Policy Research Working Paper 9081

Effects of a Multi-Faceted Education Program on Enrollment, Equity, Learning, and School Management

Evidence from India

Clara Delavallade
Alan Griffith
Rebecca Thornton

WORLD BANK GROUP
Africa Region
Gender Innovation Lab
December 2019

Policy Research Working Paper 9081

Abstract

The Sustainable Development Goals set a triple educational objective: improve access to, quality of, and gender equity in education. This paper documents the effectiveness of a multifaceted educational program, pursuing these three objectives simultaneously, in rural India. Using an experiment in 230 schools, the paper measures the effects of the program on students' school participation and academic performance over two years, while also examining heterogeneous impacts and sustainability. The findings show that the program increased enrollment, especially among girls (8.1 percent in the first year, 11.7 percent in the second), reducing gender gaps in school retention. The findings show large learning gains of 0.323 standard deviation due to the program in the first year and 0.156 standard deviation at the end of the second year, which did not vary by gender. There were also large effects on school management outcomes, increasing the number of meetings by 16 percent and the number of improvement plans completed by 25 percent.

This paper is a product of the Gender Innovation Lab, Africa Region. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted atcdelavallade@worldbank.org.

Effects of a Multi-Faceted Education Program on Enrollment, Equity, Learning, and School Management: Evidence from India

Clara Delavallade, The World Bank
Alan Griffith, University of Washington
Rebecca Thornton, University of Illinois[1]

JEL classification: O15, I25, J16, O19

Keywords: education, randomized experiment, school effectiveness, learning, gender equality

[1] Delavallade: Africa Gender Innovation Lab, World Bank, 1818 H Street, NW Washington, DC 20433, US (email: cdelavallade@worldbank.org), Griffith: Department of Economics, University of Washington, Savery Hall, 410 Spokane Ln, Seattle, WA 98105, US (email: alangrif@uw.edu), Thornton: Department of Economics, University of Illinois at Urbana-Champaign, 214 David Kinley Hall, 1407 W. Gregory, Urbana, IL 61801, US (email: rebeccat@illinois.edu). The authors thank the members of the NGO field team for their cooperation and contributions in all stages of the project and would especially like to acknowledge the useful comments provided by many seminar participants. Oxana Azgaldova, Kwong-Yu Wong, and Flor Paz provided valuable research assistance. All errors are our own.

1. Introduction

The United Nations' Sustainable Development Goal 4 proposes that by 2030 "all girls and boys [should] complete free, equitable, and quality primary and secondary education" (United Nations 2015). This ambitious goal sets a triple objective for educational policies: improving free access to, equity in, and quality of learning. It is unlikely that one single intervention is enough to make headway on these objectives simultaneously. For example, policies that are successful in increasing access to school may not necessarily foster quality education, and interventions that are effective among girls, may not be equally effective for boys. However, while education programs implemented by governments or NGOs often involve multiple inputs and programmatic objectives, we know very little about their effectiveness. The vast majority of education evaluations study programs that focus on only one objective or educational input – for example, only 5 percent of the studies in McEwan's 2015 meta-analysis of primary school evaluations measure the impact of a "combination" treatment.

In contrast, this paper presents the results from an evaluation of a multifaceted educational program designed and implemented by an India-based nongovernmental organization (NGO) in rural Rajasthan, India. The program had three primary objectives: to enroll marginalized girls, improve student learning, and support school management. The program involved multiple interventions: door-to-door campaigns to enroll drop-out or never-enrolled girls, volunteers trained to teach activity-based and playful learning among students grouped by ability, strengthening school management committees, and working with members of the community to promote girls' education. The multiple inputs were designed to complement and support the objectives of the program. By grouping students according to ability, the volunteer-led teaching activities could help teachers cater to students of varying abilities and mitigate potential harmful effects from the enrollment drive. Similarly, the program's community engagement and sensitization to the

importance of girls' education could mitigate potential negative effects of targeting girls for enrollment, rather than focusing on both genders.

To evaluate this program, we use a cluster-randomized experiment in 230 primary schools and individual-level panel data on enrollment, retention, attendance, and test scores in English, Hindi, and Math. We examine whether the program met its objectives: improving enrollment – especially of girls, increasing learning – across all abilities, and supporting school management committees. To measure whether the program was successful at targeting – to girls and students of lower-ability – we examine how the program differentially affected girls (vs. boys), and initially low-performing students (vs. higher performing). We also discuss threats that may arise from differential retention and enrollment.

We find large and statistically significant positive impacts on student enrollment – especially among girls – in the two years of the program. Further, these effects are large, representing increases of 8.1% and 11.7% in the first and second year of the program, respectively. While estimated effects are larger for girls, the differences between girls and boys are not statistically significant. Further, we find no statistically significant effects on the types of students (i.e., high vs. low ability) who enroll in program schools.

In terms of learning, we find large positive gains in all subjects, among both boys and girls. In the first year of the program, the impacts on post-program test scores are highest: 0.323 standard deviation across all subjects (0.317 in Hindi, 0.256 in English, and 0.369 in Math). Students in treatment schools are also 22.7 percentage points more likely to improve their scores in the first year when comparing pre-program with post-program tests. There are no significant differences in the effect of the program on learning gains by student gender. Second year post-program impacts are similar, although somewhat smaller, at 0.156 SD (0.127 in Hindi, 0.159 in English, and 0.136

in Math), but the difference between the post-program tests across the two years is not statistically significant (p=0.226).

Tests were also administered early at the beginning of the school year of the second year of the program – after the first year of programming, but prior to program implementation in the second year. We find significantly lower learning gains on this test among students who were exposed to the program in the prior year, suggesting that either the program resulted in teachers "teaching to the test" (Glewwe, Ilias and Kremer 2010), or in a loss in learning gains over school breaks, similar to summer vacation losses in the United States (Cooper et al., 1996).

We also examine how the intervention affected learning across students of varying initial academic performance and find that, at least in the first year of the program, gains increased as one moves upward in the ability distribution. This is similarto the findings in Bulh-Wiggers et al. (2019b), who also find larger effects of an education program among those of highest ability. This speaks to a related literature that finds larger sources of inequality in learning across dimensions other than gender – for example, across ability, or wealth (Crouch and Rolleston 2016, Kaffenberger and Pritchett 2017).

Lastly, we find that the program led to more School Management Committee meetings, as well as a larger number of prepared and completed improvement plans. In particular, the program increased the number of meetings held by 16%, the number of prepared improvement plans by 22% and the number of completed improvement plans by 25%.

The effects we find on learning are about twice as large as evaluations of other learning-targeted interventions that focus on one single educational input, such as using contract or volunteer teachers (average effect size = 0.10SDs) or training teachers (average effect size = 0.12SDs; McEwan 2015).

Our results speak to the effectiveness of multi-pronged interventions that aim to improve access without sacrificing quality of education. Many education programs implemented in developing countries involve a combination of interventions and target multiple issues at the same time. For example, among all registered voluntary and non-governmental organizations in India, only 16% that self-identify as working within the Education and Literacy sector operate exclusively within that sector.[2] Of those operating exclusively within the Education and Literacy sector, very few operated using a single programmatic strategy.[3] Similarly, of the four main federal "School Education & Literacy Schemes" for primary education in India, only one, the Mid- day Meal Program, focuses on one single intervention.[4]

In contrast to the breadth of activities, approaches, and objectives that are commonly implemented to improve education in developing countries, a majority of the studies of education programs involve measuring the effect of one type of intervention in isolation. McEwan (2015) studied 76 RCTs and finds only 5 percent evaluate a "combination" of treatments (authors' calculations).[5] In addition to RCTs, "high quality" studies involving regression discontinuity (RD) and difference in difference (DD) evaluations also tend to focus exclusively on programs with one intervention implemented in isolation. Damon et al. (2016) reviewed 39 "high quality" RD and

[2] Authors' Calculations. We extracted the list of NGOs/VOs through NGO-DARPAN at https://ngodarpan.gov.in Registered VOs and NGOs are required to sign-up online using the portal, run by the Planning Commission of India (PTI 2017). Accessed 12/3/2018.
[3] After coding all NGOs in the state of Rajasthan that focused exclusively on Education and Literacy we found 2.3% of NGOs focused on research or academia, 4.1% focused on giving financial support to students, 56.1% involved running a school or set of schools, 14.62% provided vocational training to youth or adults, 18.7% were coded as providing multiple types of education programs, and 4.1% were not known.
[4] See http://mhrd.gov.in/schemes-school.
[5] McEwan's literature search of randomized experiments conducted in developing country primary schools from the mid-1970s to 2013 resulted in the following studies that evaluated multiple treatment: He, Linden, and MacLeod (2008), Osendarp et al. 2007, Pradhan, Suryadarma, and Beatty (2011), and Ngyuen (2008).

DD studies. We coded each study to find that only 10.2 percent evaluate a combination of interventions.[6]

There are a large number of papers that have studied, in isolation, the specific interventions that are implemented in the program we evaluate. First, while we are unaware of any study that evaluates the effectiveness of an "enrollment drive," a sizable literature has evaluated various approaches to improve student enrollment (See JPAL Policy Bulletin 2017). Second, the activity-based learning among students grouped by ability is similar to "teach at the right level" or targeted interventions, that have seen overwhelming success across several settings (Banerjee, Cole and Duflo 2007, Banerjee et al 2010, Banerjee et al. 2016, Duflo, Dupas and Kremer 2011, Muralidharan Singh and Ganimian, 2018). Third, the use of volunteers to deliver their programmatic activities has also been evaluated, with mixed evidence (Banerjee, Cole, Duflo, and Linden 2007, Lakshminarayana et al. (2013), Banerjee et al. 2010, Torgerson, King and Sowden 2002). Lastly, the NGO's focus on school management and community engagement is similar to interventions that have evaluated community management or parent involvement (Barr et al. 2012, Lassibille et al. 2010, Gertler, Patrinos and Rodriguez-Oreggia 2012, Beasley and Huillery 2017, Banerjee et al 2010, Proadhan et al. 2014, and Blimbo, Evans and Lahire 2015, Glewwe and Maïga 2011).

This paper contributes to a number of additional literatures. For example, there is a growing body of evidence on the effectiveness of girl-focused interventions to address the gender gap in schooling. Evans and Yuan (2018) provide a review of these interventions but find that

[6] The papers contained in Damon et al. (2016) consist of 115 studies – 76 RCTs and 38 "High-Quality" regression discontinuity or difference in difference studies. The search involved studies conducted from 1990 to 2014, published in (peer-reviewed) academic journals from 1990 to 2014, and unpublished academic working papers written from 2010 to 2014 were included. Papers with multiple interventions include: Chey, McEwan, and Urquiola (2005), Gertler, Patrinos, and Rubino-Codina (2012), Kazinga et al. (2013), Santibanez, Abreu-Lastra, and O'Donoghue (2014).

interventions targeting girls result in no real advantage over education programs targeting both boys and girls.[7] While we find some differences in enrollment and retention gains among girls, there were no differences in the effect of the program on learning by gender. Kremer, Miguel, and Thornton (2009) found that boys benefited from a merit-based scholarship program even if they were not eligible themselves. In contrast, boys responded negatively to being excluded from a gender-based life skills program (Delavallade, Griffith, and Thornton, 2016).

Finally, our paper adds to the literature on the sustainability of treatment effects over time (Banerjee et al. 2007; Kremer and Miguel 2007; Andrabi, Das, and Khwaja 2008; Kremer, Miguel, and Thornton 2009; Duflo, Dupas, and Kremer 2011; Baird et al. 2016). Most evaluations of education programs conduct just one follow-up after the intervention to measure impact: McEwan (2015) found that the average number of follow-ups per experiment was just 1.4. Our multiple rounds of follow-up data help document some fade-out of the program effects on post-program tests in the second year, as well as document large declines in the learning gains on pre-program tests in the second year. These findings suggest there may be learning losses between academic years similar to the literature on summer vacation loss found in the United States (Cooper et al., 1996). We next present the background of the setting and education intervention in Section 2. Section 3 presents the research design, Section 4 presents the results, Section 5 discusses and presents additional robustness and Section 6 concludes.

[7] Some examples of particularly successful strategies for girls include programs that reduce the direct costs of schooling (Bruns, Mingat, and Rakatomalala 2003; Deininger 2003; Muralidharan and Prakash 2017), reduce the indirect costs and opportunity costs (Khandker, Pitt, and Fuwa 2003; Lavinas 2001), involve communities (Herz 2002; Benveniste and McEwan 2000), make schools girl-friendly (World Bank 2001; Herz 2002), and improve the quality of education (Lloyd, Mensch, and Clark 1998; Khandker 1996).

2. Background

2.1 Schooling in Rajasthan: Participation, Quality, and Gender Equity

Despite educational advances in most developing countries, the state of Rajasthan in India has experienced limited educational gains over the past decade, especially for girls (World Bank 2011). In 2012, 4.6 percent of girls 7–10 years old were still not in school in rural Rajasthan, compared with 2.2 percent of boys (Pratham Organization 2012). This gender gap widens considerably as students age, due largely to social norms that particularly disadvantage girls. Marriage is often seen as a substitute for schooling, and girls frequently have little say in when and whom they marry. In addition, marriage often occurs at a young age, with 57.6 percent of women marrying younger than the legal age (UNICEF 2012, 173).

In addition, educational quality is low, with only 47.7 percent of children in grades three to five able to read a grade one–level text in government schools in 2011, and only 33.1 percent able to do subtraction in 2012 (Pratham Organization 2012). The availability of primary schools in remote areas is still limited, leading to high variance in student-teacher ratios.

2.2 The Intervention

We evaluate an intervention developed and implemented by an Indian NGO working with government schools in the state of Rajasthan. One of its main aims is to increase girls' educational outcomes, with a focus on increasing school participation and learning in lower primary school (grades one through five). The program consists of several components that separately target enrollment and retention, learning, and school management. Each of the components of the program is directed by a trained volunteer in each village.

To target enrollment, the NGO identifies out-of-school girls before each school year, using information from community members and government records. The program volunteers hold village meetings to prepare for a house-to-house enrollment drive, targeting girls who have never enrolled or who have dropped out of school. These efforts seek both to encourage parents to support their daughters' education and to motivate girls themselves to come to school.[8]

To target student learning, the NGO organizes in-school lessons led by the volunteer in grades three through five. The curriculum and instructional model was designed with Pratham Rajasthan and emphasizes activity-based and playful learning through games that teach English, Hindi, and Math. The methodology emphasizes group work and student involvement in the teaching and learning process. These lessons are held during school hours for approximately two hours per day, several days per week, over four to five months. This component of the program does not focus explicitly on girls, but rather aims to increase learning levels for both girls and boys.[9] In tandem with the peer group learning method, students are placed in three groups according to ability, measured by diagnostic pre-program tests similar to the Annual Survey of Education Report (ASER); the tests are designed to be quick to administer so that they can be conducted individually for each student.

The program was implemented and evaluated in the academic years of 2012 and 2013. Each year, in selected villages, village volunteers conducted the door-to-door enrollment drive,

[8] The NGO identifies out-of-school girls (aged 6 to 14) using a two-step approach. Using data from the state government child tracking system (CTS) and school records, the program develops an initial list of girls to target. To prepare for the door-to-door enrollment drive, the program then engages community members to verify records, build awareness, and increase enrollment. Volunteers organize meetings of 20 to 40 individuals, to engage village leaders and community members as champions for increasing girls' school enrollment, and jointly identify barriers and develop solutions. Volunteers also leverage local meetings with other programs.

[9] The program involves a 12-week module focused on interactive teaching methods, "Catch Up" methodology for children who are behind their grade level, and peer group learning. After training, the program provides teachers with learning kits, which includes games and creative learning materials, and access to a telephone helpline and SMS updates for extra support. In schools with poorer performance, the NGO provides additional "handholding" visits to support teachers and the trained volunteer.

carried out the learning curriculum in grades three, four, and five, and assisted with strengthening School Management Committees and community support for girls' education. Appendix A presents additional details about the intervention and implementation.

3. Research Design

3.1 School Sample and Randomization

The study consists of 230 primary schools located in 98 villages in Rajasthan. In four administrative blocks, villages with at least one government primary school were selected for the study. In 2011, prior to the implementation of the program, researchers randomly assigned villages to either treatment or control groups (49 treatment and 49 control), stratified by Administrative Block, using a random number generator. This results in 117 treatment schools and 113 control schools. On average, there are 2.3 government primary schools per village.

3.2 Data and Outcome Measures

Baseline Data

We use data collected in 2011 – prior to the implementation of the program – to check for baseline balance and as covariates in the analyses. At the school level, we use school infrastructure data (such as the presence of electricity and computers and the number of students at each school) collected by the NGO staff. At the individual level, enrollment rosters collected in 2011—prior to program implementation—list each student's gender, grade, age and whether he or she belongs to a Scheduled Caste, Scheduled Tribe, or Other Backward Caste. We also construct school-grade-level enrollment for boys and girls in 2011 that we use as controls.

Learning

Learning is measured using tests conducted in school in 2012 and 2013, conditional on the student being enrolled and present on the day of the exam. In each year, two exams were administered: a pre-program test conducted just prior to, and a post-program test conducted just after, the learning component of the program was implemented.[10] The tests were based on ASER exams, which are standardized exams validated across India testing Hindi, English, and Math.[11] The exams are short, taking 5-10 minutes per student, testing the same skills in both years. For Hindi and English, students are tested on letters, words, a short paragraph, and a longer story, while math tests knowledge of single- and double- digit number recognition, two-digit subtraction with borrowing, and three-digit by one-digit division. Enumerators assess the highest level a student can comfortably perform. Following the ASER criteria, the tests are scored categorically from A (highest score) to E (lowest).

We construct two measures of learning from the exam scores. First, we normalized test scores to the control after assigning each letter grade a numeric value from one (E, lowest) to five (A, highest). Second, we create an indicator for whether the student's test score improved from pre-treatment to post-treatment. These two measures are motivated by the observation that, with the raw scores on a categorical scale, the difference between two adjacent scores may not be constant in terms of measuring learning outcomes. The indicator for a student's score increasing represents a way of looking at non-linear effects.

Tests were administered conditional on the student being present in school on the day of the exam. We address the possible bias due to differential test-taking in our analysis.

[10] In 2012, pre-program tests were administered in Nov-Dec 2012, while post-program tests were administered in March 2013. In 2013, pre-program tests were administered in Sept-Oct 2013, while post-program tests were administered in March 2014.
[11] See www.asercentre.org for more information about the ASER tool.

Enrollment

To measure enrollment, we use school enrollment rosters collected each year. From these data, we also construct school-grade-level total enrollment, by gender and pooled across girls and boys.

School Management

To understand the impact of the program on school management outcomes, we collected the number of School Management Committee (SMC) meetings held, and the number of School Improvement Plans (SIPs) prepared and implemented at each school. These were collected monthly from July 2012 to January 2013, for a total of seven observations per school.

3.3 Baseline Sample Students and Pre-Program Balance

Students who were enrolled in grades three and four at the beginning of 2011 (prior to the intervention) comprise our *Baseline Sample* (N=7,327). Table 1 presents average statistics and balance tests of pre-program student and school-level characteristics. On average, 55 percent of the students are girls, with the majority of students belonging to a scheduled caste (31 percent), scheduled tribe (15 percent) or other backward caste (40 percent). Less than half – 38 percent – of the students are in schools with electricity, few – 10 percent – have access to a computer, and just over 80 percent have access to clean drinking water.

Column 2 presents the regression coefficient testing the difference in means of pre-program variables between the treatment and control groups. We see no significant differences between treatment and control students across gender, caste, or school characteristics such as type of school (upper or lower secondary), having electricity, a computer, or drinking water, although students in treatment schools are younger on average. There are slightly fewer students enrolled in treatment

schools than in control either in all grades or within cohort, although the differences are not statistically significant. A joint test fails to reject equality of means by treatment group for all baseline variables in Table 1 (p=0.285 for all students pooled).

Table 1 -- Pre-program Characteristics of Students: Baseline Sample

	All		Girls		Boys	
	Mean in Control	Difference Treat - Control	Mean in Control	Difference Treat - Control	Mean in Control	Difference Treat - Control
	(1)	(2)	(3)	(4)	(5)	(6)
Girl	0.545	-0.016				
	(0.014)	(0.019)				
Grade 4 at Baseline	0.332	-0.009	0.334	0.006	0.331	-0.022
	(0.008)	(0.012)	(0.010)	(0.016)	(0.011)	(0.015)
Age	10.070	-0.324***	10.003	-0.260**	10.127	-0.377***
	(0.066)	(0.095)	(0.076)	(0.105)	(0.081)	(0.110)
Scheduled Caste	0.305	-0.014	0.323	-0.028	0.289	-0.002
	(0.030)	(0.038)	(0.034)	(0.045)	(0.029)	(0.036)
Scheduled Tribe	0.150	0.010	0.169	0.029	0.135	-0.008
	(0.022)	(0.036)	(0.027)	(0.046)	(0.022)	(0.032)
Other Backward Caste	0.400	-0.015	0.394	-0.041	0.405	0.008
	(0.034)	(0.052)	(0.040)	(0.059)	(0.032)	(0.050)
Upper Primary School (UPS)	0.582	-0.022	0.567	-0.008	0.594	-0.035
	(0.073)	(0.090)	(0.078)	(0.096)	(0.072)	(0.089)
Baseline Students Enrolled in Same Grade	23.003	-1.474	23.293	-1.653	22.762	-1.331
	(1.817)	(2.310)	(1.890)	(2.371)	(1.806)	(2.327)
Baseline Students Enrolled in All Grades	64.985	-4.913	66.066	-5.639	64.083	-4.328
	(5.318)	(6.575)	(5.633)	(6.843)	(5.174)	(6.509)
Electricity	0.396	0.014	0.386	0.026	0.404	0.005
	(0.066)	(0.095)	(0.065)	(0.097)	(0.069)	(0.096)
Computer	0.117	-0.023	0.107	-0.020	0.126	-0.026
	(0.041)	(0.054)	(0.040)	(0.054)	(0.044)	(0.057)
Drinking Water	0.806	0.022	0.803	0.017	0.808	0.027
	(0.059)	(0.076)	(0.065)	(0.082)	(0.057)	(0.074)
Teacher/Pupil Ratio	0.069	-0.002	0.070	-0.004	0.068	0.000
	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)	(0.009)

Notes: Robust standard errors in parentheses, clustered by village. For differences, *** p<0.01, ** p<0.05, * p<0.1. Baseline Sample consists of students enrolled in 2011 in grades 3 and 4. N=7,327 students in 230 schools.

3.4 Empirical Strategy

We measure the effects of the program in its first two years – 2012 and 2013 – on the program's three main objectives: enrollment of girls, student learning, and school management. We focus on students in grades three, four and five, who would have received the program in treatment schools in 2012 and 2013. To understand if the program met its objectives of enrolling girls, we report differential effects on school participation across gender. We then measure the

effect of the program on learning and explore whether the program had differential effects on girls and low-ability students. Lastly, we estimate effects of the program on school management outcomes.

Effect of the Program on Enrollment

To measure the effects of the program on enrollment, we test the difference in average grade-level enrollment by estimating the following:

$$Y_{gsj} = \beta_0 + \beta_1 T_j + \gamma' X_{gsj} + \epsilon_{gsj} \quad (1)$$

where Y_{gsj} identifies the number of students enrolled in grade g in school s in village j, in either 2012 or 2013. T_j indicates whether village j was assigned to the treatment group. The vector X_{gsj} is included in some specifications for statistical power and includes the number enrolled by gender prior to program implementation and cohort fixed effects. We estimate the equation with a linear probability model and cluster standard errors by village, the unit of randomization. We estimate the equation for all students and separately by gender.

Selection into Test-Taking

Before estimating the effect of the program on learning outcomes, we examine whether the program affected the type of students we observe for learning outcomes. Since test scores are only observed conditional on enrollment and attendance on the day of the exam, if the program was successful at targeting low-performing students, the estimates of the program's impact on test scores would be biased downward. Conversely, estimates may be biased upward if higher-performing students were more likely to be retained or attend on the test day in treatment schools.

To address potential threats due to differential test-taking, we examine the effects of the

program on pre-program and post-program test-taking in 2012 and 2013, as well on the likelihood of taking both tests, with Equation (1). In this context, Y_{gsj} is the number of students who were present and took the test. Using demographic data and pre-program test scores, we also characterize the types of students who are test-takers in the treatment and control groups.

Effect of the Program on Learning

Our sample for the analysis on learning outcomes includes students who took both pre-program and post-program tests in either 2012 or 2013. We estimate the effect of the program on exams in 2012 and 2013, pooling all subjects—Hindi, English, and Math—with the following regression:

$$(2) \qquad \boldsymbol{Y_{isjz} = \beta_0 + \beta_1 T_j + \gamma' X_{isj} + \epsilon_{isjz}}$$

for individual *i*, in school *s*, village *j*, on subject *z*. Learning, Y_{isjz}, is measured with either normalized test scores or with an indicator for whether the student's test score improved from pre-treatment to post-treatment in a given year. We include subject and cohort fixed effects as well as controls for school size prior to program implementation, captured by $\gamma' X_{isj}$ in the above equation. We cluster standard errors by village.

We run Equation (2) separately for pre-program and post-program test scores in each of 2012 and 2013. Recall that the pre-program test in 2012 was administered prior to the implementation of the learning component of the program and we would not expect any learning gains on this test. Pre-program tests in 2013 were administered prior to program implementation in 2013, but students in treatment schools who were enrolled in 2012 would have had exposure to the program. The results on 2013 pre-program tests indicate whether any learning gains experienced in year one were sustained into the beginning of year two.

Because all students—both boys and girls—received the in-class learning component of the program, there is unlikely to be an additional effect on girls' performance, unless girls respond differently than boys to the program. Still, we estimate differential effects of the program by gender with:

$$(3) \qquad \boldsymbol{Y_{isjz} = \beta_0 + \beta_1 T_j + \beta_2 Girl_{isj} + \beta_3 T_j \times Girl_{isj} + \gamma' X_{isj} + \epsilon_{isjz}}$$

We conduct three additional sets of analysis to understand the effects of the program on learning. First, we estimate whether the effects of the program significantly differed across tests at different points in time, with:

$$(4) \qquad \boldsymbol{Y_{isjzt} = \sum_{t=1}^{4} admin_t(\beta_{0t} + \beta_{1t} T_j) \quad + \gamma X_{isj} + \epsilon_{isjzt},}$$

Where the variable $admin_t$ is an indicator

for each of the four test administrations: 2012 pre- program test, 2012 post-program test, 2013 pre-program test, and 2013 post-program test. Note the addition of the *t* subscript in this specification, and that we estimate different constants and treatment effects for each of the four tests.

To test differences of effect across subject, we estimate the following specification:

$$(5)\ \boldsymbol{Y_{isjz} = \beta_{1Hindi}\, Hindi_z \times T_j + \beta_{1Math}\, Math_z \times T_j + \beta_{1English}\, English_z \times T_j + \gamma X_{isjz} + \epsilon_{isjz}}$$

where $Hindi_z$, $Math_z$, and $English_z$ are indicators for each subject. Again, γX_{isj} includes fixed effects for subject and cohorts, as well as controls for school size at baseline.

To test for differential effects of the program by pre-program ability, we estimate:

$$(6) \qquad Y_{isjz} = \sum_{v=1}^{5} 1\{Pretest_{isjz} = v\}(\beta_v + \delta_v T_j) + \gamma X_{ij} + \epsilon_{isjz}$$

where $Pretest_{ijs}$ indicates the pre-program test score for student i in village j on subject s. Differences in the coefficients δ_v identify different treatment effects by pre-program test score, while differences in the coefficients β_v indicate different means among students in control schools by pre-program score. We include cohort and subject fixed effects in all specifications as well as controls for school size at baseline.

We estimate Equations (4), (5), and (6), for all students pooled together as well as boys and girls separately. All specifications include three outcomes per student for each test administration, because students took tests in three subjects.

Effects on School Management Committee Outcomes

Lastly, since one aim of the program is to build capacity through School Management Committees (SMCs), we measure the program's effect on indicators of SMC activity. To do this, we compare the average number of SMC meetings, and number of improvement plans prepared and completed among treatment and control schools. In this manner, we test whether the program led to more SMC activity.

4. Effect of the Program on Enrollment, Learning and School Management

4.1 Enrollment

Table 2 shows the estimates of the program's impact on the number of enrolled students in 2012 and 2013. Columns 1-2 show results for all students, while Columns 3-6 disaggregate the results by gender. In 2012 – the first year of the program, the program increased the number of students per grade by 0.655 (Panel A, Column 2). There is no detectable effect on boys (Columns 3-4), while the effect on girls is large and statistically significant (Column 6). The program effect on girls amounts to an increase of 0.684 additional students, representing an increase of 8.1% from

the mean enrollment of girls at baseline. Further, we run a fully-interacted model analogous to Columns 4 and 6 to test whether the difference across gender is significant, which returns a p-value of 0.039 (not shown).

In 2013, the total effects on enrollment were even larger than in 2012 (Panel B). The total program effect is an increase by 1.3 students per grade, 0.4 additional boys and 0.9 girls (Columns 2, 4, 6). The difference in the effect of the program by gender is not statistically significant (p=0.215, not shown).

Table 2 -- Treatment Effects on Student Enrollment: School-Grade Level

Panel A: 2012						
Population Group	All Students		Boys		Girls	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.162	0.655	-0.043	-0.030	0.206	0.684**
	(1.302)	(0.539)	(0.716)	(0.303)	(0.737)	(0.336)
Controls	NO	YES	NO	YES	NO	YES
Observations	690	690	690	690	690	690
R-squared	0.000	0.694	0.000	0.653	0.000	0.633
Mean of dep var in control	15.643	15.643	7.188	7.188	8.455	8.455
Panel B: 2013						
Population Group	All Students		Boys		Girls	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.801	1.286*	0.352	0.399	0.450	0.887**
	(1.276)	(0.682)	(0.696)	(0.392)	(0.716)	(0.395)
Controls	NO	YES	NO	YES	NO	YES
Observations	690	690	690	690	690	690
R-squared	0.001	0.583	0.001	0.525	0.001	0.532
Mean of dep var in control	14.185	14.185	6.586	6.586	7.598	7.598

Notes: Robust standard errors in parentheses, clustered by village. *** p<0.01, ** p<0.05, * p<0.1. The sample includes students in grades 3, 4, and 5 in 230 schools. Observations are at the school-grade-level. Controls include 2011 (pre-program) enrollment by gender and cohort fixed effects where indicated, in Columns 2, 4, and 6.

4.2 Program Effects on the Composition of Test-Takers

Before turning to the results on learning, we first examine the potential for bias due to differential test-taking caused by the program. If more marginal students were more likely to take the learning assessments in treatment schools, our program effect estimates might be biased downward.

Appendix Table A1 presents the effect of the program on school-grade-level test-taking in 2012 and 2013. We present the effect of the program on taking each of the pre- and post-program tests, as well as for taking both tests. While noting the relatively large standard errors, we see no significant effects of the program on test-taking in 2012 (Panel A) – despite the effects on enrollment found in Panel A of Table 2. This result suggests that the program could enroll students, but not necessarily increase attendance or test-taking.

We do find significant effects on the number of students who take tests in 2013 (Panel B). In 2013, there were 0.6 more girls and 0.3 more boys in program schools who took the pre-program test (Columns 2 and 3). This is consistent with the increase in student enrollment in 2013 (Panel B, Table 2). Girls were also more likely to take the post-program tests (Columns 6 and 9).

Next, we examine what types of students are more likely to be test-takers across the treatment and control in Appendix Table A2. This analysis is restricted to our test-taker sample, defined as those who took both tests in 2012 (Columns 1-2) or 2013 (Columns 3-4). Since we have pre-program data for all these students, we can analyze whether students in treatment schools were weaker on average. We find no evidence of this. Test-takers in treatment schools were marginally more likely to be girls, but we see very little evidence that their test scores were lower. We interpret this as a lack of evidence that the program caused differential test-taking, suggesting limited scope for bias in our analysis of effects on learning.

4.3 Learning

Table 3 presents the effects of the program on learning in 2012 (Panel A) and 2013 (Panel B). We find no significant differences between treatment and control on the 2012 pre-program test – further indication of both the pre-program balance we saw in Table 1 and the lack of differential test-taking in Appendix Tables A1 and A2.

Table 3 -- Treatment Effects on Learning Outcomes

Panel A: 2012

Outcome:	Normalized Test Scores				Increased from pre- to post-program test*	
	Pre-program test		Post-program test			
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.007	0.009	0.323***	0.321***	0.227***	0.230***
	(0.090)	(0.096)	(0.094)	(0.097)	(0.059)	(0.061)
Girl		-0.117***		-0.084**		0.034**
		(0.035)		(0.035)		(0.014)
Treatment × Girl		-0.008		-0.012		-0.017
		(0.056)		(0.056)		(0.020)
Controls	NO	YES	NO	YES	NO	YES
Observations	17,460	17,460	17,460	17,460	16,347	16,347
R-squared	0.000	0.173	0.024	0.233	0.055	0.067
Mean of dep var in control	0.000	0.000	0.000	0.000	0.512	0.512
P-value: T + T × Girl = 0		0.991		0.002		0.000

Panel B: 2013

Outcome:	Normalized Test Scores				Increased from pre- to post-program test	
	Pre-program test		Post-program test			
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.021	0.028	0.156	0.153	0.077	0.078
	(0.091)	(0.097)	(0.106)	(0.110)	(0.051)	(0.051)
Girl		-0.033		-0.033		0.009
		(0.039)		(0.034)		(0.018)
Treatment × Girl		-0.044		-0.022		-0.005
		(0.059)		(0.052)		(0.028)
Controls	NO	YES	NO	YES	NO	YES
Observations	15,984	15,984	15,984	15,984	14,314	14,314
R-squared	0.000	0.204	0.006	0.242	0.006	0.031
Mean of dep var in control	0.000	0.000	0.000	0.000	0.536	0.536
P-value: T + T × Girl = 0		0.860		0.226		0.190

Notes: Robust standard errors in parentheses, clustered by village. *** p<0.01, ** p<0.05, * p<0.1. All analyses are restricted to test-takers, defined as students who were present for pre- and post-program tests in a given year (2012 or 2013). Observations are at the student-subject-level. Controls include 2011 (pre-program) enrollment by gender, cohort fixed effects, and subject (Math, English, and Hindi) fixed effects where indicated, in Columns 2, 4, and 6. Columns 5-6 restricted to students who scored less than the maximum possible score (A) on the pre-program test.

Columns 3-6 show that the program had a large and highly significant effect on test scores at the post-program test in 2012: those in treatment schools performed 0.32 standard deviation higher than those in control schools (Column 4) and students were 22.7 percentage points more likely to improve from the pre-program test (Column 6). The results for 2013 post-program tests, in Panel B, are qualitatively similar, albeit somewhat smaller. On average, students in treatment schools performed 0.15 standard deviation higher than those in the control (Column 4), and are 8 percentage points more likely to improve from the pre-program test (Column 6). There are no significant difference in the program's effects by gender in either 2012 or 2013 (Columns 4-6).[12]

Panel B, Columns 1 and 2 present results on the 2013 pre-program test, where we see no significant differences between treatment and control, suggesting that the learning gains from 2012 did not persist to the next academic school year. We return to this result in our discussion below.

Table 4 pools the analysis across years and all four test administrations—the pre- and post-program tests in each of 2012 and 2013. Tests of coefficient equality reveal that the treatment effect for the 2012 post-program test is significantly higher than either the preceding test (2012 pre-program test) or immediately subsequent test (2013 pre-program test). While the treatment effect on the 2012 post-program test is larger than the 2013 post-program test, the difference is not statistically significant for all students pooled (p-value=0.226, Column 1) or for boys (p-value=0.270, Column 2) or girls (p-value=0.207, Column 1) separately.[13]

[12] Results are qualitatively similar if we aggregate learning results by school-grade, as shown in Appendix Table A3, in which the dependent variable is the average within school-grade among the relevant population.

[13] We have also performed the analysis pooling across years. We find strongly positive estimated treatment effects on post-program tests for boys, girls, and both genders pooled.

Table 4 -- Treatment Effects on Normalized Test Scores across Test Administrations

	All Students	Boys	Girls
	(1)	(2)	(3)
Treatment × 2012 Pre-program test	-0.000	0.008	-0.008
	(0.088)	(0.094)	(0.090)
Treatment × 2012 Post-program test	0.288***	0.295***	0.281***
	(0.084)	(0.088)	(0.088)
Treatment × 2013 Pre-program test	0.006	0.029	-0.018
	(0.087)	(0.096)	(0.087)
Treatment × 2013 Post-program test	0.133	0.144	0.122
	(0.100)	(0.105)	(0.099)
Observations	66,888	33,404	33,484
R-squared	0.297	0.298	0.299
Mean of dep var in control	-0.001	0.028	-0.030
P-value: T × 2012 Pre-program test = T × 2012 Post-program test	0.000	0.000	0.000
P-value: T × 2012 Post-program test = T × 2013 Pre-program test	0.013	0.035	0.008
P-value: T × 2012 Post-program test = T × 2013 Post-program test	0.226	0.270	0.207
P-value: T × 2013 Pre-program test = T × 2013 Post-program test	0.073	0.102	0.075

Notes: Robust standard errors in parentheses, clustered by village. *** p<0.01, ** p<0.05, * p<0.1. All analyses are restricted to test-takers, defined as students who were present for pre- and post-program tests in the year the test was administered (2012 or 2013). Observations are at the student-subject-level. All specifications include controls for 2011 (pre-program) enrollment by gender, cohort fixed effects, subject (Math, English, and Hindi) fixed effects, and test administration (2012 pre-program test, 2012 post-program test, 2013 pre-program test, and 2013 post-program test) fixed effects.

One possible concern with the results in Table 4 is that the sample of test-takers in 2012 is different from the sample in 2013. For example, the results on the 2013 pre-program test may be driven by new students in 2013, such as third graders or newly-enrolled students, rather than students from 2012 who had been already exposed to the program. We run the analyses in Table 4 restricting the sample to students who were present for all four tests. We do this in two ways. First, in Appendix Table A4, we pool all four test administrations. Alternatively, in Appendix Table A5, we pool the latter three tests while controlling for normalized score on the 2012 pre-program test. While noting that we have less power to detect differences due to the restricted sample, results here are consistent with the full sample results shown in Table 4.

We next test for differences in treatment effect across test subjects. Table 5, Panel A shows little evidence that the effect varied by subject in 2012, while Panel B gives similar results for 2013. While scores in general are lower on English than Hindi and Math, we see large and significant treatment effects on all three subjects in 2012, with limited evidence of differences in treatment effect across tests. Similar results hold for 2013 but with smaller and insignificant

treatment effects for each subject.

Table 5 -- Treatment Effects on Normalized Test Scores by Subject

Panel A: 2012	All Students	Boys	Girls
	(1)	(2)	(3)
Treatment × Hindi	0.317***	0.323***	0.306***
	(0.087)	(0.092)	(0.092)
Treatment × English	0.256**	0.244*	0.261**
	(0.121)	(0.124)	(0.127)
Treatment × Math	0.369***	0.394***	0.340***
	(0.095)	(0.098)	(0.096)
Observations	17,460	8,509	8,951
R-squared	0.250	0.268	0.240
Mean of dep var in control	0.000	0.040	-0.039
P-value: T × Hindi = T × English	0.446	0.314	0.618
P-value: T × Hindi = T × Math	0.331	0.232	0.562
P-value: T × English = T × Math	0.169	0.092	0.351
Panel B: 2013	All Students	Boys	Girls
	(1)	(2)	(3)
Treatment × Hindi	0.127	0.139	0.119
	(0.099)	(0.106)	(0.096)
Treatment × English	0.159	0.172	0.151
	(0.125)	(0.127)	(0.129)
Treatment × Math	0.136	0.159	0.118
	(0.109)	(0.114)	(0.109)
Observations	15,984	8,193	7,791
R-squared	0.247	0.248	0.249
Mean of dep var in control	0.000	0.004	-0.004
P-value: T × Hindi = T × English	0.652	0.646	0.679
P-value: T × Hindi = T × Math	0.814	0.677	0.980
P-value: T × English = T × Math	0.746	0.844	0.676

Notes: Robust standard errors in parentheses, clustered by village. *** p<0.01, ** p<0.05, * p<0.1. All analyses are restricted to test-takers, defined as students who were present for pre- and post-program tests in the year the test was administered (2012 or 2013). Observations are at the student-subject-level. All specifications include controls for 2011 (pre-program) enrollment by gender, cohort fixed effects, and subject (Math, English, and Hindi) fixed effects.

Lastly, in Table 6, we disaggregate the effects of the program on learning by pre-program test scores.[14] We find that the positive effects on learning in 2012 are concentrated among students toward the top of the distribution, decreasing as one moves downward in score, with the lowest-

scoring students failing to experience the program's learning benefits (Panel A, Columns 1-3).[15] The results are somewhat different in 2013, where there are large increases in test scores for those scoring at the middle of the distribution on pre-program tests (Panel B).[16]

Table 6 -- Treatment Effects on Normalized Test Score: Heterogeneity by Pre-Program Test Score

Panel A: 2012	Normalized Score		
	All Students	Boys	Girls
	(1)	(2)	(3)
Treatment × Pre-program test D/E	0.242**	0.279***	0.205**
	(0.099)	(0.104)	(0.099)
Treatment × Pre-program test C	0.352***	0.334***	0.364***
	(0.071)	(0.075)	(0.072)
Treatment × Pre-program test B	0.466***	0.462***	0.466***
	(0.084)	(0.083)	(0.092)
Treatment × Pre-program test A	0.056	0.065	0.031
	(0.046)	(0.044)	(0.064)
Observations	17,460	8,509	8,951
R-squared	0.614	0.628	0.601
Mean of dep var in control	0.000	0.040	-0.039
P-value: Effect for D/E = Effect for C	0.154	0.534	0.043
P-value: Effect for C = Effect for B	0.082	0.067	0.153
P-value: Effect for B = Effect for A	0.000	0.000	0.000
Panel B: 2013	Normalized Score		
	All Students	Boys	Girls
	(1)	(2)	(3)
Treatment × Pre-program test D/E	0.152	0.146	0.164
	(0.133)	(0.136)	(0.139)
Treatment × Pre-program test C	0.205**	0.225***	0.189**
	(0.085)	(0.085)	(0.093)
Treatment × Pre-program test B	0.081	0.071	0.091
	(0.060)	(0.065)	(0.062)
Treatment × Pre-program test A	0.004	0.009	-0.005
	(0.022)	(0.023)	(0.033)
Observations	15,984	8,193	7,791
R-squared	0.549	0.551	0.548
Mean of dep var in control	0.000	0.004	-0.004
P-value: Effect for D/E = Effect for C	0.571	0.429	0.808
P-value: Effect for C = Effect for B	0.035	0.011	0.159
P-value: Effect for B = Effect for A	0.199	0.342	0.128

Notes: Robust standard errors in parentheses, clustered by village. *** p<0.01, ** p<0.05, * p<0.1. All analyses are restricted to test-takers, defined as students who were present for pre- and post-program tests in a given year (2012 or 2013). Observations are at the student-subject-level. All specifications include controls for 2011 (pre-program) enrollment by gender, indicators of baseline pre-program test score, cohort fixed effects, and subject (Hindi, English, Math) fixed effects.

There is no evidence that the treatment effect on test scores, disaggregated by pre-program test scores, is different for boys and girls in any specification, which is consistent with the fact that there was no gender-specific aspect to the learning component of the program.

4.4 School Management

Our last set of results presents the effects of the program on the third prong of the NGO program—school-level SMC outcomes (Table 7). Effects each month are presented in Appendix Table A4. Over the course of the seven months in which data were collected, treatment schools held an additional 0.66 committee meetings on average, an increase of 15.6 percent (p-value=0.019). In addition to holding meetings, school committees produced significantly more output, as measured by the number of improvement plans prepared and completed. While the figures are only marginally significant, committees in treatment schools prepared 22.4 percent (p-value=0.086) more improvement plans and completed 24.7 percent (p-value=0.100) more such plans. Appendix Table A4 disaggregates results by month.

Table 7 -- Treatment Effects on School Management Committee Outcomes

	Treatment (1)	Control (2)	Difference (Treatment - Control) (3)	P-value of test of Equality (4)
Number of Meetings	4.898	4.239	0.659	0.019
	(0.192)	(0.202)	(0.278)	
Number of Improvement Plans Prepared	8.263	6.752	1.510	0.086
	(0.591)	(0.647)	(0.875)	
Number of Improvement Plans Completed	5.924	4.752	1.172	0.100
	(0.476)	(0.527)	(0.708)	
Observations	118	112		

Notes: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Observations are schools.

[14] Due to a small number of students scoring E, we pool D and E in running this analysis.
[15] Learning impacts for best-performing students (those scoring an A on pre-program tests) are not significant. A significant effect here would be implausible, as it would require students in control schools to drop from the highest score at a higher rate than students in treatment schools. We do find that those who receive A on the pre-program tests are 3.3 percentage points more likely to retain that score in 2012, but this result is not significant (p=0.218).
[16] Those who received an A on the 2012 pre-program test scores are just 1.1 percentage point more likely to score an A on the post-program test (p=0.356).

5. Discussion and Further Results

5.1 Enrollment vs. Retention

We found relatively large effects of the program on enrollment – especially for girls – in 2012 and 2013. How much of the program effect on total enrollment was driven by new students enrolling versus students being retained?

To answer this question, we match student enrollment records in 2011, to 2012 and 2013 and identify three samples of students: *Baseline*, *Newly Enrolled* (students not enrolled in grades three or four in 2011 or 2012, but enrolled in grades four or five in 2012 or 2013, respectively), and *New Cohort* (students enrolled in grade three in either 2012 or 2013). To match students across 2011, 2012 and 2013 enrollment records, we use information regarding each student's school, village, gender, age, last and first name, and father's name. In some cases – for example, when a student name is common or when we are missing demographic information – there are multiple possible matches. Any student who matches to subsequent enrollment records (uniquely or otherwise), is coded as retained to the relevant year.[17]

Appendix Table A7 estimates Equation (1) on enrollment numbers, disaggregating total enrollment by sample of student. In 2012, we find that a portion of the treatment effects on enrollment are driven by baseline sample girls – a coefficient of 0.28 (Panel A, Column 3), but the majority of the effects are coming from newly enrolled fourth and fifth grade girls – a coefficient of 1.1 (Panel A, Column 6). We detect no enrollment effects on new third graders (Panel A, Columns 7-9). In 2013, there are positive coefficients on the treatment effect on enrollment for

[17] Matching baseline students to 2012 yields 97.1 percent unique matches (99.1 percent in the treatment and 95.1 in the control). Similarly, there are 97.6 percent unique matches from the baseline to 2013 (99.3% in the treatment and 95.7% in the control). The difference in the match rate is statistically significant in both years, p=0.001 in 2012, p=0.002 in 2013. The difference in the match rates across treatment and control schools is due entirely to enrollment rosters missing students' father's name: 40 schools have 100 percent missing father's names, of these 37 are control schools.

baseline girls (0.194) and for newly enrolled fourth and fifth grade girls (0.626), with the largest effects on new third grade girls (1.036).

As further evidence of retention, we use individual-level data and restrict the sample to baseline students who were enrolled in grades three and four in 2011. We estimate:

$$(7) \qquad Enrolled_{isj} = \beta_0 + \beta_1 T_j + \beta_2 Girl_{isj} + \beta_3 T_j \times Girl_{isj} + \gamma' X_{isj} + \epsilon_{isj}$$

for student i in school s in village j. We include cohort fixed effects and controls for number of boys and girls enrolled in each school prior to program implementation. Appendix Table A6 presents these results.

Within relevant cohorts, among students in the Baseline Sample in control schools, 77.0 percent were enrolled in 2012, and 58.7 percent were enrolled in 2013. Baseline Sample students in treatment schools were 4.6 percentage points more likely to be enrolled in 2012, a 6 percent increase in retention (Column 1). This effect is statistically significant and concentrated among girls in treatment schools, who were 6.2 percentage points more likely to still be enrolled in school in 2012 (p-value = 0.010) than girls in control schools (Column 2). The effects of retention are similarly high in 2013. The main treatment effect is 6.7 percentage points, or an 11.4 percent increase in the likelihood of being retained over those in the control (Column 3). Again, this effect is higher for girls in treatment schools, who were 8.4 percentage points more likely to re-enroll in school than girls in control schools (p-value = 0.044).

Taken together, these results indicate that the NGO program resulted in higher rates of both retention and new enrollment over the first two years of implementation for girls. Note however, the differences in retention effects are not statistically different across genders.

5.2 Potential Threat to Validity – Differential Enrollment and Test-Taking

We found large increases in enrollment and retention among girls due to the program (Table 2, Appendix Tables A7 and A8), and found some positive – and in 2013, statistically significant – program effects on test-taking (Appendix Table A1). While we find no evidence of differences in test-taker characteristics by treatment group in either 2012 or 2013 (Appendix Table 2), we may still worry about differences in unobservables that might bias our learning results.

Appendix Table A9 presents the treatment effects on test-taking (defined as taking both pre- and post-program tests in a given year) by sample type: Baseline, Newly Enrolled, or New Cohort. We find that the main increases in test-taking due to the program are among Newly Enrolled girls in 2012 (Column 6), and Newly Enrolled and New Cohort girls in 2013.

Examining the difference in pre-program learning outcomes between the treatment and control provides some additional insight in terms of selection. Appendix Table A10 presents the estimates from equation (3) on pre-program normalized test score, conditional on taking both pre- and post-program test scores and disaggregated by sample. We find no evidence that students were negatively (or positively) selected in 2012 (Panel A). There is some suggestive evidence, however, of negative selection among Newly Enrolled girls in treatment schools, who scored somewhat worse on pre-program tests (-0.256 standard deviations lower, p=0.136, Panel B, Column 2). This might suggest that our estimates of the treatment effects on learning would be a lower bound estimate for the true impacts, had the program not had a differential effect on enrollment and test-taking. For completeness, we present learning outcomes in 2012 and 2013 separately by sample in Appendix Table A11.

5.3 Fade-Out of Program Effects

The effect of the program immediately after the first year of implementation did not fully persist into the second year of implementation. Among students who were enrolled and test-takers in both 2012 and 2013, the treatment effect on post-program tests in 2012 was 0.278, while the treatment effect on pre-program tests in 2013 was 0.084 (Appendix Table A4, Column 2). The difference between these effects is statistically significant (p=0.090).

Studies on long-run effects have shown that the positive effect on test scores of even successful interventions often fades over time (Banerjee et al. 2007; Andrabi, Das, and Khwaja 2008; Buhl-Wiggers et al. 2019). This may be due to tendencies for programs to "teach to the test", reinforcing the specific skills that are required for successful test-taking; in this case, education interventions will be more likely to have long-lasting effects when they target core skills (Duflo, Dupas, and Kremer 2011). On the other hand, there may be less enthusiasm for the program over time, or newly learned skills – either by teachers or students – may naturally revert back to pre-program performance without refresher courses or retraining. Longer term gains from an intervention may also re-appear later in life (Chetty et al. 2014).

In our program, Baseline Sample students were exposed to two consecutive years of the program. Our findings—that the program's positive impacts are smaller and not statistically significant during the second year, despite the continuation of the intervention—contrasts with the other studies that find that learning gains can be cumulative with multiple years of treatment (Buhl-Wiggers et al., 2019).

Field observations from the research staff's random visits to treatment schools suggest that two main factors were at play in in the second program year. First, political economy factors changed in the second year, altering volunteers' incentives in the classroom and reducing their overall

engagement. Volunteers were not paid but offered the possibility of future employment as NGO staff. Thus, they may have put in more effort during the first year to keep their options open, until a decision was made regarding whether or not they would receive a position with the NGO. In the second year, after the staffing decision was made, volunteers' motivation may have dropped both among returning volunteers not granted access to a staff position, as well as among volunteers newly appointed as staff, once their contract was secured. This explanation is in line with the findings of Bold et al. (2016), who show strong positive impacts of a contract teacher intervention in Kenya when implemented by an NGO (0.18 standard deviations), contrasted with null impact when contract teachers were hired by the Ministry of Education.

Second, the somewhat larger classes – as a result of the program's enrollment drive – may have reduced the effectiveness of the program in the second year (Krueger and Whitmore, 2001). Third, we find significantly lower program gains in the pre-program tests of the second year – at the start of the academic year. There is an established literature in the United States of the decline in academic performance over summer vacation – especially among lower income students (Atteberry and McEachin, 2006; Cooper et al., 1996), yet there is less known in developing countries. Our results suggest that the learning losses over school vacation may be an issue in the context of rural India as well.

6. Conclusion

This paper presents the results of a multi-faceted education intervention conducted in rural Rajasthan, India. The program had three primary aims: increasing participation, learning outcomes, and gender equity at school among a particularly vulnerable population.

Many randomized experiments have looked into the effectiveness of programs aimed at either increasing school participation (Bobonis, Miguel, and Puri-Sharma 2006; Duflo, Hanna, and Ryan 2012;) or improving teaching quality (Banerjee et al. 2007; Borkum, He, and Linden 2012; Das et al. 2013; Muralidharan and Sundararaman 2010). Though most learning-targeted interventions, based on an improvement of teaching methods rather than inputs, have shown significant impacts on learning (Banerjee et al. 2007; He, Linden, and MacLeod 2008; Linden 2008), enrollment-targeted interventions have shown zero or small learning effects (Miguel and Kremer 2004; Petrosino et al. 2012). For instance, conditional cash transfers have been shown to improve school enrollment and educational attainment (Galiani and McEwan 2013) but to have limited effects on learning (Behrman, Parker, and Todd 2009).

Our findings – of increased enrollment and learning – suggest that multi-faceted interventions may be able to overcome the possible trade-offs between these two objectives. While a quantity-quality trade-off for schooling would arise if the increase in pupil-teacher ratios hampered a teacher's ability to improve the learning of all students (Duraisamy et al. 1998), there is a surprising scarcity of rigorous empirical evidence of these types of trade-offs or complementarities in education. Challenging the trade-off assumption, a small number of studies, including Banerjee et al. (2007), have shown an absence of correlation between class size and test scores. Our findings concur with the absence of such trade-off between access and learning. One plausible explanation is that the innovative curriculum, which lies at the core of the program we evaluate here, may be particularly effective at targeting the pedagogical needs of students and counterbalancing the possible harmful effects of enrollment on class size, in line with Banerjee et al. (2007) and Muralidharan, Singh, and Ganimian (2017). The effects of the program on school

management outcomes may also be important for its success – evidence has shown that successful school-based management improves both participation and learning (JPAL Policy Bulletin 2017). Although our study is one of the first to evaluate interventions combining enrollment and learning targets at the same time, it is important to emphasize that we do not compare an intervention that is multi-faceted with one that is not. However, this paper suggests that the dual objective of improving both access and learning in primary education may well be achieved through educational policies combining enrollment-targeted interventions with interventions aimed at tailoring curricula to the individual needs of students by teaching at the right level. Rigorous evidence on this statement could emanate from further research investigating the relative effectiveness of interventions targeting (i) enrollment only, (ii) enrollment and learning through additional inputs, and (iii) enrollment and learning through need-tailored curricula.

References

Andrabi, T., J. Das, and A. I. Khwaja. 2008. "A Dime a Day: The Possibilities and Limits of Private Schooling in Pakistan." *Comparative Education Review* 52 (3): 329–355.

Angrist, J. D., and V. Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement." *Quarterly Journal of Economics* 114 (2): 533–575.

Atteberry, A., & McEachin, A. (2016). School's out: Summer learning loss across grade levels and school contexts in the United States today. In Alexander, K., Pitcock, S., & Boulay, M. (Eds). Summer learning and summer learning loss, pp35-54. New York: Teachers College Press.

Baird, S., J. H. Hicks, M. Kremer, and E. Miguel. 2016. "Worms at Work: Long-Run Impact of Child Health Gains." *Quarterly Journal of Economics*. 131 (4): 1637–1680.

Banerjee, A. V., S. Cole, E. Duflo, and L. Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122 (3): 1235–1264.

Banerjee, Abhijit, Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani. 2010. "Pitfalls of participatory programs: evidence from a randomized evaluation in education in India." *American Economic Journal: Economic Policy*. 2 (1), 1–30.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton. "Mainstreaming an effective intervention: Evidence from randomized evaluations of "Teaching at the Right Level" in India." No. w22746. National Bureau of Economic Research, 2016.

Banerji, Rukmini, James Berry, and Marc Shotland. 2017. "The Impact of Maternal Literacy and Participation Programs: Evidence from a Randomized Evaluation in India." American Economic Journal: Applied Economics, 9 (4): 303-37. DOI: 10.1257/app.20150390

Barr, Abigail, Frederick Mugisha, Pieter Serneels, and Andrew Zeitlin. 2012. "Information and collective action in the community monitoring of schools: Field and lab experimental evidence from Uganda." Unpublished manuscript.

Beasley Elizabeth, and Elise Huillery, 2017. "Willing but Unable? Short-term Experimental Evidence on Parent Empowerment and School Quality," *World Bank Economic Review*, World Bank Group, vol. 31(2), pages 531-552.

Behrman, J. R., S. W. Parker, and P. E. Todd. 2009. “Schooling Impacts of Conditional Cash Transfers on Young Children: Evidence from Mexico.” *Economic Development and Cultural Change* 57 (3): 439–477.

Benveniste, L. A., and P. J. McEwan. 2000. “Constraints to Implementing Educational Innovations: The Case of Multigrade Schools.” *International Review of Education* 46 (1–2): 31–48.

Blimpo, Moussa P., David K. Evans, and Nathalie Lahire. “Parental Human Capital and Effective School Management: Evidence from The Gambia.” World Bank Policy Research Working Paper 7238, April 2015

Bobonis, G. J., E. Miguel, and C. Puri-Sharma. 2006. “Anemia and School Participation.” *Journal of Human Resources* 41 (4): 692–721.

Bold, T., M. Kimenyi, G. Mwabu, A. Ng’ang’a, and J. Sandefur. 2016. *Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education*. Working Paper WPS/2013-04. Oxford, UK: Centre for the Study of African Economies.

Borkum, E., F. He, and L. L. Linden. 2012. *The Effects of School Libraries on Language Skills: Evidence from a Randomized Controlled Trial in India*. Working Paper 18183. Cambridge, MA, US: National Bureau of Economic Research.

Bruns, B., A. Mingat, and R. Rakatomalala. 2003. *Achieving Universal Primary Education by 2015: A Chance for Every Child*. Washington, DC: World Bank.

Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2014. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." American Economic Review, 104 (9): 2633-79.

Cooper H., Nye B., Charlton K., Lindsay J., Greathouse S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. Review of Educational Research, 66(3), 227–268.

Damon, Amy, P. Glewwe, S. Wisniewski, and B. Sun. 2016. “Education in Developing Countries. What Policies and Programmes affect Learning and Time in School?” Report for the Expert Group for Aid Studies.

Das, J., S. Dercon, J. Habyarimana, P. Krishnan, K. Muralidharan, and V. Sundararaman. 2013. “School Inputs, Household Substitution, and Test Scores.” *American Economic Journal: Applied Economics* 5 (2): 29–57.

Deininger, K. 2003. “Does Cost of Schooling Affect Enrollment by the Poor? Universal Primary Education in Uganda.” *Economics of Education Review* 22 (3): 291–305.

Delavallade, C., A. Griffith, and R. Thornton. 2016. “Network Partitioning and Social Exclusion under Different Selection Regimes.” Unpublished, International Food Policy Research Institute, Washington, DC.

Duflo, E., P. Dupas, and M. Kremer. 2011. “Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya.” *American Economic Review* 101:1739–1774.

———. 2015. “School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools.” *Journal of Public Economics* 123:92–110.

Duflo, E., R. Hanna, and S. P. Ryan. 2012. “Incentives Work: Getting Teachers to Come to School.” *American Economic Review* 102 (4): 1241–1278.

Duraisamy, P, E James, J Lane, an JP Tan. 1998. “Is there a quantity–quality trade-off as pupil–teacher ratios increase? Evidence from Tamil Nadu, India” *International Journal of Educational Development* 18 (5), 367-383

Epple, D., E. Newlon, and R. Romano. 2002. “Ability Tracking, School Competition, and the Distribution of Educational Benefits.” *Journal of Public Economics* 83 (1): 1–48.

Evans, David and Fei Yuan. 2018. “What We Learn about Girls’ Education from Interventions that Don’t Focus on Girls. Working Paper.

Exeter, D., S. Ameratunga, M. Ratima, S. Morton, M. Dickson, D. Hsu and R. Jackson. 2010. “Student engagement in very large classes: the teachers’ perspective”, *Studies in Higher Education*, 35 (7): 761-775.

Filmer, D., and L. Pritchett. 1999. “The Effect of Household Wealth on Educational Attainment: Evidence from 35 Countries.” *Population and Development Review* 25 (1): 85–120.

Galiani, S., and P. J. McEwan. 2013. “The Heterogeneous Impact of Conditional Cash Transfers.” *Journal of Public Economics* 103:85–96.

Gertler, Paul, Harry Patrinos, and Eduardo Rodríguez-Oreggia. “Parental Empowerment in Mexico: Randomized Experiment of the "Apoyos a la Gestion Escolar (AGE)" in Rural Primary Schools in Mexico.” Washington, DC: The World Bank. Preliminary draft, August 2012.

Gibbs, G. and A. Jenkins (eds). 1992. *Teaching Large Classes in Higher Education*. London: Kogan Page.

Glewwe, P., N. Ilias, and M. Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2(3): 205-227.

Glewwe, P., M. Kremer, and S. Moulin. 2009. "Many Children Left Behind? Textbooks and Test Scores in Kenya." *American Economic Journal: Applied Economics* 1 (1): 112–135.

Glewwe, P., M. Kremer, S. Moulin, and E. Zitzewitz. 2004. "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya." *Journal of Development Economics* 74 (1): 251–268.

Glewwe, P., and E.W. Maïga. 2011. "The impacts of school management reforms in Madagascar: do the impacts vary by teacher type?" *Journal of Development Effectiveness*, *3*(4), pp.435-469.

He, F., L. L. Linden, and M. MacLeod. 2008. "How to Teach English in India: Testing the Relative Productivity of Instruction Methods with Pratham English Language Education Program." Unpublished, Columbia University, New York.

Herz, B. 2002. *Universal Basic Education: What Works*. Prepared for the Coalition for Basic Education. Washington, DC: Academy for Educational Development.

J-PAL Policy Bulletin. 2017. "Roll Call: Getting Children into School.

Kerwin, Jason and Thornton, Rebecca L., Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures (January 30, 2018). Available at SSRN: https://ssrn.com/abstract=3002723

Khandker, S., M. Pitt, and N. Fuwa. 2003. "Subsidy to Promote Girls' Secondary Education: The Female Stipend Program in Bangladesh." Unpublished, Munich Personal RePEc Archive.

Khandker, S. R. 1996. *Education Achievements and School Efficiency in Rural Bangladesh*. Discussion Paper 319. Washington, DC: World Bank.

Kiessel, J. and A. Duflo. 2014. "Cost Effectiveness Report: Teacher Community Assistant Initiative (TCAI)." IPA Brief, Innovations for Poverty Action, New Haven, CT.

Kremer, M., C. Brannen, and R. Glennerster. 2013. "The Challenge of Education and Learning in the Developing World." *Science* 340 (6130): 297–300.

Kremer, M., and A. Holla. 2009. "Improving Education in the Developing World: What Have We Learned from Randomized Evaluations?" *Annual Review of Economics* 1:513–545.

Kremer, M., and E. Miguel. 2007. “The Illusion of Sustainability.” *Quarterly Journal of Economics* 112:1007–1065.

Kremer, M., E. Miguel, and R. Thornton. 2009. “Incentives to Learn.” *Review of Economics and Statistics* 91 (3): 437–456.

Krueger, Alan, and Diane M. Whitmore. 2001. “The Effect of Attending Small Class in Early Grades on College Test-Taking and Middle School Test Results: Evidence from Project STAR,” *The Economic Journal*, CXI, 1–28.

Lakshminarayana, R., Eble, A., Bhakta, P., Frost, C., Boone, P., Elbourne, D., & Mann, V. 2013. The Support to Rural India's Public Education System (STRIPES) Trial: A Cluster Randomised Controlled Trial of Supplementary Teaching, Learning Material and Material Support. *PloS one*, 8(7), e65775.

Lassibille, Gérard, Jee-Pang Tan, Cornelia Jesse, and Trang Van Nguyen. 2010. “Managing for Results in Primary Education in Madagascar: Evaluating the Impact of Selected Workflow Interventions.” *The World Bank Economic Review*. 24 (2): 303-329.

Lavinas, L. 2001. *The Appeal of Minimum Income Programmes in Latin America*. Geneva: International Labour Organization.

Linden, L. L. 2008. *Complement or Substitute? The Effect of Technology on Student Achievement in India*. InfoDev Working Paper 17. Washington, DC: World Bank.

Lloyd, C. B., B. Mensch, and W. Clark. 1998. *The Effects of Primary School Quality on the Educational Participation and Attainment of Kenyan Girls and Boys*. Working Paper 116. New York: Population Council.

Lucas, A. M., P. J. McEwan, M. Ngware, and M. Oketch. 2014. “Improving Early-Grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda.” *Journal of Policy Analysis and Management* 33:950–976.

Mbiti, Isaac, M. Romero, Y. Schipper, C. Manda, and Rakesh Rajani. 2018. “Inputs, Incentives, and Complementarities in Education: Experimental Evidence from Tanzania” NBER Working Paper 24876.

McEwan, P. 2015. “Improving Learning in Primary Schools of Developing Countries: A Meta-analysis of Randomized Experiments.” *Review of Educational Research* 85 (3): 1–42.

Miguel, E., and M. Kremer. 2004. “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities.” *Econometrica* 72 (1): 159–217.

Muralidharan, Karthik, and Nishith Prakash. 2017. "Cycling to School: Increasing Secondary School Enrollment for Girls in India." *American Economic Journal: Applied Economics*, 9 (3): 321-50.

Muralidharan, K., A. Singh, and A. Ganimian. 2018. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." *American Economic Review*. Forthcoming.

Muralidharan, K., and V. Sundararaman. 2010. "The Impact of Diagnostic Feedback to Teachers on Student Learning: Experimental Evidence from India." *Economic Journal, Royal Economic Society* 120 (546): F187-F203.

Nguyen, T. 2008. Information, role models, and perceived returns to education: Experimental evidence from Madagascar. Unpublished manuscript, MIT, Cambridge, MA.

Osendarp, S. J. M., Baghurst, K. I., Bryan, J., Calvaresi, E., Hughes, D., Hussaini, M., . . . Wilson, C. 2007. Effect of a 12-mo micronutrient intervention on learning and memory in well-nourished and marginally nourished school-aged children: 2 parallel, randomized, placebo-controlled studies in Australia and Indonesia. American Journal of Clinical Nutrition, 86, 1082–1093.

Petrosino, A., C. Morgan, T. Fronius, E. Tanner-Smith, and R. Boruch. 2012. "Interventions in Developing Nations for Improving Primary and Secondary School Enrollment of Children: A Systematic Review." *Campbell Systematic Reviews* 8:19.

Pradhan, Menno, Daniel Suryadarma, Amanda Beatty, Maisy Wong, Arya Gaduh, Armida Alisjahbana, and Rima Prama Artha. 2014. "Improving Educational Quality through Enhancing Community Participation: Results from a Randomized Field Experiment in Indonesia." *American Economic Journal: Applied Economics.* 6 (2): 105-126.

Pratham Organization. 2011. *Annual Status of Education Report.* Mumbai: Pratham Resource Center.

———. 2012. *Annual Status of Education Report.* Mumbai: Pratham Resource Center.

Rosenthal R (1979). "File drawer problem and tolerance for null results". Psychol Bull. 86 (3): 638–41. doi:10.1037/0033-2909.86.3.638.

Torgerson, Carole J., Sarah E. King & Amanda J. Sowden (2002) Do Volunteers in Schools Help Children Learn to Read? A Systematic Review of Randomised Controlled Trials, Educational Studies, 28:4, 433-444, DOI: 10.1080/0305569022000042435

UNICEF. 2012. *Child Marriage in India: An Analysis of Available Data.* New Delhi.

United Nations. 2015. *Transforming Our World: The 2030 Agenda for Sustainable Development*. New York.

World Bank. *World Development Report 2012: Gender Equality and Development*. Washington, DC.

———. 2018. World Development Report 2018: Learning to Realize Education's Promise. Washington, DC: World Bank. doi:10.1596/978-1-4648-1096-1. License: Creative Commons Attribution CC BY 3.0 IGO

———. 2001. *Pioneering Support for Girls' Secondary Education: The Bangladesh Female Secondary School Assistance Project*. Washington, DC.

Appendix Table A1 -- Treatment Effects on Test-Taking: School-Grade Level

Panel A: 2012

Test	Pre-program test			Post-program test			Pre- and post-program tests		
	All Students	Boys	Girls	All Students	Boys	Girls	All Students	Boys	Girls
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	0.025	-0.253	0.279	-0.205	-0.379	0.174	-0.218	-0.324	0.106
	(0.526)	(0.269)	(0.328)	(0.515)	(0.278)	(0.314)	(0.457)	(0.247)	(0.276)
Observations	690	690	690	690	690	690	690	690	690
R-squared	0.551	0.569	0.442	0.568	0.565	0.472	0.492	0.497	0.391
Mean of dep var in control	10.821	5.238	5.583	10.869	5.292	5.577	8.655	4.274	4.381

Panel B: 2013

Test	Pre-program test			Post-program test			Pre- and post-program tests		
	All Students	Boys	Girls	All Students	Boys	Girls	All Students	Boys	Girls
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	0.903*	0.300	0.603*	0.487	0.031	0.456	0.525	0.078	0.447*
	(0.539)	(0.306)	(0.318)	(0.538)	(0.316)	(0.298)	(0.466)	(0.277)	(0.256)
Observations	690	690	690	690	690	690	690	690	690
R-squared	0.529	0.507	0.434	0.468	0.454	0.369	0.437	0.444	0.323
Mean of dep var in control	10.315	5.060	5.256	9.259	4.679	4.580	7.557	3.926	3.631

Notes: Robust standard errors in parentheses, clustered by village. *** p<0.01, ** p<0.05, * p<0.1. The sample includes grades 3, 4, and 5 in 230 schools. Observations are at the school-grade-level. All specifications include controls for 2011 (pre-program) enrollment by gender and cohort fixed effects.

Appendix Table A2 -- Pre-program Scores of Test-Takers

	2012 Test-Takers (N=17,460)		2013 Test-Takers (N=15,984)	
	Control	Difference	Control	Difference
	(1)	(2)	(3)	(4)
Girl	0.506	0.012	0.481	0.013
	(0.016)	(0.021)	(0.016)	(0.023)
Normalized Pre-Test Score	0.000	0.007	0.000	0.021
	(0.068)	(0.090)	(0.063)	(0.091)
Pre-Test E	0.084	0.003	0.042	0.001
	(0.015)	(0.022)	(0.011)	(0.014)
Pre-Test D	0.350	-0.010	0.258	-0.004
	(0.029)	(0.033)	(0.020)	(0.030)
Pre-Test C	0.307	0.004	0.342	-0.010
	(0.012)	(0.019)	(0.015)	(0.022)
Pre-Test B	0.196	0.002	0.256	0.008
	(0.021)	(0.026)	(0.017)	(0.025)
Pre-Test A	0.063	0.001	0.102	0.005
	(0.009)	(0.013)	(0.012)	(0.018)

Notes: Robust standard errors in parentheses, clustered by village. *** p<0.01, ** p<0.05, * p<0.1. Test-takers are students who took pre- and post-program exams in a given year.

Appendix Table A3 -- Treatment Effects on Learning Outcomes: School-Grade Level

Panel A: 2012

Population Group	All Students			Boys			Girls		
Outcome	Pre-test	Post-test	Increased	Pre-test	Post-test	Increased	Pre-test	Post-test	Increased
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	-0.047	0.287***	0.218***	-0.003	0.316***	0.211***	-0.040	0.287***	0.211***
	(0.091)	(0.090)	(0.054)	(0.096)	(0.091)	(0.057)	(0.093)	(0.095)	(0.053)
Observations	681	681	681	594	594	594	621	621	621
R-squared	0.148	0.226	0.153	0.122	0.212	0.119	0.126	0.188	0.146
Mean of dep var in control	0.068	0.085	0.488	0.052	0.068	0.472	0.067	0.081	0.504

Panel B: 2013

Population Group	All Students			Boys			Girls		
Outcome	Pre-test	Post-test	Increased	Pre-test	Post-test	Increased	Pre-test	Post-test	Increased
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	-0.052	0.156*	0.102**	-0.010	0.191*	0.109***	-0.021	0.164*	0.083*
	(0.080)	(0.090)	(0.042)	(0.089)	(0.098)	(0.041)	(0.079)	(0.091)	(0.044)
Observations	663	663	663	585	585	585	606	606	606
R-squared	0.214	0.193	0.046	0.198	0.157	0.059	0.164	0.176	0.026
Mean of dep var in control	0.048	0.001	0.462	0.042	-0.006	0.454	0.008	-0.025	0.481

Notes: Robust standard errors in parentheses, clustered by village. *** p<0.01, ** p<0.05, * p<0.1. All analyses are restricted to test-takers, defined as students who were present for pre- and post-program tests in a given year (2012 or 2013). Observations are at the student-subject-level. All analyses include controls for 2011 (pre-program) enrollment by gender and cohort fixed effects. Columns 3, 6, and 9 restricted to students who scored less than the maximum possible score (A) on the pre-program test.

Appendix Table A4 -- Treatment Effects on Normalized Test Scores Across Test Administrations: Students who Took All Tests

	All Students	Boys	Girls
	(1)	(2)	(3)
Treatment × 2012 Pre-program test	-0.051	-0.057	-0.044
	(0.099)	(0.104)	(0.107)
Treatment × 2012 Post-program test	0.278***	0.291***	0.265***
	(0.094)	(0.098)	(0.101)
Treatment × 2013 Pre-program test	0.084	0.119	0.047
	(0.098)	(0.106)	(0.097)
Treatment × 2013 Post-program test	0.148	0.166	0.128
	(0.100)	(0.104)	(0.104)
Observations	27,828	14,340	13,488
R-squared	0.319	0.323	0.322
Mean of dep var in control	0.001	0.037	-0.038
P-value: T × 2012 Pre-program test = T × 2012 Post-program test	0.000	0.000	0.000
P-value: T × 2012 Post-program test = T × 2013 Pre-program test	0.090	0.171	0.061
P-value: T × 2012 Post-program test = T × 2013 Post-program test	0.296	0.349	0.286
P-value: T × 2013 Pre-program test = T × 2013 Post-program test	0.354	0.533	0.252

Notes: Robust standard errors in parentheses, clustered by village. *** p<0.01, ** p<0.05, * p<0.1. All analyses are restricted to students who were present for all four test administrations. Observations are at the student-subject-level. All specifications include 2011 (pre-program) enrollment by gender, cohort fixed effects, subject (Math, English, and Hindi) fixed effects, and test administration (2012 pre-program test, 2012 post-program test, 2013 pre-program test, and 2013 post-program test) fixed effects.

Appendix Table A5 -- Treatment Effects on Normalized Test Scores Across Test Administrations: Students who Took All Tests, Conditional on 2012 Pre-test score

	All Students		Boys		Girls	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment × 2012 Post-program test	0.288***	0.308***	0.303***	0.326***	0.274**	0.291***
	(0.097)	(0.071)	(0.101)	(0.073)	(0.105)	(0.078)
Treatment × 2013 Pre-program test	0.086	0.106	0.125	0.148	0.048	0.065
	(0.102)	(0.097)	(0.110)	(0.105)	(0.101)	(0.098)
Treatment × 2013 Post-program test	0.153	0.173	0.173	0.196*	0.133	0.150
	(0.103)	(0.109)	(0.107)	(0.113)	(0.107)	(0.111)
2012 Pre-program test score		0.384***		0.381***		0.385***
		(0.023)		(0.025)		(0.025)
Observations	20,871	20,871	10,755	10,755	10,116	10,116
R-squared	0.273	0.396	0.281	0.403	0.270	0.390
Mean of dep var in control	0.003	0.003	0.031	0.031	-0.027	-0.027
P-value: T × 2012 Post-program test = T × 2013 Pre-program test	0.090	0.090	0.171	0.171	0.061	0.061
P-value: T × 2012 Post-program test = T × 2013 Post-program test	0.296	0.296	0.349	0.349	0.286	0.286
P-value: T × 2013 Pre-program test = T × 2013 Post-program test	0.354	0.354	0.533	0.533	0.252	0.252

Notes: Robust standard errors in parentheses, clustered by village. *** p<0.01, ** p<0.05, * p<0.1. All analyses are restricted to students who were present for all four test administrations. Observations are at the student-subject-level. All specifications include 2011 (pre-program) enrollment by gender, cohort fixed effects, subject (Math, English, and Hindi) fixed effects, and test administration (2012 pre-program test, 2012 post-program test, 2013 pre-program test, and 2013 post-program test) fixed effects.