and regular attendance of children in school, regular prenatal visits by pregnant women, and regular health checkups for young children.

aim to encourage the performance of students who are already in the educational system. Two recent evaluations in the United States showed mixed (Fryer, 2011) or no impacts (Bettinger, 2010) on standardized tests; an evaluation of a program in Israel showed positive impacts on exit-exams, take-up rates, and college matriculations (Angrist and Lavy, 2009). Finally, a mixed approach using merit-based scholarships in Kenya yielded positive effects on test scores in one of two districts (Kremer, Miguel, and Thornton, 2009).⁴ The question of how to effectively turn incentives for schooling into incentives for learning remains open.

In Cambodia, two evaluations of the impact of scholarships for lower secondary school have shown substantial increases in school enrollment and attendance as a direct consequence of the programs (Filmer and Schady, 2008, 2011, 2014). Recipients were 20–30 percentage points more likely to be enrolled and attending school as a result of the scholarships. These studies also show that scholarships targeted to lower secondary school students led to more family expenditure on education and less work for pay among recipients. Impacts on learning outcomes were limited. The authors argue that the limited impacts point to potential issues in the quality of education and the match between students' skill levels and instruction (particularly among students induced to stay in school as a result of the scholarship).

Three main explanations are typically put forward for mixed findings on test scores. First, the positive results may reflect the capacity of monetary incentives to act as an extrinsic motivator (for the student and family), especially among individuals from low-income families. Such short-term motivation may be important when information on the returns to education is imperfect and

⁴ Besides giving scholarships concurrent with studies, some programs “promise” students a scholarship if they perform well on a test administered in the future. In theory, the promise of a reward elicits increased effort from students whose abilities place them within reach of a scholarship. Kremer, Miguel, and Thornton (2009) is an evaluation of one such program.

the discount rate is very high. Bettinger (2010) argues that these considerations may be highly relevant for primary school students in the United States. Second, scholarship programs can induce a negative impact by reducing intrinsic motivation. A person's motivation to perform well can decline if she comes to view performing well as an obligatory activity for achieving a certain goal, in this case a scholarship (for example, see Lepper et al., 1973; Deci, Koestner, and Ryan, 2001). Finally, these programs may have no impact on academic achievement if students are unable to respond to the incentives. Students may not know how to convert the incentive into actions that influence achievement (Fryer, 2011). A related issue is that the incentive may have strong complementarities with other inputs that are out of a student's control and that the scholarship program does not alter, such as the quality or appropriateness of the teaching (Fryer, 2011): a student or her family may have a substantial amount of control over whether she attends school but far less control over the factors that enable that schooling to be converted into learning.

We explore an additional explanation that is consistent with the contrasting impacts of the poverty-based and merit-based targeted scholarships we study here. If the targeting approach itself changes the frame within which various actors are responding to the incentive, and that frame matters for impact, then the impact of the program on test scores will be dependent on the actual targeting mechanism. This line of reasoning is aligned with a (large) psychology literature on "stereotype threat" (for a general theoretical framework, see Schmader, Johns, and Forbers, 2008). When certain individuals are faced with a specific label that carries a social stigma, labeling interferes with intellectual performance. The mechanisms by which labeling can affect performance are diverse: anxiety, stereotype activation, self-doubt, working memory, and arousal (Schmader, Johns, and Forbes, 2008). For example, African Americans responded to stereotype

threats by lowering performance on tests (Steele and Aronson, 1995). Labeling associated with caste in India affected the performance of lower-caste individuals when the caste was publicly announced (Hoff and Pandey, 2006). Aronson et al. (1999) shows that this effect can be present, even in the absence of a social stigma attached to the labeling; white males underperform in relation to Asian students. But there is a broader set of ways that labeling can matter: it sends signals and messages to the recipients and others and different signals and messages may induce different behaviors. After describing the program, the evaluation and the results we return to this issue in the interpretation of our findings.

3. Program and evaluation design

Cambodia has a recent record of using demand-side incentives to raise school enrollment and attendance rates. Some of these programs operate at the primary school level—such as school feeding programs or small-scale programs that offer incentives for children to attend school—but most are targeted at lower secondary school (Filmer and Schady, 2008, 2014). The programs do not simply waive school fees; instead, the families of children selected for a “scholarship” receive a small transfer, conditional on school enrollment, regular attendance, and satisfactory grades.

One important finding from the evaluations of earlier incentive programs is that, relative to the population as a whole, the targeting only mildly favored the poor. Filmer and Schady (2009) show that one program, despite reaching the poorest children who *applied* for scholarships, did not reach the poorest of the poor, who had already dropped out of school before Grade 6—when they would have applied for secondary school scholarships. Figure 1 presents the proportion of children ages 15–19 nationally who completed each grade at the time the program we study in

this paper was launched. The figure shows that children from the poorest quintiles are the least likely to reach 6th grade. This finding suggests that it is hard for a program that targets children at the end of Grade 6 to be strongly pro-poor—and that a program targeting poor students earlier in the schooling cycle is needed when the goal is to reach the poorest of the poor.

Primary School Scholarship Pilot Program

Based in part on these findings, and the desire to assess the viability, effectiveness, and optimal design of such a program, the Government of Cambodia implemented a new pilot scholarship program in 2008. The program's stated goal was to offset the direct and opportunity costs of schooling, and increase educational attainments as a result.⁵ An implicit goal was also to improve learning outcomes through the additional education. This paper reports the results of the impact evaluation of that pilot program.

The basic design of the primary school scholarship pilot was to select participating schools; randomly assign them to offer scholarships based on poverty or merit; and then, within each school, identify candidates for scholarships based on transparent, clearly articulated criteria. Once selected, scholarship recipients were required to stay enrolled, attend school regularly, and maintain passing grades to keep the scholarship until they graduated from primary school.⁶ In Cambodia, primary school consists of Grades 1 through 6, and the program targeted students entering the upper-primary level (Grades 4, 5, and 6). The scholarship was equivalent to US\$ 20

⁵ Primary schools are officially non-fee based. Opportunity costs include various forms of child labor which are relatively prevalent in the study area—although typically labor is combined with schooling at the primary school ages.

⁶ These requirements are moderately enforced. Students absent for many days are followed up by school officials and if they return to school they remain eligible for the scholarship. After a student is absent for too many days they are classified as having dropped out and no longer are eligible for the scholarship.

per student annually.⁷ Through a follow-up survey, we determined that the mean household expenditure per capita per year in our sample was US\$ 610;⁸ as such, the scholarship represents 3.3% of the yearly per capita expenditure (the average household size in our sample is 6.99). The scholarships were intended to be disbursed in two tranches of US\$ 10—the first in the beginning of the year, and the second in the middle of the school year. In the first year of the program, however, scholarships were distributed in one lump sum due to delays in implementation.⁹

The pilot program was implemented in three provinces—Mondulkiri, Ratanakiri, and Preah Vihear—where average dropout rates in the upper-primary grades were highest, according to an analysis of Cambodia’s Education Management Information System (EMIS). To narrow the geographic scope of the program, only seven districts in Ratanakiri (those with the highest dropout rates) were selected for participation, out of a total of nine districts. In the other two provinces, all districts were included. Within these selected districts, all primary schools that offered classes through Grade 6 participated in the program.

Of the 207 program schools, 103 were randomly assigned to join the program in Phase 1 (the program’s first year, 2008/09) and the other 104 in Phase 2 (the second year of the program, 2009/10). In each phase, schools were randomly assigned to either poverty-based or merit-based scholarship targeting. In total, the program offered scholarships to approximately 5,162 students from a pool of 12,066 individuals in the 207 program schools—although our analysis focuses only on the subset (Grade 4 students) that were offered them in the first year of implementation.

⁷ The previous lower-secondary scholarships were in the amounts of \$45 and \$60 per year (Filmer and Schady, 2011)

⁸ GNI per capita in Cambodia was, approximately, UD\$700 for 2008 (World Development Indicators).

⁹ Scholarship distributions for the cohort of recipients analyzed here (Phase 1) took place in July 2009 (US\$ 20), November 2009 (US\$ 10), April 2010 (US\$ 10), November 2010 (US\$ 10); and April 2011 (US\$ 10).

Schools using poverty-based targeting selected students based on a poverty index. All students in Grade 4 filled out a simple form with questions regarding their household and family socioeconomic characteristics.¹⁰ These forms were scored according to a strict formula based on weights derived from an analysis of household survey data to derive a poverty index for each student.¹¹ The poverty index ranged from 0 (richest household) to 292 (poorest household). The application forms were scored centrally by a firm contracted specifically for this purpose, thereby reducing the probability of any manipulation of program eligibility. Within each school, the applicants with the highest scores (that is, those with the highest level of poverty) were offered a scholarship.

Schools using merit-based targeting selected scholarship recipients based on scores on a test administered at baseline. The test included questions in math and Khmer (the national language). The maximum possible score was 25 points/correct answers. The test was adapted from the Grade 3 National Learning Assessment.¹² All students in Grade 4 took the test, and within each school the applicants with the highest test scores were offered scholarships. Again, the tests were scored centrally to minimize the risk of program manipulation.

Our analysis exploits the fact that all students in the targeted grade filled out both the “poverty form” as well as the math and Khmer test—regardless of which type of targeting their

¹⁰ The first 14 entries of Table 2 report the full set of variables included in the calculation of the score.

¹¹ The weights were determined by estimating a model predicting the probability that a student would drop out of school during Grades 4 to 6—since addressing this dropout was the stated goal of the program. Strictly speaking, the score should be referred to as a “dropout-risk score.” However, the risk is essentially a set of characteristics that capture the socioeconomic status of a household, weighted to capture those elements that predict dropout best. For convenience and ease of exposition, the score is referred to in this paper, as well as in program documents, as a “poverty” score. It might seem heroic to rely on the responses of 4th graders to a socioeconomic questionnaire to rank households. We believe these are reliable because (1) the items asked about are simple and easily understood by the young respondents (2) the poverty score is negatively associated with endline (log) household consumption expenditures for the control group ($\rho = -0.13$, $p < 0.01$).

¹² The National Assessment was implemented in a sample of schools nationwide in Grade 3 during the 2005/06 school year (Royal Government of Cambodia, 2006).

school would ultimately use to select scholarship recipients. In addition, we use the fact that Grade 4 students in Phase 2 schools (who would be ineligible for scholarships) also filled out the forms and took the tests.

School head-teachers, as well as other school-level stakeholders, were aware (after the initial forms were completed and before the scholarship recipients were announced) which type of scholarship was being allocated in a school. School-level spot checks suggest a high degree of understanding of the nature of the targeting, likely due to the extensive “socialization” of the program that took place.

The number of students within each school who would receive a scholarship was fixed exogenously, and set to half the number of registered students in the year prior to the program (as determined by an analysis of EMIS data).¹³

Evaluation design and data

Application forms (including baseline tests) for the sample that we study in this analysis were filled out in December 2008/January 2009. Recipient lists were circulated to schools in May 2009, which was prior to the distribution of scholarships which took place at the beginning of July 2009—i.e. the end of the 2008/2009 school year.

The identification of impacts relies on the fact that among the cohort of students studied, 4th graders in Phase 2 schools were not eligible for scholarships in the 2008/09 school year or later.

¹³ The number of scholarships is not equal to half the number of applicants for three reasons: First, the rule was to allocate scholarships to all applicants who had scores *higher than or tied with* the cutoff score; second, there were Grade 4, Phase 2 schools that did not implement the scholarship in the period of the first phase (control schools). Third, there is year-to-year variation in the number of enrolled students. Covering half the students in a class/school is relatively high compared to other programs. This might affect the external validity of our findings if they are to inform programs that cover a much smaller share of students.

Among these students, we identify the valid counterfactual group—a group that differs, on average, from the treatment group only in that it did not receive scholarships. Since students in the control schools were never exposed to the program (even after the *subsequent* cohorts in Phase 2 schools became eligible for scholarships), the two groups of students can be tracked over time and their enrollment, attendance, and other outcomes compared.¹⁴ Recipients received scholarship disbursements (conditional on remaining in school, attending regularly, and maintaining passing grades) during the 2008/09, 2009/10, and 2010/11 school years.

We use three main data sources to evaluate program impacts. First, we use the full set of data collected for program eligibility: data on baseline household characteristics (which were used to construct the poverty index), as well as mathematics and Khmer language test scores for all students in the 207 schools. Second, we use the official list of students who were offered a scholarship.¹⁵ Third, we use endline data that were collected specifically for this evaluation. These data are derived from a survey administered at the end of the 2010/11 school year, three years after the program began implementation, to a random subsample of students from the 207 program schools. The students who were in Grade 4 at the start of the program and stayed in school were finishing (or had just finished) Grade 6 at the time of the survey. The survey was administered at home (not in schools) to the child who applied for the scholarship, and it included a household module administered to the child’s mother, father, or other caregiver.

¹⁴ Because the scholarship offers are made according to an strict cut-off criteria, it is possible to implement a regression discontinuity design (RDD) approach to evaluate program impact. Below we present the main result from this strategy.

¹⁵ We use the official government declaration (“Prakas”) of recipients.

Of the 5,902 students in Grade 4 at baseline (Table 1), we provided a list of randomly selected 2,952 students¹⁶ to the survey company to apply follow up surveys, as well as a list of potential (randomly generated) “replacement” students to be interviewed if the original person selected could not be found. The survey located 2,274 (678 attritors) and 174 replacement households, yielding an overall Grade 4 attrition rate of 23%.

The endline survey includes measures of school attainment, of the “intensity” of school participation (based on questions related to time spent in school), as well as two measures of achievement and cognitive development based on a mathematics and a “Digit Span” test. The items on the mathematics test were drawn from a variety of sources, including the baseline mathematics test, questions from the Grade 6 National Assessment, as well as publicly released items from the Trends in International Mathematics and Science Study (TIMSS) Grade 4 Assessment. A pretest ensured that only items with adequate properties were retained for the final test. The final test measured both knowledge and the capacity to use this knowledge for problem solving.

Performance on the math test is a measure of the most immediate academic impact of the intervention. The program is expected to have an impact on endline test scores among students selected through each targeting mechanism for two main reasons. First, the program incentivized more enrollment and school attendance; consequently students are more exposed to school, and through that additional schooling they potentially increase their skills and knowledge. Second, because the program requires that all treated students—those with merit and those with poverty scholarships—maintain passing grades, the program is expected to incentivize students to study more, which in turn should affect their ability to solve mathematical problems.

¹⁶ Includes all Grade 4 students (phase 1 and phase 2, treated and non-treated students).

In a Digit Span test, a series of numbers are read to a respondent who is then asked to repeat the numbers back to the enumerator. The series increase from 2 digits to a larger and larger number, up to 9-digit numbers. Respondents are also asked to repeat the numbers back in reverse order. The test is often included in batteries of psychometric tests, has been used in previous analyses of development programs (for example, see Kazianga, de Walque, and Alderman, 2012), and is typically interpreted as a measure of short term memory and working memory capacity (Schmader, Johns, and Forber, 2008). Nevertheless, recent articles suggest that higher academic achievement can also impact cognitive ability (Cascio and Lewis 2006; Hanushek and Woessman 2008; Carlsson, Dahl, and Rooth 2012). We interpret impacts on both of these tests as measures of academic achievement. These tests exhibit a high degree of internal consistency: the Cronbach alpha statistic for the math test is 0.71 and for the Digit Span test it is 0.65.¹⁷ For the multiple-choice math questions, there is also a positive correlation between the correct answer to each item on the test and the overall test score, and a negative correlation between each incorrect answer and the overall test score.¹⁸

Two particular characteristics of the data are noteworthy. First, we use follow-up information collected from home visits rather than at the school. This strategy allows us to avoid the problems that arise when follow-up information is collected at school, where students who do not attend would not be present. Second, we collect follow-up data three years after the students

¹⁷ In Classical Test Theory this statistic is typically used as a measure of the reliability of a test and captures the covariance among the various items of a test as a share of the sum of the variances and covariances of the items. We interpret it here mostly as a measure of the extent to which the various items on the tests capture a single underlying construct (such as ability). Values above 0.6 are typically considered “acceptable,” and above 0.7 are typically considered “good.”

¹⁸ There is one item for this this is not true. Our results are robust to excluding this item from the analysis. This “item-total” correlation is used in Classical Test Theory to ensure that each item is measuring the same construct as the other items included. Details on these correlations are available from the authors on request.

first receive the scholarship, which allows us to capture longer-run effects than most evaluations of the impact of school-related interventions capture.

Our analytical sample consists of Grade 4 students selected for follow up who were offered a poverty- or merit-based scholarship, or would have been offered a poverty- or merit-based scholarship had they attended a Phase 1 school (Table 1). In total, 570 and 525 treated students in the merit and poverty groups respectively were randomly selected for follow up. The firm found 440 students in the merit group and 436 students in the poverty group. The firm received a list of 579 merit control students and 531 poverty control students to follow; they found, respectively, 421 and 378 students (merit attrition rate 25%; poverty attrition rate 23%). Finally, the household survey collected information from 26 and 18 individuals in replacement households for the merit and poverty treatment groups, and 53 in each of the control groups. All in all, the sample we analyze includes 940 students in the merit group and 885 students in the control group. Analytical issues related to attrition, and adjustments for it, are discussed below.

4. Empirical strategy

Empirical strategy

We estimate a reduced-form model relating the program to enrollment and attendance outcomes, test scores, and other outcome/process indicators.

We estimate separate and pooled models. For poverty-based treatment, the separate estimation is based on the equation

$$Y_{i,t1} = \beta_0 + \beta_1 T_{i,s}^{Sch} + X'_{i,t0} B + \eta_{i,s,t1} \quad (1)$$

where $Y_{i,t1}$ denotes the value of the outcome variable (e.g., school attainment) for individual i at follow-up ($t1$); $X'_{i,t0}$ is a vector of control variables measured at baseline (e.g., household characteristics); and $\eta_{i,t1}$ captures unobserved student characteristics and idiosyncratic shocks. $T_{i,s}^{Sch}$ is equal to one if the student was offered a poverty-based scholarship, and zero for students in the control schools. An analogous model is estimated for the merit-based scholarships. In each case, the sample consists only of (a) recipients in treatment schools and (b) individuals who would have been recipients of scholarships had they attended a treatment school of that type:

$$T_{i,s}^{Sch} = \begin{cases} 1 & \text{if } Index_i \geq Index_{50 Perc, Scholarship School} \\ 0 & \text{if } Index_i \geq Index_{50 Perc, Control School} \end{cases}$$

where $Index$ refers to the baseline poverty index in case of the poverty-based scholarship or the baseline test score in the case of the merit-based scholarship. The pooled model combines the two samples (hereafter “pooled regression”) and includes both treatment dummy variables:¹⁹

$$Y_{i,t1} = \beta_0 + \beta_{Pov} T_{i,s}^{Pov} + \beta_{Merit} T_{i,s}^{Merit} + X'_{i,t0} B + \eta_{i,s,t1} \quad (2)$$

The control variables ($X'_{i,t0}$) included in the estimation consist of an indicator for gender; the number of children 14 and under in the household; indicators for whether the household owns a motorcycle, a car/truck, an ox/buffalo, a pig, an ox or buffalo cart; indicators for whether the house has a hard roof, a hard wall, a hard floor, an automatic toilet, a pit toilet, electricity, piped water; and the poverty index and test scores.²⁰ Note that since we are relying on randomized

¹⁹ We thank an anonymous referee for suggesting this specification.

²⁰ We use dummy indicative variables to impute values for a very small number of missing observations in the control variables. Given the random nature of the program, the imputation does not make any substantive difference in the estimation, although standard errors are slightly smaller since sample size is maximized.

assignment for identification, the inclusion of control variables should not affect the estimate of impacts. We include them because they may potentially increase the precision of the estimates.²¹

Given that the treatment variable identifies, at baseline, all individuals who were offered the scholarship, the estimator is an intent-to-treat estimator (ITT).²² Errors are clustered at the school level, and each model includes district-level fixed effects.

Before turning to the results, we present validity checks to rule out effects driven by attrition and imbalance on observable characteristics across treatment and control groups.

Attrition

Recall that the endline analysis is based on a household survey administered to a random subsample of scholarship applicants. Annex A presents, for the analytical sample, differences in baseline characteristics between non-attriters and attriters and between non-attriters and students coming from replacement households. As mentioned above, overall attrition is 23%. Four out of 16 variables are statistically different between attriters and non-attriters. There are no discernable differences between non-attriters and replacement students, or between attriters and replacement students.

²¹ Additionally, we ran different specifications of the pooled model: first, a model that included as covariate a dummy indicating joint controls of poverty and merit treatments; second, a fully saturated model, without a constant term, that includes both treatment variables, a dummy variable that is one only for merit controls, a dummy variable that is one only for poverty controls, and a dummy variable that is one for the joint controls. Results, available from the authors on request, are very similar to the ones presented here.

²² Our analysis is based on an ITT approach starting from the official list of students which maps one-to-one to a list we generate independently from the baseline data. Note that official recording forms for scholarship receipt were generated centrally and pre-populated with student names (and space for a thumb print on collection). If a student had dropped out and could not collect the scholarship the funds could not be reassigned to another student but would be returned to a central fund for use in a subsequent distribution round. This limits the potential for reallocation of scholarships at the school level. We treat as recipients all those who were initially offered a scholarship, whether or not they kept the scholarship for all three years.

More important, however, is to establish that the difference in characteristics between attritors and non-attritors is not systematically different between treatment groups. Columns 1 and 2 of Table 2 report the means of household baseline characteristics for attrited students in control and treatment groups respectively, while columns 3 and 4 report the analogous means for the non-attrited sample, control and treatment (standard deviations in parentheses). Column 5 reports the estimate (standard errors in parentheses) and statistical significance of the “double-difference.” This is derived from a regression of each baseline characteristic on a dummy variable for attrition, a dummy variable for treatment, and the interaction of the two. A statistically significant estimate would suggest that the characteristics of attritors is different by treatment group.

The attrition rates for the two types of scholarships are very similar to the overall rate (23% poverty scholarship, 25% merit scholarship). The table shows balance between attritors and non-attritors across treatment groups. Out of the 32 double-difference estimates, only 2 are statistically significantly different from zero: the probability of being a girl (for poverty) and presence of hard floor in the dwelling (for merit). These findings are consistent with pure chance, and we therefore do not think that attrition is a likely driver of our results. In addition, we test whether double-differences are jointly equal to zero (based on estimating a seemingly unrelated regression model). For both treatments, we fail to reject the null hypothesis of coefficient equal to zero jointly (the p-values of the test are 0.57 and 0.32 for poverty and merit treatment respectively). While these results suggest that selective attrition is not a problem, we nevertheless check that our estimates are robust to how we handle attrition, and discuss those findings in the results section.

Baseline balance and characterization of the study sample

This section presents the general characteristics of the study sample and describes the validation of the experimental design. Table 3 presents average characteristics for students in the control and treatment groups, by type of scholarship. It includes only treated students in treatment schools —“treatment”, Columns (2) and (5)—and untreated students in control schools who would have been eligible for treatment—“control”, Columns (1) and (4)—based on their poverty index score or baseline test score, had they attended a treatment school. Columns (3) and (6) report the estimated difference (with standard errors in parenthesis) between treatment and control applicants by fitting a regression model of each baseline characteristic against the treatment dummy variable (robust standard errors are clustered at the school level). We also include predicted values for the baseline test score and poverty index, based on an OLS regression of each variable against the other covariates in the table.²³

Overall, characteristics differ very little between treatment and control groups. Only a few coefficients in Columns (3) and (6) are statistically significant: of the 32 differences reported, 5 are statistically significant. The table also reports a joint test for equality of means. We fail to reject the joint hypothesis of equality of means between treatment and control groups at conventional levels of significance (p-value of 0.29 for the poverty-based scholarship and 0.40

²³ As a further check, we used controls’ information at baseline to create predicted values of each of the main outcomes of interest for the whole sample (treated and control). Over these predicted values we ran tests of differences across treatment and control groups. We fail to reject the hypothesis of equality for all the variables. Results are available from the authors on request.

for the merit-based scholarship). The results from these tests confirm the validity of the random assignment.^{24,25}

An important feature of the data is that the distributions of the poverty index and the test score are statistically indistinguishable between the treatment and control groups. Figures 2 and 3 present the density of poverty index and the test score at baseline for treatment and control schools. There is a clear overlap throughout the whole distribution. A Kolmogorov-Smirnov test does not allow us to reject equality in both cases.

Table 3 also indicates that the applicants who are offered scholarships on the basis of merit tend to own more assets, have a lower poverty index score, and performed better on the baseline test than students offered a poverty-based scholarship. For instance, the poverty index, which ranges from 0 (wealthiest family in the sample) to 292 (poorest family in the sample), has a mean of 245.1 for (control) students targeted based on poverty and 218.2 for (control) students targeted based on merit. Likewise, the baseline test (ranging from 0 to 25) has a mean of 19.77 for (control) merit-based students and 17.74 for (control) poverty-based students. We return to this issue below when we discuss the tradeoffs between the two targeting approaches. The main finding from Table 3, however, is that the samples are balanced at baseline, allowing us to identify casual effects.

5. Results

²⁴ Given that the randomization was performed at the school level, we also analyzed the balance of baseline's characteristics at this level between treatment and control (e.g., all students in treated schools in comparison to all students in control schools—not just those who did, or would have, received a scholarship). The results of this exercise (not presented here) are very similar to the ones reported in Table 2.

²⁵ To the extent that differences in these variables is cause for concern, this is mitigated by their inclusion as control variables in the regressions.

Impacts on enrollment and attendance

The intervention is aimed at directly incentivizing higher enrollment and attendance rates, so scholarship students must stay enrolled, attend school regularly, and maintain passing grades until they graduate from primary school (6th grade) to continue receiving scholarships. We focus on three enrollment and attendance proxies: the probability of reaching 6th grade, the highest grade completed, and the hours of school attended in the past seven days.

Table 4 reports the program’s impact on these enrollment and attendance proxies. Column (1) presents the estimates for the poverty-based intervention (this is coefficient β_1 in Equation 1) using only the poverty sample and Column (2) presents the analogous estimates for the merit-based intervention, using only the merit sample (separate regressions). Column (3) presents the results from the pooled regression and the tests for different impacts across scholarship types (β_{POV} and β_{Merit} in Equation 2). All regressions control for the baseline characteristics presented in Table 3—including baseline test score and poverty index—and district fixed effects; robust standard errors, clustered at the school level, are reported.

The mean of the control group for each outcome variable is a reference point for assessing the magnitude of the impacts (these means are presented in the Table): 59.2% of the control poverty students reported reaching at least 6th grade, and the average grade completion for that group was 5.37 grades. The equivalent numbers for the merit control group were 61.8% and 5.44. The third outcome variable is based on the number of hours students had attended school during the past seven days, conditional on being enrolled. On average, students reported having

attended school for about 8.767 hours (the poverty-targeted control group) and 9.215 hours (the merit-targeted control group) in the past week.²⁶

Given that the separate regression (Columns 1 and 2) and the pooled regression show very similar results (Column 3), we will discuss only the estimates from the pooled sample. Impacts on all enrollment and attendance variables were significant and positive in the poverty-targeted treatment. Poverty-based scholarship recipients are about 19.1 percentage points more likely to reach Grade 6 than control students and they complete just over 0.34 more grades. The estimates for hours in school have large standard errors. The point estimate suggests that poverty-based scholarship recipients spend around 2.75 hours more than the control group, but this estimate is not statistically significant.²⁷ Impacts are similar for the merit-based intervention: students in the merit-targeted treatment group are 13.7 percentage points more likely to reach Grade 6 than control students; they complete about 0.225 more grades than control students; and they attend about 0.74 additional hours of school compared to the control students (with this last coefficient not statistically significant). These impacts are comparable to those found in the context of the Lower Secondary School scholarship program, where enrollment increased on the order of 20–25 percentage points (Filmer and Schady, 2014). These impacts are larger than the majority of impacts documented in countries elsewhere in the world (Fiszbein and Schady, 2009), and they are also large when assessed against the small size of the transfer (US\$ 20 per year). The relatively large impacts are likely related to the very low counterfactual levels, suggesting

²⁶ In cases where the school was closed because classes had ended for the year, we assigned a value of zero, because we did not know exactly when the school year had ended. Given the randomization, we do not expect that school closures were systematically different in any of the groups studied.

²⁷ Since this indicator measures a different margin than the others we should perhaps not be surprised that the impacts could be slightly different.

substantial “room to grow.” In Mexico, for example, counterfactual enrollments at the primary school level for the Oportunidades intervention were above 90% (Schultz 2004).

Based on the estimates from the pooled sample, we cannot reject the null hypothesis that the impacts are equal between poverty and merit scholarships (p-values of 0.18 for probability of reaching Grade 6; 0.22 for highest grade completed; and 0.27 for hours in school respectively). In sum, there is evidence that the program increased school participation regardless of whether the scholarship was awarded based on merit or poverty, and the size of the impacts were similar across the two targeting schemes.

We carry out two types of robustness checks to assess the extent to which our results are sensitive to the approach taken to handling attrition. First we estimate the model excluding the “replacement” students and use only those on the original sampling list. The results (reported in the first row of the table in Annex B) are very similar to our original estimates. This is not surprising: the number of replacement students is low and the characteristics at baseline were very similar to non-replacement students (Annex A).

Second, we implement Lee bounds (following Lee 2009) starting with the sample that excludes replacement students (results are reported in the second and third rows of Annex B).²⁸ The bounds are quite tight for the probability of reaching Grade 6 and for the highest grade completed: For both scholarship types, both lower and upper bounds are positive and statistically significant. The estimates for hours in school are noisier and it is only for the upper bound

²⁸ The approach consists of trimming the data in the treatment or control group (whichever has the least amount of attrition). “Excess observations” are trimmed either from the top or the bottom of the distribution of outcomes (resulting in “upper” and “lower” bounds) to achieve equivalent attrition in both treatment and control groups.

estimates that there is a statistically significant impact (for both scholarship types). Lower bounds are negative, so a zero impact lies within the estimated bounds for this outcome.

Results on test scores

Table 5 presents impacts of the interventions on the mathematics and Digit Span tests. The table has the same structure as Table 4: it presents impact estimates from models after controlling for baseline characteristics, the first two columns by running the model separately for each treatment, and the third column by pooling the data. All achievement measures are standardized such that the control schools have a mean of zero and a standard deviation of one (when averaged across all students, not just those who serve as the counterfactual for the treatment group). Impacts can therefore be interpreted as changes in a standard deviation of the achievement measure.

In contrast to the results on enrollment and attendance, the impacts on test scores differ between recipients of poverty-based and merit-based scholarships. Poverty-based scholarships had no impact on test scores, whereas merit-based scholarships had positive impacts. The impacts of the poverty-based scholarships on the mathematics and Digit Span test scores are estimated to be very small (-0.045 to -0.056 standard deviations respectively; pooled regression), and they are not statistically significantly different from zero. On the other hand, the estimates of the merit-based scholarships are 0.172 standard deviations for the mathematics test and 0.148 standard deviations for the Digit Span test. We reject the null hypothesis of equality across targeting approaches with 95% confidence (p-values of 0.03 for math and 0.01 for Digit Span). We note that these test score impacts are similar in magnitude to those reported for the ex-post

merit-based scholarship program evaluated in Busia, Kenya (Kremer, Miguel, and Thornton, 2009).²⁹

We carry out the same robustness checks as for the school participation indicators. Impacts estimated on the sample that excludes replacement households are very similar: for the poverty treatment impacts are small, negative and not statistically significant; for the merit treatment impacts are positive and significant (size effects are 0.187 standard deviations for math and 0.151 standard deviations for Digit Span). Lee bounds are less precise in this case, but point to a qualitatively similar finding. For the poverty scholarship, the lower bound for both tests is negative and the upper bound positive (and only significant for the Digit Span test) which means that the bounds include zero. For the merit scholarship the lower bounds for both tests are positive (although not statistically significant) and the upper bound positive (both significant) which means that the bounds do not include zero. Despite the fact that the bounding exercise is less tight for test scores, and that the lower bound estimates for the impact of the merit scholarship are not statistically significant, we see these results as consistent with our main finding. That is, whereas both poverty- and merit-based scholarships incentivized students to acquire more schooling, only students who received merit-based scholarships showed any gains in academic achievement.

As discussed, the different results for recipients of poverty-based and merit-based scholarships could potentially derive from the different skills and endowments of individuals in these groups at baseline. We can test this hypothesis directly because some recipients of poverty-

²⁹ Note that it is possible that school resources are more “diluted” in phase 1 schools than in phase 2 schools at endline (because there are two recipient cohorts in the former and only one in the latter). The bias introduced, if this is the case, would be towards finding no impact on test scores (if more resources improve learning outcomes). Moreover, we would not expect such a bias to differ for poverty-based versus merit-based scholarship recipients.

based scholarships performed well enough on the baseline test to qualify for merit-based scholarships, had they been in a merit-based scholarship school. These high performers in the poverty-based treatment are “identical” to the high performers in the merit-based treatment in every respect but one: their scholarships were labeled “poverty” scholarships rather than “merit” scholarships. Absent any labeling effects, one would expect to find the same impacts on learning outcomes among these high achievers as among recipients of merit-based scholarships who are similarly poor. This is not what we find.

In order to estimate these heterogeneous effects, we estimate the following models. For each poverty school, we identify students that were below each school’s median baseline test score (hereafter “low test”) and similarly, for merit schools we identify students below each school’s median poverty index (hereafter “low poverty”). We estimate the pooled specification with each outcome variable against the following variables: a dummy variable for poverty treatment, a dummy variable for merit treatment, a dummy variable for “low test”, a dummy variable for “low poverty”, and two interaction variables: the interaction between poverty treatment and “low test,” and the interaction between merit treatment and “low poverty”. The coefficient on the dummy variable for poverty treatment is therefore the impact *for the above-average test scoring recipients*; the coefficient on the dummy variable for the merit treatment is the impact *for the above-average poverty recipients*. If these coefficients are not equal, then this means that merit scholarship for poor students have a different impact than poverty scholarships for high scoring students—i.e. that the differently labelled scholarships have different impacts on “identical”

students. In all specifications we include baseline controls to account for any imbalances in these subpopulations.³⁰

The results of estimating this model are reported in Table 6. As it is evident, we lose some power from trying to identify heterogeneous impacts among small subgroups of the population (and we therefore have correspondingly larger standard errors).

For the two main measures of schooling attainment, the probability of Grade 6 completion and the highest grade completed, the impacts of both treatments are positive and significant. The interaction effects are not statistically different from zero suggesting limited heterogeneity in impacts of both types of scholarships on these outcomes. Most importantly, the test fails to reject that poverty scholarships for high scorers have the same impact as merit scholarships among the poorest recipients (p-values for these tests are 0.22 and 0.60 for Grade 6 completion and for highest grade respectively). The results for school hours are qualitatively consistent, but these lack statistical power.

The results are different for the measures of academic achievement. Here we find that the impacts of the poverty scholarships are not statistically significant (with no significant difference between high or low test scoring recipients). On the other hand, the merit based scholarships do have a statistically significant impact (again with no significant differential impact among poorer or less poor recipients). For both measures of achievement, we statistically reject that the impact of the poverty scholarships on high test scoring recipients is equal to that of merit scholarships

³⁰ We tested for differences at baseline between low- and high-test/poverty index for poverty and merit treatments. For the majority of cases, we failed to reject the null hypothesis of equality. Results are available from the authors on request.

for high poverty recipients (p-values of 0.07 and 0.01 for math and for Digit Span respectively).³¹

In summary, we see these estimates as pointing to the notion that the labeling or framing of the scholarships shaped their impact. Absent such shaping one would expect high testing recipients who receive a poverty scholarship to have the same outcome as merit scholarship recipients who are poor. But poverty-based scholarships do not appear to have induced better test results among recipients who also had high test scores at baseline, whereas the merit-based scholarships did induce better test scores among all individuals, poor or not poor at baseline.³²

Regression Discontinuity Design Results

Because the scholarship offers are made according to strict criteria, applicants “just above” and “just below” the cutoff for eligibility could be compared using a regression discontinuity design (RDD) approach to evaluate program impact. The implementation of the selection rules was carried out in a way closely consistent with the intent: A comparison of the official list of scholarship recipients promulgated by the Ministry of Education with the list generated by the firm hired to carry out the scoring and ranking yields no discrepancies. Spot checks at a number of schools yielded no cases of manipulation of the selection process.

Table 7 presents results of the RDD strategy.³³ It is worth noting that the RDD procedure is a local estimation, whereas the results presented so far in the paper are average treatment effects.

³¹ As in the earlier robustness checks, we re-estimate the model excluding the replacement households and find the same pattern: we cannot reject equality of impacts for the school participation effects, while we do for the math and Digit Span tests. Results are available upon request.

³² As suggested by an anonymous referee, we also fit models for heterogeneous effects by gender, without finding any significant effect. Results are available on request.

³³ Barrera-Osorio et al (2014) presents the main tests that validate the use of the RDD estimation.

Moreover, in order to gain power, we augment the RDD analysis sample with data from the second cohort (when there was no longer a “pure” control group). Both of these reasons suggest caution while comparing the RCT and RDD results. We nevertheless believe that the RDD provide a useful additional check on the robustness of our findings.

RDD results are very similar to the RCT results. The point estimates for the (local) effect of merit treatment on highest grade completed is 0.51 (and statistically significant); the analogous estimate for the poverty treatment is 0.20 (not statistically significant). In contrast, the effects of treatment on mathematics are 0.04 (not significant) and 0.34 (statistically significant) for poverty and merit treatment respectively. These results were stable to different bandwidths and control functions.³⁴ These results corroborate the asymmetry found in the RCT estimation: positive results on enrollment for both treatments, but only positive effects on mathematics for the merit treatment.

6. Pathways and program efficiency

Pathways

Both types of scholarships studied here plausibly provide incentives for students or their families to increase effort. We use two measures, based on the data available from household interviews, to assess the program’s impact on the hours that students study outside of school, and the share of family expenditures going to education-related items such as textbooks.

³⁴ The results in the table are from a parametric estimation, with a bandwidth of +/- 15 points, controlling by linear forcing variable and linear interaction at both sides of the discontinuity. For brevity, we present the results for highest grade completed and mathematics. Barrera-Osorio et al (2014) presents results for alternative specifications.

As a reference point, students in both the poverty and merit control groups spent around 3.4 hours per week doing school tasks outside of school, and total household expenditures on education averaged approximately US\$ 21. Merit-based scholarship recipients spent more time doing academic work outside of school compared to the control group (an increment of 0.64 hours); the impact for poverty-based scholarship recipients is smaller (0.48 hours) and statistically insignificant (Table 7). However, we fail to reject the null hypothesis of equality of effects (p-value 0.65). Similarly, spending on education differs by type of scholarship received. Households with a merit-based scholarship recipient spent US\$ 4.54 more on education than control students' households, whereas households that received a poverty-based scholarship did not discernibly increase expenditures on education (point estimate of -1.529, not statistically significant). In this case, we reject the null hypothesis of equal effects (p-value of 0.06).

To illustrate the magnitude of the association between the measures of effort and achievement outcomes, we estimated a simple model that regresses each test score outcome against each effort variable (hours of study outside school and household expenditure in education). The coefficient estimate of hours of study outside the school on math results is 0.0425 (with a t-statistic of 3.34) and on Digit Span is 0.028 (t-statistic of 3.35). The coefficient estimate of household expenditure and math is 0.003 (t-statistic of 3.72) and for the Digit Span test is 0.0017 (t-statistic of 2.23).³⁵ In short, there is a statistically significant association between the outcomes and the measures of effort.³⁶

³⁵ We estimated heterogeneity effects for these mechanism (analogous to Table 6), but due to low power, the estimates are highly imprecise.

³⁶ We nevertheless acknowledge that these measures of effort may not be sufficient in magnitude to explain the test score differences. We think of these as proxies for a broader set of differences in effort which we do not directly observe.

In sum, it appears that students work harder, and families spend more on education, when a merit-based scholarship is awarded. No such impacts are seen among recipients of poverty-based scholarships.

We are unable to investigate a different student- and family-level pathway that may have led to differential impacts, namely different beliefs on the part of recipients and their families in the two study arms. As mentioned above, other programs have shown that the promise of a cash transfer conditional on high endline performance on a test have led to better test scores (Kremer, Miguel and Thornton 2009). It is possible that recipients of the merit scholarships in Cambodia believed that high performance was required to retain the scholarship. While this is a possibility, we do not have any empirical or anecdotal evidence to suggest that recipients believed this. Moreover, program socialization was the same in poverty- and merit-based schools, and program rules included the exact same conditionalities. Nevertheless, if such a process was indeed playing out, we do not see this as wholly inconsistent with our interpretation of the results. If labeling a scholarship as based on merit induced the mistaken belief that it was tied to future performance, and that mistaken belief led to better learning outcomes (perhaps through the effort pathways discussed above) then we would see this as supportive of our overall findings. This is, of course, different from explanations linked to concepts such as stereotype threat but it would be consistent with different behavioral responses to differently labelled transfers.

It is important to recognize that the scholarship program could influence the behavior of not just students and their families, but also of other actors such as teachers. For instance, altruistic teachers may provide more attention to students with poverty scholarships to help them retain their scholarships. It is also possible that teachers exert more effort with scholarship students if

parents pressure them to do so. Banerjee and Duflo (2006) discuss changes in teacher motivation arising from greater accountability to families in the context of a scholarship program. Last, teachers might put more effort into students who are labeled high performers (so-called “Pygmalion effects”; see Jacobson and Rosenthal 1968). As such, school actors (such as teachers) may also differentially change their behavior depending on how the scholarships are labelled.^{37,38} Our data do not allow an analysis of these supply-side effects: we note, however, that for these supply-side effects to drive our results they would need to play out differently for recipients of poverty-targeted as opposed to merit-targeted recipients. So while the pathway may be different from the “effort” effect discussed above, it nevertheless reflects a differential behavioral response to the labeling of the scholarships.³⁹

Program equity and efficiency

Given the two targeting approaches, the Cambodia pilot scholarship program is well suited to shed light on the potential tradeoff between efficiency—defined as achieving more per dollar transferred—and equity—defined as reaching the poorest members of the population.⁴⁰ Analyses

³⁷ There is little scope for explicit “tracking” of students across classes in these schools. At baseline, 90% of schools had only one Grade 4, Grade 5 and Grade 6 classroom suggesting limited scope for reallocating students across classrooms.

³⁸ The current consensus of the literature on the Pygmalion effect is that it exists but is small. Teachers also tend to correct information as the academic year progresses (Jussim and Harber, 2005)

³⁹ An additional pathway could be differential classroom dynamics/peer-effects induced by the alternative targeting schemes. If merit scholarships draw in higher achieving students whereas poverty scholarships draw in low achieving students, then the resulting classroom composition might be different in a way that affects learning through a channel that is not labeling (e.g. teachers responding differently to a different mix of students). While this is a possibility, and we cannot rule it out conclusively, we do not think that it is at play here. First, we find no learning impacts on non-recipient peers which suggests that this effect impact would only affect recipients—which seems inconsistent with the notion that it is affecting classroom dynamics as a whole. Second, the magnitude of our estimates suggests that scholarships result in an additional 2-3 students per classroom (out of 10-12) which we think is likely insufficient to substantively alter the classroom dynamics.

⁴⁰ In a separate exercise we investigated a different dimension of equity, namely the impact of the program on non-recipients within the same school. We find that the poverty-based intervention was associated with higher enrollments among non-recipient peers, but find no other substantive or significant impacts. The results are available from the authors on request.

of the socioeconomic profile of program applicants and recipients under the two targeting schemes, and comparisons to the national distribution of socioeconomic characteristics, suggest that both targeting approaches are heavily skewed to the poor. The first panel in Figure 4 shows that 50% of those who applied to the program are in the poorest nationally-benchmarked wealth quintile (70% in the poorest two quintiles); fewer than 3% of applicants were from the richest quintile. The fact that the program was targeted to poor areas and to schools serving high concentrations of poor children in Cambodia accounts for this pattern. Unsurprisingly, scholarships targeted to the poorest from each school yields a greater pro-poor distribution of benefits than when targeting by merit. In the poverty-targeted schools, 63% of applicants were from the poorest quintile of the population (85% were in the poorest two quintiles). Merit-based targeting is less pro-poor but is still able to reach the poorest groups in the population. In the merit-targeting schools 54% of applicants were from the poorest quintile of the population (and 76% from the poorest two quintiles).

The finding that merit-based targeting did not result in an overall regressive scheme is a reassuring result. Part of this may be due to the fact that there is an overall weak correlation between baseline test scores and baseline poverty among the targeted students. We estimate that, of the merit-based scholarship recipients, 56% would have received a scholarship had the school been a poverty-based one; of the poverty-based scholarship recipients, 52% would have received a scholarship had the school been a merit-based one.⁴¹

⁴¹ One might be concerned that the lack of a strong association between the poverty score and test score performance could be due to poor measurement of either or both of these measures. We do not believe this to be the case. The “poverty score” discriminates well between households that own assets that are associated with higher consumption levels and those that don’t; it correlates well with other aspects of monetary welfare—including consumption per capita at endline in the control group (the correlation coefficient is -0.13 with log per capita expenditures which is statistically significant with $p < 0.01$); and higher poverty predicts lower grade attainment among the control group. The baseline test score has a Cronbach alpha of 0.85 (which is considered high); there is a positive correlation

It is possible that the relatively narrow geographic targeting of the program led to this result: indeed, the locations for this intervention are remote and poor and there is relatively limited heterogeneity in student poverty. This raises the question of whether our results are generalizable. In other settings—for example, where there is greater heterogeneity in student poverty levels—the tradeoff between the alternative targeting strategies might be starker. Nevertheless, to the extent that the result is somewhat generalizable, the findings suggest that there may be ways to enhance both equity and efficiency in this type of program. Indeed, if poverty-based scholarships trigger lower motivation and effort (or other behaviors that affect learning outcomes) than a merit-based scholarship, not because of differences in the underlying skills of the population, but because of framing, this will affect how the equity-efficiency tradeoff is made. If the framing of merit-based scholarship triggers high motivation and self-esteem (or other behaviors) then this targeting mechanism can induce efficiency gains.

7. Conclusions

The fact that some students were able to take better academic advantage than others from additional school exposure highlights an issue rarely addressed in previous evaluations of CCT programs. Recent evidence on monetary incentives for schooling shows that students are able to change their behavior on the margins that are under their control—for example, enrollment and attendance. These positive effects do not necessarily translate into test score gains, however. For example, despite the fact that Mexico’s Oportunidades program—a rigorously evaluated CCT program—induced students to enroll and attend additional years of school, the program did not

between the correct answer to each item on the test and the overall test score, and a negative correlation between each incorrect answer and the overall test score; and it correlates with other measures of performance—including math and Digit Span test scores at endline in the control group (the correlation coefficients are 0.12 and 0.11 with math and Digit Span respectively, both of which are statistically significant with $p < 0.01$).

induce better test scores. A recent set of papers has argued that education systems in developing countries are typically tailored towards better-off and better-skilled students. Specifically, Glewwe, Kremer, and Moulin (2009) show that only the strongest students at baseline were able to take advantage of textbooks that were provided to schools in Kenya. Furthermore, Duflo, Dupas, and Kremer (2011) show that teachers who were assigned to students at the bottom of the achievement distribution were less likely to teach.

The most important finding of our study is the asymmetry of response to the two targeting mechanisms. Both poverty-based and merit-based targeting schemes induce higher enrollment and attendance, yet only the merit-based mechanism induces positive effects on test scores, and the merit-based students (and their households) exhibit a higher level of effort in education. These results are not driven by differences in baseline skills across recipients of the two types of scholarships. Poorer students who are academically prepared do not gain significantly in skills and knowledge (measured by test scores) from the additional schooling they receive as a result of the simple poverty-targeted incentive. Clearly more work is needed—in Cambodia and elsewhere—to establish the best approach to ensure that additional schooling translates into learning. Remedial lessons for students in the early grades, or increasing school readiness among poorer students, for example through early child development programs, might be approaches to try. Indeed, data from Cambodia suggest that children suffer from substantial delays in cognitive development, which hamper their school readiness (Naudeau et al., 2011).

The experience evaluated in this paper suggests another approach, based on the finding that incentivizing school attendance in a way that recognizes academic potential can pay off in measurable learning outcomes. By changing the framing of the program, and the associated

labeling of the recipient, the merit-based scholarships distributed by the program led to better learning outcomes. We are able to explore two pathways for how this might have worked. We find that merit-scholarship recipients and their families appear to have been motivated to exert greater effort: in terms of hours spent studying and in terms of education expenditures. Poverty-based recipients did not appear to have been motivated in a similar way.

We recognize that our results are subject to a number of caveats. An ideal “labeling” experiment would have been to randomly assign scholarships and randomly assign labels within schools. This is not our approach. We interpret the fact that impacts on learning are different for recipients with “identical” characteristics, but for whom the labeling of the scholarship differs, as evidence of “labeling” effects. This is perhaps a broad understanding of the term—it encompasses behaviors of a number of agents. We document two specific behaviors of two such agents—students and their families—but there are others that we cannot—such as beliefs about scholarship retention or teacher behaviors. Understanding what is driving the different responses, and assessing the response of teachers to these different targeting approaches, would be useful for future research to shed light on.

Our findings nevertheless suggest that it should be possible to frame demand-side incentive programs in a way that maximizes their impact on learning outcomes. Such an approach could be undertaken in a way that does not necessarily imply a stark tradeoff between equity and efficiency. Indeed, the results show that among poorer children who received merit-based scholarships, the impacts on school participation as well as test scores were large. Scaling up an approach that targets students with an incentive that recognizes their high academic potential—

while ensuring that the poorest student are among that set—is likely to maximize both the equity and efficiency objectives of the program.

References

- Angrist J. and V. Lavy. 2009. "The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial" *American Economic Review*, 99:4, pp. 1384-1414
- Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. 1999. "When white men can't do math: Necessary and sufficient factors in stereotype threat." *Journal of experimental social psychology*, 35(1), 29-46.
- Baird, S., Ferreira, F.H.G., Özler, B. and Woolcock, M. (2013) Relative effectiveness of conditional and unconditional cash transfers for schooling outcomes in developing countries: a systematic review. *Campbell Systematic Reviews* 2013: 8, pp. 1–124. DOI: 10.4073/csr.2013.8.
- Baird, Sarah, Craig McIntosh and Berk Özler, 2011. "Cash or Condition? Evidence from a Cash Transfer Experiment," *The Quarterly Journal of Economics*, 126(4), pages 1709-1753.
- Banerjee, Abhijit, and Esther Duflo. 2006. "Addressing Absence." *Journal of Economic Perspectives*, 20(1): 117–132.
- Bardhan, P. 1996. "Efficiency, equity and poverty alleviation: policy issues in less developed countries." *The Economic Journal*, 106(438), pages 1344-1356.
- Barham, T., Macours, K., and Maluccio, J. A. 2013. "More schooling and more learning? Effects of a Three-Year Conditional Cash Transfer Program in Nicaragua after 10 years." IDB Working Paper Series no IDB-WP-432.

- Barrera-Osorio, F., D. Filmer and J. McIntyre (2014) "Randomized Controlled Trials and Regression Discontinuity Estimations: An Empirical Comparison" processed, Harvard University.
- Behrman, Jere R., Susan W. Parker, and Petra E. Todd. 2005. "Long-Term Impacts of the Oportunidades Conditional Cash Transfer Program on Rural Youth in Mexico." Discussion Paper 122, Ibero-America Institute for Economic Research, Göttingen, Germany.
- Behrman, Jere R., Susan W. Parker, and Petra E. Todd. 2011 "Do Conditional Cash Transfers for Schooling Generate Lasting Benefits? A Five-Year Followup of PROGRESA/Oportunidades." *Journal of Human Resources*, 46, no. 1: 93-122.
- Behrman, Jere R., Piyali Sengupta, and Petra Todd. 2000. "The Impact of PROGRESA on Achievement Test Scores in the First Year." Unpublished manuscript, International Food Policy Research Institute, Washington, DC.
- Bettinger, Eric. 2010 "Paying to Learn: The Effect of Financial Incentives on Elementary School Test Scores," NBER Working Paper 16333.
- Carlsson, M., Dahl, G. B., & Rooth, D. O. (2012). *The effect of schooling on cognitive skills* (No. w18484). National Bureau of Economic Research.
- Cascio, E. U., & Lewis, E. G. (2006). Schooling and the Armed Forces Qualifying Test Evidence from School-Entry Laws. *Journal of Human Resources*, 41(2), 294-318.
- Chaudhury, Nazmul and Dilip Parajuli. 2008. "Conditional Cash Transfers and Female Schooling: The Impact of the Female School Stipend Programme on Public School Enrolments in Punjab, Pakistan." *Applied Economics*, 42(28): 3565-3583.

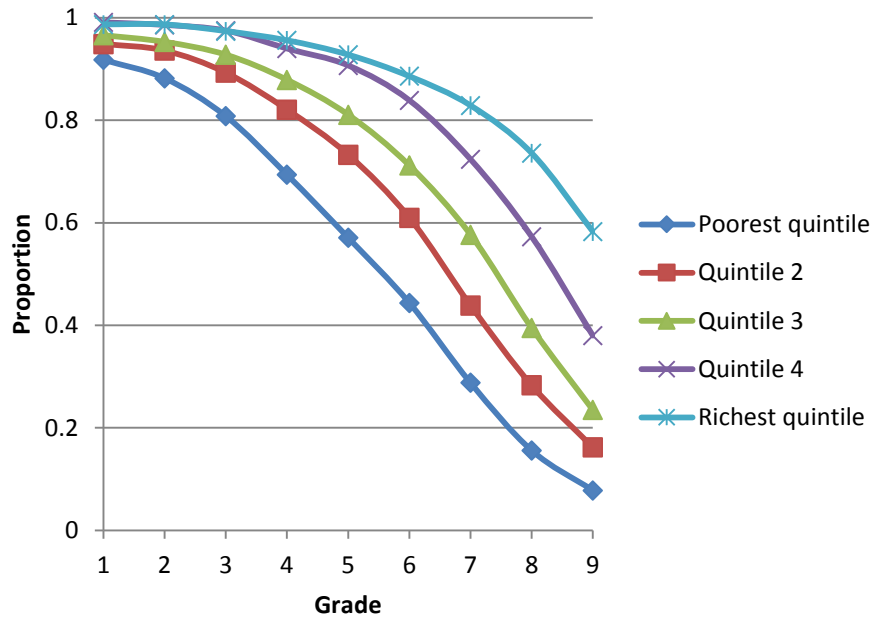
- Coady, D., Grosh, M., and Hoddinott, J. 2002. *The targeting of transfers in developing countries: Review of experience and lessons*. Social Safety Net Primer Series, World Bank, Washington DC.
- Deci, E. L., Koestner, R., and Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, 71(1), 1-27.
- Duflo, Esther, Pascaline Dupas and Michael Kremer. 2011. "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya." *American Economic Review*. 101(5): 1739-74.
- Filmer, Deon, and Norbert Schady. 2008. "Getting Girls into School: Evidence from a Scholarship Program in Cambodia." *Economic Development and Cultural Change*, 56(2): 581–617
- Filmer, Deon and Norbert Schady. 2009. "School Enrollment, Selection and Test Scores." World Bank Policy Research Working Paper No. 4998. The World Bank, Washington, DC.
- Filmer, Deon and Norbert Schady. 2011. "Does more cash in conditional cash transfer programs always lead to larger impacts on school attendance?" *Journal of Development Economics*. 96(1): 150–157.
- Filmer, Deon and Norbert Schady. 2014. "The Medium-Term Effects of Scholarships in a Low-Income Country." *Journal of Human Resources*. 2014. 49(3):663-694.
- Fiszbein, Ariel and Norbert Schady. 2009. *Conditional Cash Transfers: Reducing Present and Future Poverty*. The World Bank. Washington, DC
- Fryer Jr., Roland G. 2011. "Financial Incentives and student Achievement: Evidence from randomized trials." *The Quarterly Journal of Economics* 126, 1755–1798.

- Glewwe, Paul, Michael Kremer and Sylvie Moulin. 2009. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics*. 1(1): 112-135.
- Hanushek, E. A., & Woessmann, L. (2008). The role of cognitive skills in economic development. *Journal of economic literature*, 607-668.
- Hoff, K. and P. Pandey. 2006. "Discrimination, Social Identity, and Durable Inequalities" *American Economic Review*, Vol 96, No. 2, pp 206-211
- Independent Evaluation Group. 2011. "Evidence and Lessons Learned from Impact Evaluations on Social Safety Nets." Washington, DC, United States: World Bank.
- Jussim, L., & Harber, K. D. (2005). "Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies." *Personality and Social Psychology Review*, 9(2), 131-155.
- Kazianga, Harounan, Damien de Walque and Harold Alderman. 2012. "Educational and Child Labour Impacts of Two Food-for-Education Schemes: Evidence from a Randomised Trial in Rural Burkina Faso." *Journal of African Economies*. 21(5): 723-760.
- Kremer, Michael, Edward Miguel, and Rebecca Thornton. 2009. "Incentives to Learn." *Review of Economics and Statistics*. 91(3): 437-456.
- Krishnaratne, Shari, Howard White and Ella Carpenter. 2013. "Quality education for all children? What Works in Education in Developing Countries" 3ie Working Paper No. 20.
- Lee, D. S. 2009. "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects." *Review of Economics and Statistics* (76): 1071-1102.

- Lepper, M. R., D. Greene, and R. Nisbett. 1973. "Undermining Children's Intrinsic Interest with Extrinsic Rewards: A Test of the "Overjustification" Hypothesis" *Journal of Personality and Social Psychology*, Vol. 28, No. 1, pp. 129-137
- McEwan, Patrick. Forthcoming. "Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments." *Review of Educational Research*.
- Murnane, Richard J. and Alejandro J. Ganimian. 2014. "Improving Educational Outcomes in Developing Countries: Lessons from Rigorous Evaluations." National Bureau of Economic Research Working Paper No. 20284. <http://www.nber.org/papers/w20284>.
- Naudeau, Sophie, Sebastian Martinez, Patrick Premand, and Deon Filmer. 2011. "Cognitive Development among Young Children in Low-Income Countries" in Alderman, Harold ed. *No Small Matter: The Impact of Poverty, Shocks, and Human Capital Investments in Early Childhood Development*. The World Bank. Washington, DC.
- Orfield, Gary, "Foreword," in Donald E. Heller and Patricia Marin (Eds.), *Who Should We Help? The Negative Social Consequences of Merit Aid Scholarships* (2002) (Papers presented at the conference "State Merit Aid Programs: College Access and Equity" at Harvard University). Document found in <http://civilrightsproject.ucla.edu/research/college-access/financing/who-should-we-help-the-negative-social-consequences-of-merit-scholarships/>
- Royal Government of Cambodia. 2006. Student Achievement and Education Policy: Results from the Grade Three Assessment—Final Report. Cambodia Education Sector Support Project—National Assessment Component. Phnom Penh, Cambodia.

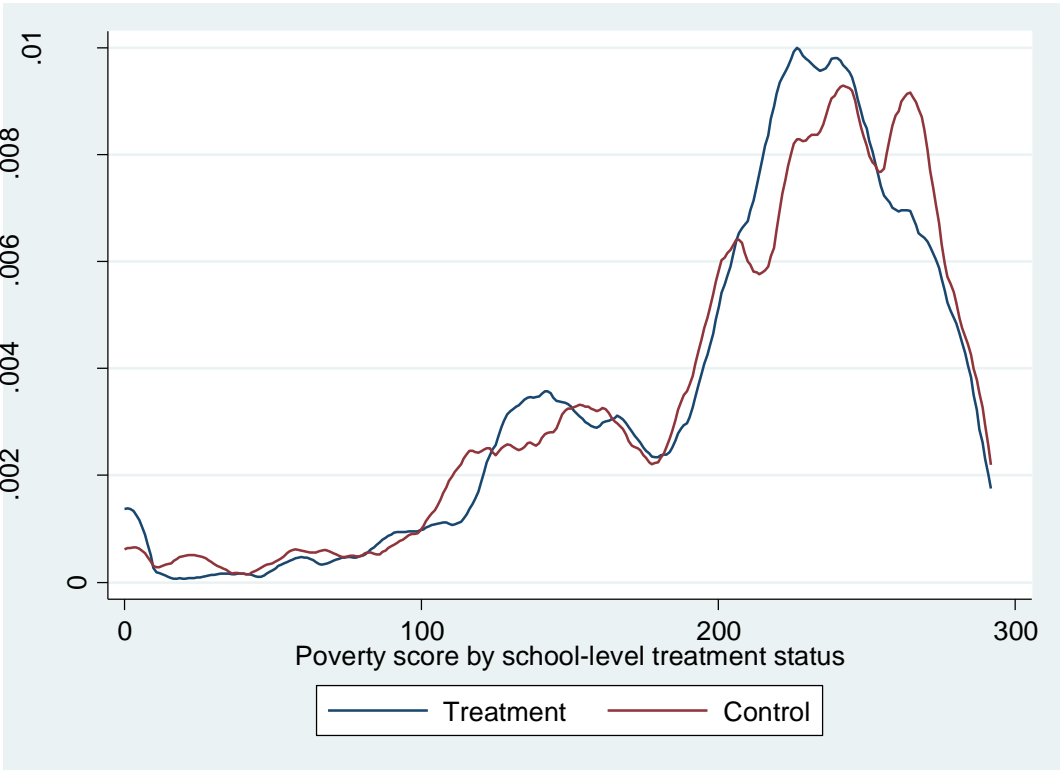
- Saavedra, Juan Esteban and Sandra Garcia. 2013. "Educational Impacts and Cost-Effectiveness of Conditional Cash Transfer Programs in Developing Countries: A Meta-analysis". University of Southern California Center for Economic and Social Research Working Paper No. 2013-07. cesr.usc.edu.
- Schmader, T., Johns, M., and Forbes, C. 2008. "An integrated process model of stereotype threat effects on performance." *Psychological Review*, 115(2), 336.
- Schultz, T. Paul. 2004. "School subsidies for the poor: evaluating the Mexican Progresa poverty program." *Journal of Development Economics*. 74:199-250.

Figure 1: Proportion of 15 to 19 year olds who have completed each grade, by quintile.



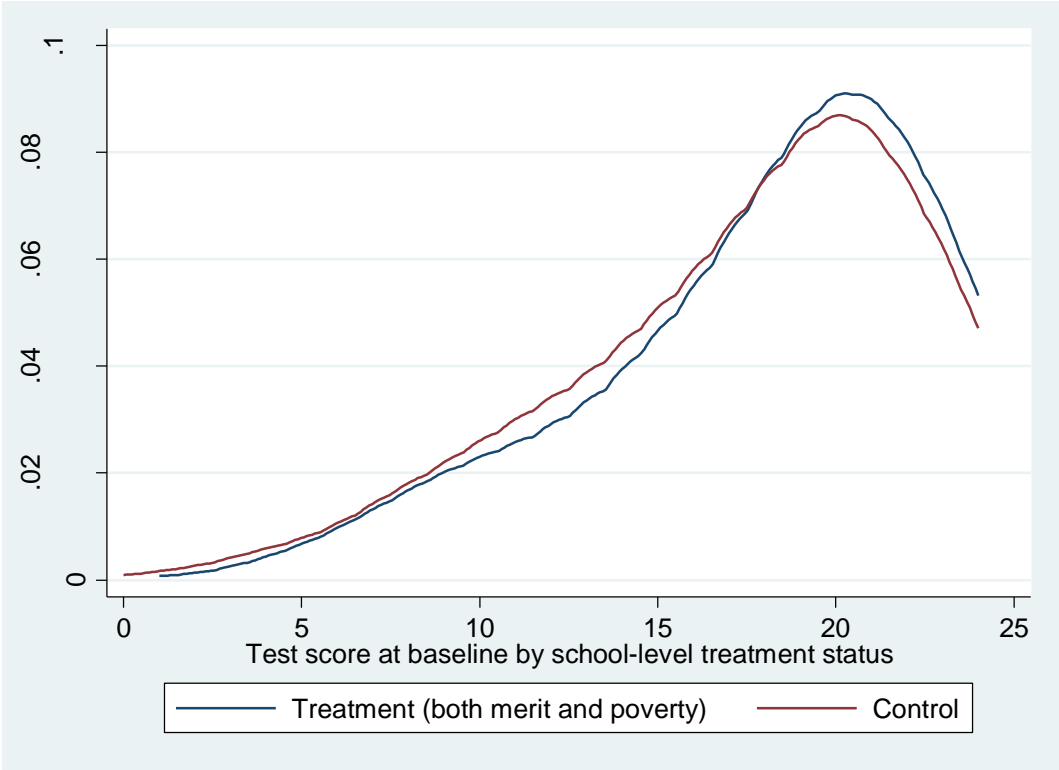
Source: Authors' analysis of data from Cambodia Demographic and Health Survey 2010.

Figure 2: Poverty score at baseline, Treatment versus Control



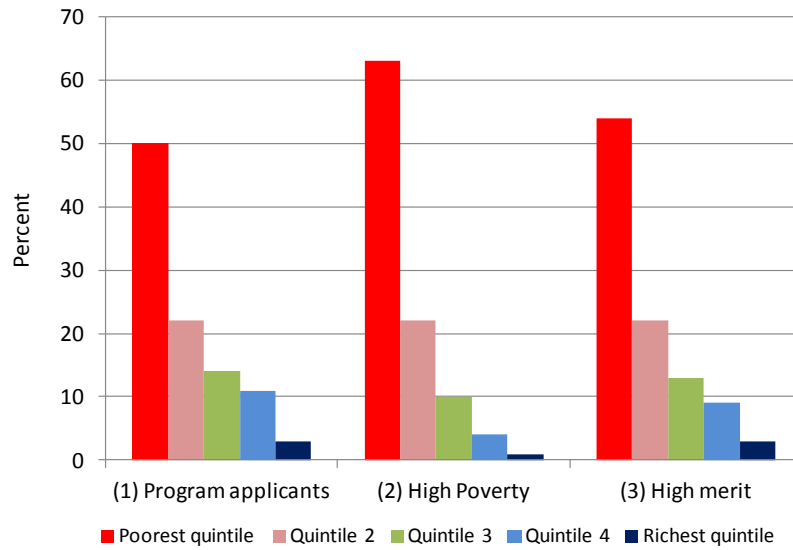
Source: Analysis of baseline application forms.

Figure 3: Test scores at baseline, Treatment versus Control



Source: Analysis of baseline Math and Khmer language tests.

Figure 4: Distribution of selected populations across nationally benchmarked quintiles



Source: Analysis of Cambodia DHS 2010 and Primary Scholarship Application forms. Quintiles are defined on the basis of an index of household wealth-related variables that are collected in both the DHS 2010 as well as on the scholarship program application forms.

Table 1. Analytical Sample

	Total	Merit Schools ("Phase 1")	Poverty Schools ("Phase 1")	Non-Treated Schools ("Phase 2")	
		Merit Treatment Students	Poverty Treatment Students	Merit Control Students	Poverty Control Students
A. Assignment of schools to groups (February 2008, not public)	207 schools	52 schools	51 schools	104 schools	
B. Baseline (December 2008/January 2009)					
Grade 4 students in school	5902	1496	1650	1208	1584
Number of students actually treated	1708	835	873	0	0
C. Follow-up (May-September 2011)					
Sample given to survey to follow-up	2952	570	525	579	531
Non-attriters	2274	440	436	421	378
Attriters	678	130	89	158	153
Replacement	174	26	18	53	53
Final analytical sample of students		466	454	474	431

Table 2. Attrition analysis, analytical sample

Variable	Poverty scholarship					Merit scholarship				
	Attritor control	Attritor treatment	Non-	Non-	DD	Attritor control	Attritor treatment	Non-	Non-	DD
			attritor control	Attritor treatment				attritor control	Attritor treatment	
(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	
Girl	0.47 (0.50)	0.42 (0.50)	0.52 (0.50)	0.62 (0.49)	-0.153** (0.08)	0.39 (0.49)	0.41 (0.49)	0.48 (0.50)	0.46 (0.50)	0.043 (0.07)
No of minors	1.9 (1.11)	1.86 (1.15)	1.78 (1.13)	1.86 (1.07)	-0.119 (0.19)	1.74 (1.16)	1.5 (1.04)	1.72 (1.12)	1.64 (1.11)	-0.167 (0.18)
Own motorcycle	0.25 (0.44)	0.26 (0.44)	0.28 (0.45)	0.24 (0.43)	0.045 (0.07)	0.33 (0.47)	0.38 (0.49)	0.42 (0.49)	0.42 (0.49)	0.057 (0.08)
Own car/truck	0.03 (0.18)	0.03 (0.18)	0.04 (0.20)	0.05 (0.21)	-0.006 (0.03)	0.11 (0.31)	0.12 (0.32)	0.13 (0.34)	0.17 (0.37)	-0.026 (0.05)
Own oxen/buffalo	0.37 (0.49)	0.39 (0.49)	0.38 (0.49)	0.5 (0.50)	-0.107 (0.10)	0.46 (0.50)	0.46 (0.50)	0.53 (0.50)	0.57 (0.50)	-0.044 (0.08)
Own pig	0.46 (0.50)	0.44 (0.50)	0.42 (0.49)	0.56 (0.50)	-0.157 (0.10)	0.5 (0.50)	0.47 (0.50)	0.55 (0.50)	0.58 (0.49)	-0.053 (0.07)
Own ox or buffalo cart	0.16 (0.37)	0.24 (0.43)	0.18 (0.39)	0.26 (0.44)	0.004 (0.09)	0.21 (0.41)	0.24 (0.43)	0.29 (0.45)	0.3 (0.46)	0.016 (0.07)
Hard roof	0.29 (0.46)	0.39 (0.49)	0.33 (0.47)	0.37 (0.48)	0.053 (0.08)	0.36 (0.48)	0.42 (0.49)	0.47 (0.50)	0.58 (0.49)	-0.05 (0.09)
Hard wall	0.38 (0.49)	0.38 (0.49)	0.39 (0.49)	0.42 (0.49)	-0.034 (0.09)	0.48 (0.50)	0.6 (0.49)	0.56 (0.50)	0.56 (0.50)	0.115 (0.07)
Hard floor	0.84 (0.37)	0.83 (0.38)	0.79 (0.41)	0.83 (0.37)	-0.054 (0.07)	0.85 (0.36)	0.83 (0.37)	0.83 (0.37)	0.91 (0.29)	-0.094** (0.05)
Have automatic toilet	0.01 (0.11)	0.02 (0.15)	0.02 (0.14)	0.01 (0.08)	0.022 (0.02)	0.05 (0.23)	0.06 (0.25)	0.06 (0.23)	0.05 (0.22)	0.018 (0.04)
Have pit toilet	0.08 (0.27)	0.16 (0.37)	0.12 (0.32)	0.13 (0.34)	0.062 (0.06)	0.09 (0.29)	0.1 (0.30)	0.14 (0.34)	0.14 (0.35)	-0.002 (0.05)
Electricity	0.19 (0.39)	0.17 (0.38)	0.17 (0.38)	0.15 (0.36)	-0.002 (0.06)	0.25 (0.43)	0.22 (0.42)	0.24 (0.43)	0.23 (0.42)	-0.01 (0.07)
Pipe water	0.03 (0.18)	0.03 (0.18)	0.03 (0.16)	0.02 (0.13)	0.012 (0.03)	0.04 (0.20)	0.06 (0.24)	0.05 (0.23)	0.05 (0.21)	0.032 (0.04)
Poverty Index (o to 292)	248.24 (28.30)	245.24 (35.84)	244.83 (33.32)	241.94 (33.78)	-0.109 (7.48)	222.34 (55.83)	216.33 (54.02)	218.06 (50.90)	207.42 (65.50)	4.63 (10.49)
Test score (0 to 25)	17.37 (4.96)	18.33 (4.99)	17.87 (4.67)	18.64 (4.84)	0.187 (0.84)	19.57 (3.45)	19.81 (3.41)	19.84 (3.14)	19.85 (2.84)	0.234 (0.55)
Number of students	153	89	378	436	1056	158	130	421	440	1149
Attrition rate					0.23					0.25
Joint test, all coefficient equal zero:										
Chi-square					14.332					18.068
P-Value					0.57					0.32

Note: Analytical sample, Grade 4 selected for follow up. Columns (1), (2), (3), and (4) present mean and standard deviations. Columns of differences (5) shows the differences by regressing each covariate against attrition dummy, treatment dummy and the interaction, clustering by school id. Joint test using Seemingly Unrelated Regressions (SUR).

Table 3. Baseline characteristics and balance, analytical sample

	Student Level					
	Poverty			Merit		
	Control (1)	Treatment (2)	Difference (3)	Control (4)	Treatment (5)	Difference (6)
Girl	0.52 (0.50)	0.61 (0.49)	0.099*** (0.03)	0.49 (0.50)	0.46 (0.50)	-0.03 (0.04)
No of minors	1.79 (1.12)	1.86 (1.07)	0.073 (0.12)	1.73 (1.12)	1.63 (1.10)	-0.097 (0.12)
Own motorcycle	0.28 (0.45)	0.24 (0.43)	-0.034 (0.05)	0.42 (0.49)	0.42 (0.49)	0.003 (0.05)
Own car/truck	0.04 (0.19)	0.05 (0.21)	0.01 (0.03)	0.13 (0.34)	0.16 (0.37)	0.028 (0.04)
Own oxen/buffalo	0.39 (0.49)	0.49 (0.50)	0.109* (0.06)	0.53 (0.50)	0.56 (0.50)	0.033 (0.06)
Own pig	0.43 (0.50)	0.55 (0.50)	0.119** (0.06)	0.55 (0.50)	0.58 (0.49)	0.029 (0.05)
Own ox or buffalo cart	0.19 (0.40)	0.25 (0.43)	0.057 (0.05)	0.29 (0.45)	0.3 (0.46)	0.009 (0.05)
Hard roof	0.32 (0.47)	0.37 (0.48)	0.05 (0.05)	0.48 (0.50)	0.58 (0.49)	0.102** (0.05)
Hard wall	0.38 (0.49)	0.43 (0.50)	0.045 (0.06)	0.55 (0.50)	0.57 (0.50)	0.018 (0.05)
Hard floor	0.79 (0.41)	0.83 (0.38)	0.037 (0.05)	0.84 (0.37)	0.91 (0.29)	0.068* (0.04)
Have automatic toilet	0.02 (0.13)	0.01 (0.08)	-0.01 (0.01)	0.05 (0.22)	0.06 (0.23)	0.005 (0.02)
Have pit toilet	0.11 (0.32)	0.13 (0.34)	0.02 (0.03)	0.13 (0.34)	0.14 (0.34)	0.001 (0.04)
Electricity	0.16 (0.37)	0.15 (0.36)	-0.011 (0.04)	0.23 (0.42)	0.23 (0.42)	-0.002 (0.05)
Pipe water	0.03 (0.17)	0.02 (0.12)	-0.012 (0.01)	0.06 (0.23)	0.04 (0.20)	-0.013 (0.02)
Poverty Index (o to 292)	245.13 (32.73)	242.19 (33.75)	-2.938 (5.13)	218.2 (51.66)	206.42 (66.29)	-11.771 (8.76)
Predicted Poverly Index	248.9 (27.93)	246.48 (28.35)	-2.419 (3.91)	226.88 (43.85)	222.72 (45.08)	-4.161 (5.47)
Test score (0 to 25)	17.74 (4.71)	18.62 (4.84)	0.876 (0.68)	19.77 (3.22)	19.8 (2.88)	0.028 (0.48)
Predicted Test score	18.66 (0.58)	18.63 (0.58)	-0.028 (0.07)	18.43 (0.67)	18.43 (0.69)	-0.001 (0.09)
Number of students	431	454	885	474	466	940
Joint test, all coefficient equal zero:						
Chi-square						18.878
P-Value						0.399

Table shows average (and standard deviations) in students characteristics of the analytical sample. Columns (1), (2), (4), (5): means and standard deviation. Columns (3), (6): difference between control-treatment, estimated by regressing each variable against corresponding treatment variable. Robust standard errors clustered at the school level are presented in parentheses. Predicted poverty index/text is done by regression the index / test against all the other 14 variables at baseline. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively. Joint test using Seemingly Unrelated Regressions (SUR).

Table 4. Impact on enrollment

	Reach Grade Six			Highest Grade Completed			Hours in school, last 7 days		
	Poverty (1)	Merit (2)	Pooled (3)	Poverty (1)	Merit (2)	Pooled (3)	Poverty (1)	Merit (2)	Pooled (3)
Treatment Poverty	0.180*** (0.04)		0.191*** (0.04)	0.338*** (0.11)		0.340*** (0.10)	2.872 (1.88)		2.755 (1.69)
Treatment Merit		0.122*** (0.04)	0.137*** (0.04)		0.186* (0.10)	0.225** (0.10)		0.674 (1.55)	0.742 (1.56)
Control mean (sd)	0.592 (0.49)	0.618 (0.49)	0.6 (0.49)	5.372 (1.23)	5.444 (1.24)	5.394 (1.24)	8.767 (12.94)	9.215 (13.21)	8.779 (12.89)
Ho: Coef Poverty=Coef Merit, F(Stat) Prob			1.78 0.184			1.531 0.217			1.217 0.271
Covariates	✓	✓	✓	✓	✓	✓	✓	✓	✓
No. Obs	833	897	1485	833	897	1485	666	712	1194
F()	3.624	2.495	2.924	2.28	1.806	1.978	1.223	2.038	1.435
R2 Adj	0.17	0.146	0.134	0.146	0.123	0.104	0.131	0.199	0.138

Regression coefficient of dependent variable against treatment indicator. Column (1): poverty sample; Column (2): merit sample; Column (3): Pooled regression, all controls. All regressions controlling for Table 1 baseline characteristics and district fixed effects. Robust standard errors clustered at the school level are presented in parentheses. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively.

Table 5. Impact on test scores

	Mathematics			Digit Span		
	Poverty (1)	Merit (2)	Pooled (3)	Poverty (1)	Merit (2)	Pooled (3)
Treatment Poverty	-0.019 (0.08)		-0.045 (0.08)	-0.055 (0.07)		-0.056 (0.08)
Treatment Merit		0.195* (0.10)	0.172* (0.09)		0.152** (0.08)	0.148** (0.07)
Control mean (sd)	0.029 (1.00)	0.14 (1.04)	0.073 (1.02)	0.014 (0.98)	0.08 (0.99)	0.029 (1.01)
Ho: Coef Poverty=Coef Merit, F(Stat) Prob			4.81** 0.03			6.57** 0.011
Covariates	✓	✓	✓	✓	✓	✓
No. Obs	885	940	1570	885	940	1570
F()	3.298	3.453	3.692	2.722	2.178	2.357
R2 Adj	0.19	0.17	0.15	0.114	0.094	0.08

Regression coefficient of dependent variable against treatment indicator. Column (1): poverty sample; Column (2): merit sample; Column (3): Pooled regression, all controls. All regressions controlling for Table 1 baseline characteristics and district fixed effects. Robust standard errors clustered at the school level are presented in parentheses. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively.

Table 6: Impacts on test scores by baseline test score and poverty index

	Dependent variable:				
	Completed grade 6	Highest grade completed	School hours	Mathematics	Digit Span
Treatment Poverty (T_Poverty)	0.184*** (0.05)	0.288** (0.12)	2.89 (1.96)	-0.045 (0.10)	-0.132 (0.10)
Treatment Merit (T_Merit)	0.122** (0.05)	0.230** (0.11)	1.092 (1.76)	0.188* (0.11)	0.155* (0.09)
Below baseline poverty mean (Low_Poverty)	-0.045 (0.04)	-0.105 (0.11)	2.169 (1.53)	-0.091 (0.10)	-0.021 (0.09)
Below baseline test score mean (Low_Test)	0.085* (0.05)	0.155 (0.11)	-1.198 (1.53)	-0.01 (0.09)	-0.057 (0.10)
T_Poverty*Low_Test	0.016 (0.06)	0.138 (0.16)	-0.176 (2.73)	-0.002 (0.12)	0.191 (0.13)
T_Merit*Low_Poverty	0.028 (0.05)	-0.033 (0.12)	-0.685 (2.30)	-0.037 (0.13)	-0.039 (0.12)
<i>Test that impact of poverty scholarship for high scorers is different from impact of of merit scholarship for high poverty recipients</i>					
Ho: T_Poverty = T_Merit: F-Stat	1.536	0.282	0.767	3.396*	6.975***
Prob	0.217	0.596	0.382	0.067	0.009
Covariates	✓	✓	✓	✓	✓
No. Obs	1485	1485	1194	1570	1570
F()	2.688	1.815	1.397	3.727	2.199
R2 Adj	0.137	0.108	0.141	0.157	0.082

Regression coefficients of dependent variable against treatment indicator, baseline indicators for above the median poverty or tests and interactions. Pooled regression. All regressions controlling for Table 1 baseline characteristics and district fixed effects. Robust standard errors clustered at the school level are presented in parentheses. ***, **, * indicates significance at the 1%, 5% and 10% levels

Table 7. Regression Discontinuity Design: main effects

	Highest Grade Completed		Mathematics	
	Poverty (1)	Merit (2)	Poverty (1)	Merit (2)
Treatment Poverty	0.2 (0.15)		0.04 (0.11)	
Treatment Merit		0.51* (0.28)		0.34* (0.19)
No. Obs	302	384	302	384

Parametric estimation; bandwidth: +/- 15 percentiles. Control function: linear plus linear interaction. Source: Barrera-Osorio, Filmer and McIntyre (2014)

Table 8. Impact of program on schoolwork (outside of school) and expenditures on education

	Time doing homework, studying or private lessons, last week			Total yearly expenditure in education, hh (US \$)		
	Poverty (1)	Merit (2)	Pooled (3)	Poverty (1)	Merit (2)	Pooled (3)
Treatment Poverty	0.4 (0.38)		0.484 (0.35)	-1.12 (3.07)		-1.529 (2.67)
Treatment Merit		0.700** (0.33)	0.643* (0.34)		5.734* (3.21)	4.538 (3.17)
Control mean	3.404 (3.92)	3.358 (3.55)	3.42 (3.86)	21.232 (42.63)	20.529 (34.27)	21.073 (40.14)
Ho: Coef Poverty=Coef Merit, F(Stat) Prob			0.211 0.647			3.74* 0.055
Covariates	✓	✓	✓	✓	✓	✓
No. Obs	600	656	1086	623	683	1130
F()	7.406	3.374	3.635	2.15	1.587	1.77
R2 Adj	0.145	0.176	0.128	0.125	0.155	0.106

Regression coefficient of dependent variable against treatment indicator. Column (1): poverty sample; Column (2): merit sample; Column (3): Pooled regression, all controls. All regressions controlling for Table 1 baseline characteristics and district fixed effects. Robust standard errors clustered at the school level are presented in parentheses. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively.

Annex A. Attrition and household replacement, analytical sample

Variable	Baseline			Differences		
	Non-Attritors	Attritors	Replacement Households	Attritors- Non	Replacement- Non attritors	Replacement- Attritors
	(1)	(2)	(3)	(2)-(1)	(3)-(1)	(3)-(2)
Girl	0.52 (0.50)	0.42 (0.49)	0.5 (0.50)	-0.099*** (0.03)	-0.022 (0.05)	0.077 (0.05)
No of minors	1.74 (1.10)	1.73 (1.11)	1.71 (1.02)	-0.008 (0.07)	-0.027 (0.10)	-0.02 (0.11)
Own motorcycle	0.35 (0.48)	0.33 (0.47)	0.38 (0.49)	-0.018 (0.03)	0.031 (0.05)	0.05 (0.05)
Own car/truck	0.11 (0.31)	0.08 (0.27)	0.09 (0.29)	-0.025 (0.02)	-0.019 (0.03)	0.007 (0.03)
Own oxen/buffalo	0.52 (0.50)	0.44 (0.50)	0.49 (0.50)	-0.081** (0.03)	-0.03 (0.05)	0.05 (0.05)
Own pig	0.55 (0.50)	0.47 (0.50)	0.54 (0.50)	-0.077** (0.03)	-0.007 (0.05)	0.07 (0.05)
Own ox or buffalo cart	0.27 (0.44)	0.22 (0.42)	0.3 (0.46)	-0.05 (0.03)	0.027 (0.04)	0.077 (0.05)
Hard roof	0.46 (0.50)	0.37 (0.48)	0.49 (0.50)	-0.086*** (0.03)	0.028 (0.05)	0.114** (0.06)
Hard wall	0.5 (0.50)	0.47 (0.50)	0.51 (0.50)	-0.028 (0.03)	0.01 (0.05)	0.038 (0.05)
Hard floor	0.85 (0.35)	0.84 (0.37)	0.83 (0.38)	-0.016 (0.02)	-0.029 (0.04)	-0.012 (0.04)
Have automatic toilet	0.04 (0.19)	0.04 (0.20)	0.04 (0.19)	0.007 (0.01)	0 (0.02)	-0.007 (0.02)
Have pit toilet	0.13 (0.34)	0.11 (0.31)	0.08 (0.27)	-0.022 (0.02)	-0.052* (0.03)	-0.029 (0.03)
Electricity	0.2 (0.40)	0.22 (0.41)	0.2 (0.40)	0.015 (0.03)	-0.001 (0.04)	-0.016 (0.05)
Pipe water	0.04 (0.19)	0.05 (0.22)	0.04 (0.20)	0.013 (0.01)	0.007 (0.02)	-0.006 (0.02)
Poverty Index (o to 292)	224.64 (52.39)	228.98 (49.86)	222.22 (58.93)	4.345 (3.73)	-2.414 (4.98)	-6.759 (5.77)
Test score (0 to 25)	18.92 (4.14)	18.62 (4.47)	18.06 (4.56)	-0.295 (0.31)	-0.862** (0.44)	-0.566 (0.44)
Number of students	1450	437	120	1887	1570	557
Attrition rate				0.28		

Note: Analytical sample, Grade 4 selected for follow up. Columns (1), (2), and (3) present mean and standard deviations. Columns of differences (2-1; 3-1; 3-2) shows the differences by regressing each covariate against respective dummy variable, clustering by school id

Annex B. Robustness checks: Estimation without replacement students and Lee bounds

	Reach grade 6		Highest grade completed		School hours		Mathematics		Digit Span	
	Poverty (1)	Merit (2)	Poverty (1)	Merit (2)	Poverty (1)	Merit (2)	Poverty (1)	Merit (2)	Poverty (1)	Merit (2)
Treatment	0.197*** (0.05)	0.149*** (0.05)	0.361*** (0.12)	0.263** (0.12)	3.2 (1.99)	1.491 (2.11)	-0.041 (0.09)	0.187* (0.11)	-0.05 (0.08)	0.151* (0.08)
Lee Lower bound	0.166*** (0.04)	0.135*** (0.03)	0.223*** (0.08)	0.205** (0.09)	-2.523 (1.67)	-2.707 (1.57)	-0.312 (0.08)	0.106 (0.10)	-0.361 (0.08)	0.043 (0.10)
Lee Upper bound	0.301*** (0.05)	0.196*** (0.04)	0.617*** (0.10)	0.398*** (0.12)	6.148*** (1.58)	3.274** (1.37)	0.091 (0.08)	0.281*** (0.09)	0.178** (0.08)	0.225** (0.09)
No. Obs	764	824	764	824	611	652	814	861	814	861
Trimming Portion	0.128	0.062	0.128	0.062	0.207	0.151	0.143	0.058	0.143	0.058
Tight bounds	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Treatment: regression coefficient of dependent variable against treatment indicator without controls; sample WITHOUT household replacement. Lower and Upper bound: Lees bounds. ***, **, * indicates significance at the 1%, 5% and 10% levels respectively.