Scaling up remedial education in India: Evidence from two RCTs of the same program at different scales

Lucas Kitzmüller [a], Jeffery McManus [a,*], Neil Buddy Shah [a], and Kate Sturla [a]
[a] *IDinsight, P.O. Box 689, San Francisco, CA 94104, USA*

April 20, 2024

Abstract

Experimental evidence from India has shown effects of remedial education programs range from zero to +0.7 standard deviations, with implementation details affecting impact. A key question is whether the effects of successful programs implemented at small scale can persist when those programs are scaled-up. This paper reports on the impacts of a remedial program implemented in government schools in northern India, first when the program was implemented in fewer than 1,000 schools, and later when the same program was implemented in more than 50,000 schools. In the first randomized controlled trial of the program, students in treatment schools gained +0.44 SD on tests of foundational literacy and numeracy relative to control students. Heterogeneity analysis suggests that the implementer updated program activities in the last two years of the evaluation in response to results from the first year. Following the first RCT, the program design was largely unchanged, and the program was expanded to three additional states and 50x as many schools. A second RCT of the program at scale was designed and implemented using the same measurement tools and methodological approach. Results from this second RCT are pending.

Note to reviewers: Results from the second RCT will be available in June 2024 and incorporated into the paper and presentation. The pre-analysis plan for this RCT has been registered in the AEA RCT Registry #10873 (link). As of April 20, 2024, the paper below reports the results from the first RCT only.

[*] Corresponding author. *jeffery.mcmanus@idinsight.org*, +1-941-961-0711

1. Introduction

For the past ten years enrollment of children age 6 to 14 in schools in India has been consistently above 95% (ASER 2021). Yet learning remains stubbornly low. According to the Annual Status of Education Report of 2018, the most recent year that the survey included learning assessments of children, only 27% of children in grade 3 and 50% of children in grade 5 can read at the grade 2-level (ASER 2019). 72% of children in grade 5 are unable to complete a subtraction problem from the grade 2 curriculum (ibid).

Many culprits have been blamed for the lack of learning in Indian schools, including high levels of teacher absenteeism, low levels of teacher effort, and asymmetric information about educational performance between providers and parents and communities (Kremer, Chaudhury, Halsey Rogers, Muralidharan, & Hammer, 2005; Banerjee, Banerji, Duflo, Glennerster, & Khemani, 2010). A symptom and further cause of stagnant outcomes is the high variance in student learning levels within classes (Glewwe & Muralidharan, 2016). The introduction of first-generation students into the primary school system over the past few decades has further exacerbated differences between students in the same classroom (Muralidharan, 2017). Until 2018 students in primary government schools in India could not be held back, even if they failed end-of-year exams, creating classrooms with many students who had not acquired foundational learning skills and were increasingly left behind.

Heterogeneity in student abilities within the same class makes effective teaching difficult. To ensure that students at the bottom of the distribution are being adequately supported, remedial education programs have been introduced across India. Several studies have found that these programs can have large effects on helping lagging students catch up. But the magnitude of impact varies by program and context. At the high end of treatment effects, researchers found that a program that trains community volunteers to deliver two hours of after-school remedial instruction per day led to 0.75 standard deviation (SD) increases in test scores relative to a control group after 18 months (Lakshminarayana, Elbe, Bhakta, Frost, Boone, Elbourne, et al., 2013). Banerjee, Cole, Duflo, and Linden (2007) find similarly large effects from a program that recruits young women ("Balsakhis") from local communities to tutor students during school hours in basic literacy and numeracy; students who received tutoring improved test scores by 0.6 SD at the end of two years relative to control students. Reading camps run by volunteers in Uttar Pradesh led to similarly large gains in reading skills (Banerjee, Banerji, Duflo, et al., 2010).

Other remedial education programs have had more modest effects. A series of five experiments assessing ten different remedial education interventions in India that found effects on test scores ranging from near-zero and statistically insignificant (training camps in Uttarakhand) to 0.7 SD (in-school learning camps in Uttar Pradesh) (Banerjee, Banerji, Berry, Duflo, Kannan, Mukerji, et al., 2016). The authors observe that exact replications of a previously successful model can generate similar impact, but that deviations from that model – implementing during school hours versus after school, running summer camps versus programs during the school year,

implementation by volunteers versus paid contract teachers versus government teachers – can greatly influence the magnitude of the effect.

These findings are consistent with a recent literature that questions the information value of point estimates from a single experiment (e.g. Vivalt, 2015; Vivalt, 2017; Bold, Kimenyi, Mwabu, Ng'ang'a, & Sandefur, 2013; Pritchett & Sandefur, 2013). As it becomes increasingly clear that implementation details and environmental factors influence an intervention's impact, there has been a corresponding call for development programming that encourages experiential learning by implementers to figure out what works in their specific context (Pritchett, Samji, & Hammer, 2013; Andrews, Pritchett, & Woolcock, 2013; Wild & Ramalingam, 2018).

Development Impact Bonds (DIBs) have been proposed as one tool that can promote experiential learning in social programs (Gustafsson-Wright, Bogglid-Jones, Segell, & Durland, 2017). DIBs are a financing instrument in which an investor provides capital up-front to an implementer and earns a return from a donor based on the effectiveness of the program, as measured by a third-party evaluator. This set up shifts the focus away from paying for inputs – such as more teachers or more textbooks – to paying for outputs – such as student learning. Donors only pay if impact is achieved, investors receive a financial return to cover their investment risk, and implementers receive flexible funding to scale programming. In theory this flexibility could lead to more effective program innovation if implementers have the data and capacity to adapt their program over the course of the DIB.

Given the novelty of these financing instruments, however, there is little evidence on whether DIBs encourage program adaptation and ultimately improve outcomes. In this paper, we present the results of a three-year, randomized controlled trial of an education program funded by a DIB. The program, run by the Mumbai-based nonprofit Educate Girls, provided remedial instruction from 2015 to 2018 to students in grades 3 to 5 in government primary schools in Bhilwara District, India. Educate Girls recruited, trained, and managed volunteers who delivered a basic reading, math, and English curriculum two to three times per week. The UBS Optimus Foundation, acting as the investor, supplied the capital to Educate Girls in Year 1, while the Children's Investment Fund Foundation paid for educational outcomes in Year 3, and Instiglio managed the DIB. Our research team measured learning gains each year in Hindi, Math, and English through assessments administered to treatment and control students in school and at home.

While the terms of the DIB itself were not exogenously varied across treatment schools, the results of the evaluation are consistent with the hypothesis that the DIB created an environment conducive for program innovation. After one year of programming, students in the treatment group gained a modest 0.07 SD in learning levels relative to the control group, equivalent to 0.27 additional years of business-as-usual schooling in this environment. These effects were driven primarily by gains in math (+0.11 SD) and English (+0.07 SD). After extensive program innovations, by the end of Year 3 treatment students had substantially outpaced their peers, gaining on average 0.44 SD relative to control students, or 1.14 additional years of schooling. Treatment effects were largest in Math (+0.64 SD) and English (+0.60 SD). These gains accrued

even for students who only participated in the program for the final year: students who were in grade 3 in Year 3 gained 0.55 SD relative to control students, representing an additional 1.2 years of schooling across subjects, and 1.6 and 2.0 additional years of schooling in Math and English respectively.

Heterogeneity analysis suggests that some of the specific refinements that Educate Girls made in response to midline evaluation results may have made the program more effective. Besides spending more time in schools each week, in Years 2 and 3 Educate Girls volunteers shifted focus from higher-performing students to lower-performing students. We observe some evidence of a corresponding increase in treatment effects in Years 2 and 3 for students who were at the bottom of the distribution at baseline. Educate Girls also added home tutoring visits in the final year of the evaluation to complement in-school sessions. We observe that students who were more likely to be absent experienced greater treatment effects in the final year of the program. However, since these changes accompanied other changes to program implementation, and since student observable and unobservable characteristics are correlated, we interpret these results as suggestive rather than causal.

This paper provides further evidence that volunteer-based remedial education in schools can be a highly effective way of improving student test scores, but that implementation details matter. The experience of the DIB suggests that flexibility in programming, together with incentives and access to high-quality outcome data early in the evaluation may provide the enabling factors necessary for implementers to learn what works in their specific context. But many questions remain around which specific conditions created by DIBs spur greater program impact, and whether DIBs are the optimal financing instrument for generating and sustaining those conditions.

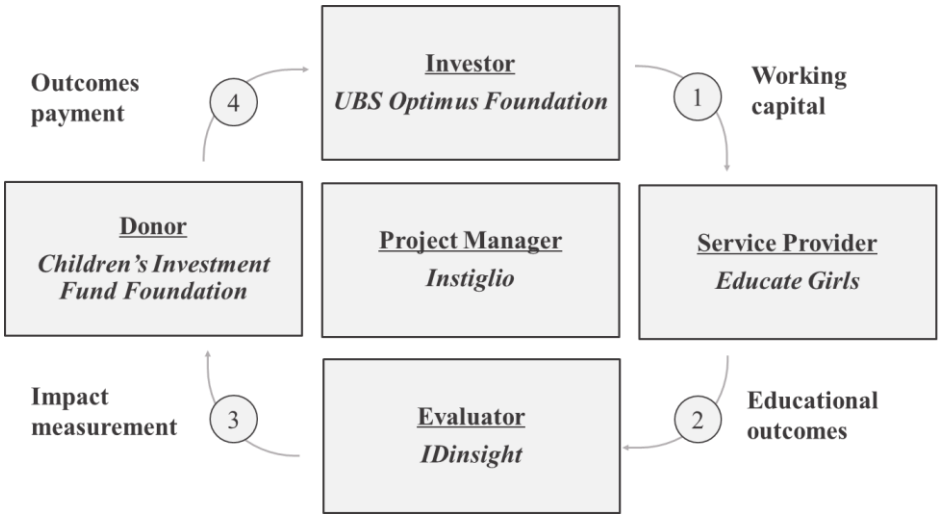
The rest of this paper is structured as follows. In Section 2 we describe the DIB and Educate Girls' program. In Section 3 we describe the design of the RCT and the data collection protocol, and in Section 4 we present the results. Section 5 concludes.

2. The Educate Girls Development Impact Bond

2.1 DIB structure

The Educate Girls DIB was launched in 2015 with two objectives: to test the DIB model as a proof of concept and to improve educational outcomes in Bhilwara district in Rajasthan, India. These goals brought together the five main partner organizations in the DIB, which are listed in **Figure 1** along with their respective roles and relationships to each other.

Figure 1: Educate Girls DIB stakeholder map



Notes: Fig. 1 lists the major stakeholders involved in the Educate Girls DIB, their roles, and their relationships to each other.

The DIB set three-year impact targets for enrollment of out-of-school girls and learning gains of boys and girls in grades 3 to 5. Under the contract terms, the UBS Optimus Foundation would disburse payments to Educate Girls in the first year of programming and would be repaid by the Children's Investment Fund Foundation proportional to Educate Girls' success against targets at the end of the third year. If Educate Girls exactly met the DIB targets then the UBS Optimus Foundation would earn a 10% internal rate of return on their investment; if they exceeded targets the maximum IRR was set at 15%. A subcontract between Educate Girls and the UBS Optimus Foundation further stipulated that the UBS Optimus Foundation would pass on an incentive payment of 32% of any return on their investment to Educate Girls. More details on how the targets were set, the financial structure of the DIB, and the roles of different stakeholders are provided on the Educate Girls DIB website (Instiglio, 2022).

For the purposes of this paper we focus on the impact of the Educate Girls' program on the second outcome – learning gains – for two reasons. First, learning gains represented the bulk of DIB payments (80%), reflecting the priorities of the Working Group to respond to the learning crisis in Indian government schools. Second, whereas learning gains were estimated through an RCT, enrollment was measured by verifying Educate Girls' documentation of newly-enrolled girls off of lists of eligible out-of-school girls in treatment villages. Due to costs, the DIB Working Group decided against preparing and verifying comparable lists of out-of-school girls and newly-enrolled girls in control villages, and so enrollment estimates may not reflect the causal effect of Educate Girls' program.

The Educate Girls DIB thus created several conditions that deviated from business-as-usual programming and could in theory influence program effectiveness, including:

- Funding that permitted (and encouraged) program adaptation, rather than contractually requiring Educate Girls to deliver the program in a certain way.
- Financial incentives to maximize impact, in the form of payments from the investor to Educate Girls conditional on the return on the investment.

- Reputational incentives to maximize impact, which were augmented by the high-visibility of the DIB.
- A three-year timeline that gave Educate Girls time to react to early year results and strengthen program implementation.
- Annual data on learning outcomes in Educate Girls' schools and in an experimental counterfactual, which showed the extent to which the program was achieving the expected results and which subgroups were benefitting the most and least. In addition to submitting detailed annual evaluation reports to Educate Girls and the Working Group, we shared de-identified evaluation data with Educate Girls to enable further subgroup analysis.

2.2 Program Description

Educate Girls' core program involved volunteer-based in-school remedial instruction. Educate Girls field coordinators recruited and trained community volunteers ("Team Balika") who delivered a basic reading, math, and English curriculum to students in grades 3 to 5 in government primary schools.

Over the course of the DIB the details of Educate Girls' delivery model evolved. Although implementation monitoring was outside the scope of our role on the DIB, we met with Educate Girls staff at various points throughout the evaluation and received updates on program implementation. Some of the key changes to the program over time, as reported by Educate Girls before the release of the Year 3 results, included the following:

- *Shift in organizational focus and resources from enrollment to learning*. Since Educate Girls stayed ahead of their yearly benchmarks to meet the three-year enrollment target, but lagged behind on learning targets, staff and volunteers gradually shifted time and resources away from enrolling out-of-school girls and toward the learning component of the program. One manifestation of this shift was that the average number of days that volunteers spent in school each week increased from two in Year 1 to three in Year 2.
- *Longer implementation period*. In Year 1, due to implementation delays, Educate Girls was only able to implement their in-school volunteer program from October to February. Implementation started two months earlier in Year 2, in August, and by Year 3 volunteers were delivering sessions in schools from July to February.
- *Curriculum overhaul*. At the end of Year 1 Educate Girls collaborated with pedagogy experts to create a new curriculum that involved sorting students into groups by learning level, rather than age, so that volunteers could deliver customized lessons to each group, with more support for the lowest-performing students. Educate Girls also incorporated more teaching aids, such as worksheets and games.
- *Home visits*. When Year 1 and Year 2 results showed that students who were frequently absent from school did not benefit from the program, volunteers added in-home tutoring sessions in Year 3 to reach serially absentee students.

Educate Girls staff later reported similar changes to programming in an interview shortly after the end of the DIB (Slobig, 2018).

3. Evaluation design

We designed a village-level clustered RCT to estimate the causal impact of the Educate Girls program on student learning.

3.1 Village sampling and randomization

Educate Girls' remedial instruction program was designed to be implemented in all schools in a village and included several non-school, village-based activities, such as enrollment of out-of-school children and door-to-door community mobilization. To mirror the program design and minimize crossovers and spillovers between treatment and control groups, we randomized treatment assignment at the village-level.

Through discussions with Educate Girls and the DIB Working Group we established criteria to identify eligible schools and villages for the program. These criteria were selected to facilitate program implementation and reduce the risk of school or village attrition. We used publicly available school records from the District Information System for Education (DISE) for the most recent school year (2014-15) to identify all active schools in the three study blocks (Mandalgarh, Jahajpur, and Bijoliya) in Bhilwara district, Rajasthan, and to filter out ineligible schools, according to the following criteria:

Program eligibility criteria for schools and villages

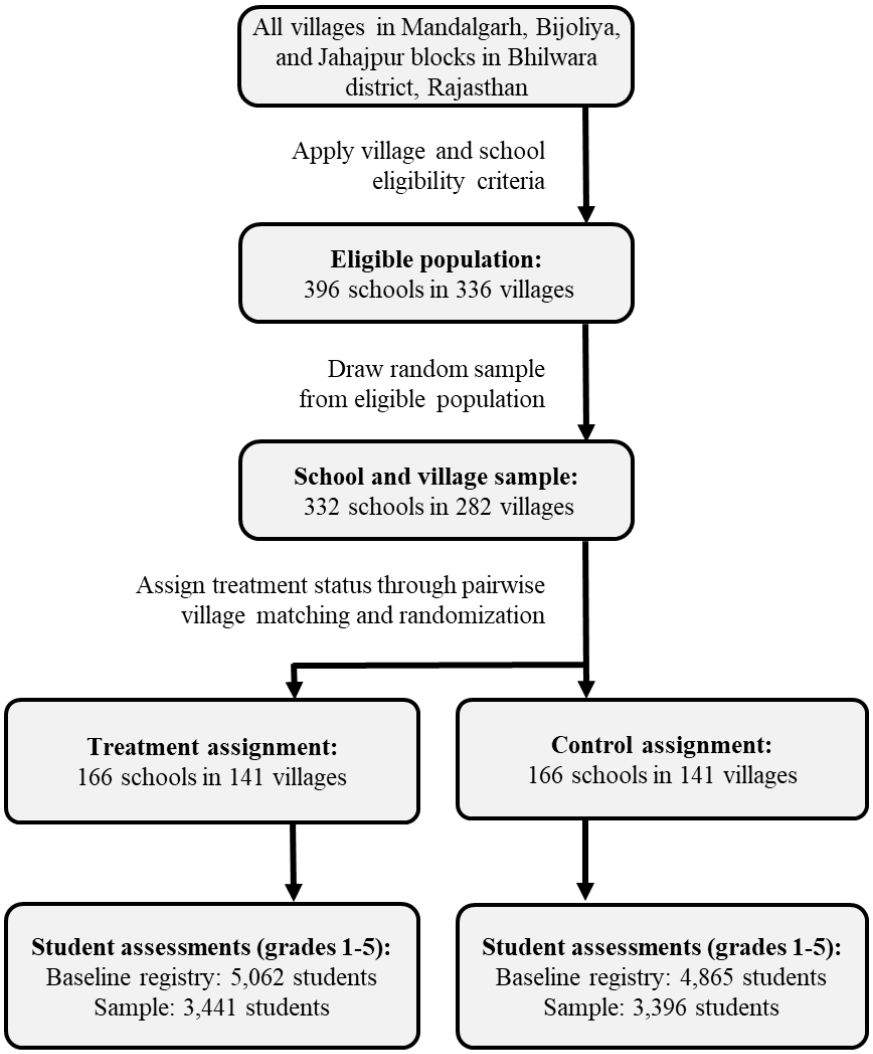
1. The school must exist in the previous year's DISE database (2013-14) since first-year schools may not have the infrastructure to accommodate the Educate Girls program.
2. The school must include grades 1 to 5 and not include secondary or higher secondary grades since the program is targeted at primary-school students (combined primary/upper primary schools were also eligible).
3. The school must be managed by the Department of Education or Local Government Bodies since securing permission from individual private schools, madrasas, Sanskrit schools, and Shiksha Karmi schools would not be possible.
4. The school must have between 10 and 60 students total in grades 3 to 5 to ensure sufficient students for the program without overburdening volunteers.
5. The medium of instruction must be Hindi since all program materials were developed and delivered in Hindi.
6. The village must have between 1 and 4 schools that meet the preceding criteria so that one volunteer can manage a full village case load.
7. All schools in the village must be labeled "rural" in the DISE dataset since the program is designed for implementation in rural areas.
8. The village must be deemed "operationally feasible." This criteria was defined by Educate Girls field staff, not DISE. Once the short list of villages from the previous criteria had been prepared, Educate Girls field staff reviewed the list and marked any villages where their program could not be delivered, primarily due to remoteness or unsafe areas. This criteria excluded 8 (2.3%) otherwise eligible villages.

Applying these criteria resulted in 342 eligible villages containing 413 eligible schools. To maximize balance and improve power, we constructed eight strata defined by the block where the village is located (Jahajpur or Mandalgarh/Bijoliya – the latter two are combined by DISE into one block labeled "Mandalgarh") and by the number of eligible schools in each village (1, 2, 3, or 4). If a stratum contained an odd number of schools, we randomly selected one school to drop from the stratum; this dropped 6 villages and 17 schools, reducing the eligible population to 336 villages containing 396 schools.

In order to further improve power, we matched villages within strata on the first component from a principal components analysis that included four school characteristics, aggregated to the village-level, that we anticipated could be correlated with learning outcomes: (i) student enrollment, (ii) presence of an upper primary school, (iii) female-to-male student ratio, and (iv) student-to-teacher ratio. During initial discussions, Educate Girls planned to implement their program in 168 villages, or half of the eligible sample. However, due to cost constraints the DIB Working Group decided to reduce the scope of implementation to 141 villages. We randomly selected 141 of the 168 village pairs to participate in the evaluation.

Within each village pair, we randomly selected one village to receive Educate Girls' program and assigned the other village to the control group. Our analytical model includes village pair fixed effects to account for this pairwise random assignment. All sampling, eligibility filtering, and randomization were conducted in Stata/IC V.13.1. The evaluation design is summarized in **Figure 2** below.

Figure 2: Sampling and randomization design



Notes: Fig. 2 shows the sampling and randomization design of the RCT, including the sizes of the student population and sample assessed for this study.

3.2 Student sampling

Over the course of the three-year evaluation we tracked five different grades of students as they progressed through school. At baseline we assessed students in grades 1 through 5. In each subsequent endline we assessed students who were then in grades 3, 4, and 5 (the target grades for Educate Girls' programming). Since a student's grade changes year to year, student cohort labels can be ambiguous; for instance, "grade 3" could refer to three different cohorts of students in the evaluation: students who were 3rd graders in Year 1, Year 2 or Year 3 of the evaluation. To remove this ambiguity we refer to student cohorts according to their grade in Year 1, unless explicitly noted otherwise. For instance, Cohort 2 refers to students who were in grade 2 during the first year of the evaluation, and who progressed to grade 3 in Year 2 and grade 4 in Year 3. **Table 1** lists all five cohorts and shows how each cohort progressed through school during the evaluation and, for the treatment group, how many years each cohort was exposed to Educate Girls' program. **Underlined** grades indicate when the cohort was assessed.

Table 1: Student cohorts during the evaluation

Student cohort label (grade in Y1)	Grade level at each year of evaluation				Years of exposure to EG program (treatment group)
	Baseline (Sept 2015)	Y1 Endline (Feb 2016)	Y2 Endline (Feb 2017)	Y3 Endline (Feb 2018)	
1	**1**	1	2	**3**	1
2	**2**	2	**3**	**4**	2
3	**3**	**3**	**4**	**5**	3
4	**4**	**4**	**5**	6	2
5	**5**	**5**	6	7	1

Notes: **Underlined** values indicate when a cohort was assessed.

Our study population consists of all students in these cohorts who were present in the 332 evaluation schools during the baseline assessment. In September 2015 our enumerators visited each of the evaluation schools and made a complete list of students who were in attendance (the 'baseline registry'). In order to improve balance and ensure sufficient samples from relevant subgroups, enumerators stratified by grade and gender in each school and randomly selected 50% of students from each stratum to assess. If a stratum had fewer than 4 students then enumerators were instructed to assess all students in that stratum. The resulting baseline sample consisted of 6,837 students (3,396 in control villages and 3,441 in treatment villages), or 69% of students on the baseline registry. **Table 2** lists the population sizes and evaluation samples for each cohort.

Table 2: Student populations (present at baseline) and samples by cohort

	Population – Present at baseline			Sample – Sampled at baseline		
Cohort	**Control**	**Treatment**	**All**	**Control**	**Treatment**	**All**
1	1028	1051	2079	690	699	1389
2	954	928	1882	660	645	1305
3	960	1006	1966	683	687	1370
4	933	1029	1962	679	696	1375
5	990	1048	2038	684	714	1398
All	4865	5062	9927	3396	3441	6837

Due to odd numbers of students in some strata and the rule that enumerators should assess all students in strata with fewer than 4 students, students from different schools and subgroups had different probabilities of being sampled. To recover population average treatment effects we weight each student observation by the inverse probability of being selected. 98% of student weights are between 1 and 2, and the remaining 2% of weights vary from 2 to 5 (with one

student's weight equal to 9) due to enumerator miscounting or other issues when stratifying students within schools. In the appendix we present results that weight all students equally, which are nearly identical to the weighted results.

Table 3 shows balance between sampled students in treatment and control groups across baseline characteristics, including test scores at baseline. As expected from random assignment and the large sample, the two groups are well-balanced. The p-value on the F-statistic from a joint test of orthogonality on the variables listed in this table is 0.60.

Table 3: Treatment-Control balance on baseline characteristics

Variable at Baseline	**Control Mean [Std Error]**	**Treatment Mean [Std Error]**
Hindi Level (1-6)	2.647 [0.057]	2.593 [0.050]
Math Level (1-5)	2.387 [0.032]	2.326 [0.029]
English Level (1-5)	1.905 [0.038]	1.871 [0.035]
Total Level (3-16)	6.939 [0.118]	6.789 [0.104]
Child Grade	2.980 [0.028]	3.019 [0.026]
Age	8.100 [0.035]	8.102 [0.038]
Female	0.482 [0.010]	0.497 [0.010]
SC_ST	0.471 [0.030]	0.478 [0.029]
Observations in Sample	3,396	3,441

In addition to the students sampled from baseline registries, during the Year 2 and Year 3 endlines we compiled a separate list of students who were enrolled but not on the baseline registry. This supplementary list consists of two types of students - those who were absent but enrolled at baseline and those who newly enrolled over the course of the evaluation – though we cannot distinguish between these two types since we do not have reliable enrollment registers from baseline. For the purposes of calculating DIB payments we assessed all 5,421 of these additional students (2,390 in control villages and 3,031 in treatment villages). However, we do not include these students in our estimation of causal effects since treatment could plausibly have induced some of these students to enroll who would not have enrolled otherwise; in fact, Educate Girls was financially incentivized to do this by the DIB. Even if there were a similar number of students in treatment and control in this group, the program could have encouraged different types of students (such as female students or students from lower castes, or students with certain unobservable characteristics) to enroll than would have otherwise.

Including these students in our sample could undermine the comparability of treatment and control, and so we exclude them from our estimates below.

3.3. Data collection

We assessed students using the Annual Status of Education Report (ASER) tool, a well-known testing instrument administered to children throughout India. The ASER assessment tests foundational literacy and numeracy, which for our study population consisted of three subjects: Hindi, Math, and English. Each subject is assessed on 5 levels, ranging from beginner to more advanced competencies in each subject, corresponding to a possible score of 1 to 5 points for each subject. In selecting an assessment tool we were particularly concerned about "ceiling effects", in which students who obtained the highest score on a section could have scored even higher if the test included more advanced competencies. This would lead to underestimates of a student's true ability (and potentially underestimates of treatment effects). To partially mitigate ceiling effects we added one additional level to the Hindi section (what we call "Story Plus") since previous ASER data indicated that a sizeable fraction of older students tend to max out on the Hindi section, which is only intended to cover Grade 1 and 2 competencies. In the Results section we discuss the implications of other ceiling effects on our estimates.

Table 4 lists the competencies/levels for each subject. Different versions of the assessment, all from ASER's online database, were administered each year. The specific version administered each year was not announced to teachers or Educate Girls in advance to minimize the risk of teaching-to-the test. **Appendix A** contain images from the ASER assessment administered during the Year 3 endline.

Table 4: Learning levels measured by ASER

Level	Hindi	Math	English
1	Beginner	Beginner	Beginner
2	Letters	Numbers 1-10	Capital letters
3	Words	Numbers 11-99	Lowercase letters
4	Paragraph	Subtraction	Words
5	Story 1	Division	Sentences
6	Story Plus	—	—

To assess students, we recruited, trained, and managed local enumeration teams. All data were recorded in SurveyCTO modules on Android-based smartphones, though materials for the ASER assessment were printed on paper and shown to students during the test. We obtained verbal informed consent from all headmasters, teachers, and students to conduct the assessment. If a student in our sample was not present in school on the day of the endline assessment then enumerators went to their homes, obtained parental consent, and assessed them there. 17.8% of assessments over the course of the three endline surveys were conducted at students' homes, and 34.2% of students were assessed at home at least once. We exploit this heterogeneity to estimate the difference in treatment effects for students who were present at school versus habitually absent.

Due to this intensive home follow-up protocol for students who were absent, as well as the relative stability of the student population, attrition from the evaluation was low. **Table 5** compares student attrition in treatment and control groups at each endline. Attrition is highest in Year 3 when we were unable to assess 12.9% of students. Higher attrition that year was primarily driven by 252 students (129 in control villages, 123 in treatment villages) who were held back from advancing to Grade 3 – likely influenced by the Ministry of Education's change in policies regarding student detention – and instructions to enumerators to only assess students in Grades 3 to 5. Student detention from Grade 3 is not correlated with treatment status (p-value = 0.71). Omitting these students, attrition in Year 3 was 7.2% (7.2% in control villages, 7.1% in treatment villages).

Table 5: Student attrition by year

Year	Attrition rate		p-value of difference in attrition rates (T-C)	p-value of F-stat on joint orthogonality test with non-attrited students
	Treatment	Control		
1	1.6%	2.0%	0.41	0.63
2	4.9%	4.7%	0.91	0.63
3	12.8%	13.1%	0.53	0.48

The difference in attrition rates between treatment and control is always small and statistically insignificant. We also ran treatment-control balance checks on the non-attrited students each year across the same variables listed in **Table 3**. The p-value on the F-statistic from a joint test of orthogonality is never statistically significant, indicating that in addition to attrition rates being similar in treatment and control, the same types of students attrited from treated and control groups.

3.4 Analytical model

To estimate the effect of the Educate Girls program on learning outcomes we run the following regression specification:

$$Y_{iscvn} = \beta_0^* + \beta_1^* T_v + \beta_2^* Y_{iscv0} + X'_{iscv0}\beta^* + \alpha'_p\beta^* + \varepsilon^*_{iscvn}$$

where

- Y_{iscvn} denotes student *i*'s test score in subject *s* in cohort *c* in village *v* after *n* years of the evaluation. Test scores are normalized relative to the standard deviation of test scores for control students in subject *s* in cohort *c* in village *v* after *n* years of the evaluation.
- T_v denotes the treatment status of village *v* (1 = treatment, 0 = control).
- Y_{iscv0} denotes student *i*'s test score at baseline, normalized relative to the control group.
- X'_{iscv0} denotes a vector of student covariates at baseline, containing whether the student is from a scheduled caste/scheduled tribe; whether the student is female; and dummy variables for a student's age in years.

- α_p' denotes a vector of dummy variables corresponding to village pairs, which were defined for pairwise randomization. β_0^* denotes the coefficient of the omitted village pair.
- ε_{iscvn}^* denotes the error term for student *i*, clustered at the village level *v*, which was the level of treatment assignment.
- $*$ denotes the sampling weights applied to each student observation, which is equal to the inverse probability of being sampled from all eligible students in their grade-gender-school strata among students present at baseline. As shown the appendix, results are very similar when students are assigned equal weights.

We estimate the effect of the program on each cohort after n-years (i.e. after 1 year, 2 years, and 3 years). Although nth-year effects (such as the effect of the program in Year 2) are inherently interesting, given the experimental set-up we cannot obtain consistent nth-year point estimates, except for in Year 1, for two reasons. First, as a result of program impact in earlier years, the learning levels of students in treatment and control are different at the beginning of Years 2 and 3. Second, we did not assess students at the beginning of each school year after Year 1, and so Years 2 and 3 cover longer time periods, including school break for summer. Even if we had test scores at the beginning of Years 2 and 3, we would not obtain consistent estimates by controlling for start-of-year test scores: students who start at the same level in Years 2 and 3 may not be comparable across treatment and control. That being said, we infer large differences in n-year point estimates as suggestive of changes in program effectiveness.

We defined the experimental design and analytical model in a pre-analysis plan posted on 3ie's RIDIE registry before baseline data collection had ended.[1] While the final design parameters are consistent with the pre-analysis plan – including the program being evaluated, the treatment arms, the randomization procedure, student sampling, and how outcomes were defined and measured – we note a few deviations in the analytical model. The pre-analysis plan was prepared for the purposes of the DIB, and so the primary estimator was an *aggregate* treatment effect, which was calculated by summing the learning gains of students in treatment schools and subtracting the learning gains of student in control schools. This estimator accounts for differences in the number of students in treatment and control schools each year and thereby incentivized Educate Girls to increase enrollment. The baseline levels of all newly-enrolled students were imputed as the lowest possible score on the ASER test, which likely overstated growth but doubly incentivized Educate Girls to try and enroll children into school.

For the analysis in this paper we deviate from the pre-analysis plan by estimating average treatment effects using an ANCOVA model (i.e. including baseline learning levels on the right hand side of the regression), which gives us more power than a difference-in-differences estimator. To further improve power we control for the covariates described above, and to mirror the randomization design we include village pair fixed effects. We define outcomes in terms of standard deviations (SDs) of the control group, rather than raw ASER levels, for comparability with other programs. Finally, as described in the student sampling section, we

[1] RIDIE Study ID: 56042ebbd220d

focus our analysis on the population of students present in baseline register, the largest group for whom we can consistently estimate causal effects.

4. Results

4.1 Main results

Table 6 presents the results from estimating the regression specification above for each cohort and subject 1, 2, and 3 years after baseline. Although we report results in terms of control-group SDs, our Year 2 and 3 results may understate impact relative to other 1-year evaluations reported in SDs. Since variance in test scores increases over time, the denominator in the normalization transformation increases even as the raw difference between treatment and control grows. For instance, if a program had a 1-year effect of 0.2 SD and an equivalent additional effect in the second year, the combined 2-year effect would be less than 2*0.2 = 0.4 SD due to increasing variance of test scores. In this evaluation, test scores for students in the control group who were assessed each year (i.e. Cohort 3) had virtually the same variance in Hindi across years, but variance in Math scores grew 67% and variance in English scores grew 46% between the Year 1 and Year 3 endlines.

For this reason, we also report results in terms of additional equivalent years of schooling (EYOS). This intuitive metric describes how many additional years of schooling treatment students grew in each subject relative to their business-as-usual peers in the control group. Additional EYOS are defined as:

$$\frac{Avg\ treatment\ effect\ in\ ASER\ levels}{Avg\ change\ in\ ASER\ scores\ in\ the\ control\ group} * (\#\ years\ baseline\ to\ endline\ N)$$

For instance, if the average treatment effect for cohort X in subject Y is 0.5 ASER levels after 3 years, and the average control student grew two levels after three years, then the treatment effect in terms of EYOS is (0.5/2)*3 = 0.75.

Table 6: Average treatment effects in SDs and EYOS

Subject	*Cohort*	*ATEs by the end of…*					
		Year 1		*Year 2*		*Year 3*	
		SD	*EYOS*	*SD*	*EYOS*	*SD*	*EYOS*
Hindi	**1**					0.207*** [0.052]	0.52
	2			0.066 [0.043]	0.19	0.045 [0.046]	0.12
	3	0.032 [0.033]	0.10	0.092** [0.037]	0.30	0.116*** [0.039]	0.32
	4	0.040 [0.028]	0.13	0.062* [0.032]	0.23		

	5	0.008 [0.027]	0.03				
	Pooled	0.027 [0.018]	0.10	0.070*** [0.022]	0.24	0.108*** [0.028]	0.30
Math	**1**					0.771*** [0.075]	1.63
	2			0.073* [0.044]	0.21	0.675*** [0.050]	1.91
	3	0.046 [0.035]	0.18	0.150*** [0.038]	0.49	0.607*** [0.042]	1.92
	4	0.084** [0.039]	0.38	0.192*** [0.048]	0.67		
	5	0.168*** [0.034]	0.96				
	Pooled	0.108*** [0.023]	0.54	0.150*** [0.028]	0.51	0.638*** [0.032]	1.86
English	**1**					0.707*** [0.070]	1.95
	2			0.091* [0.051]	0.29	0.578*** [0.053]	1.79
	3	0.022 [0.033]	0.07	0.148*** [0.046]	0.50	0.577*** [0.053]	1.69
	4	0.081** [0.036]	0.27	0.231*** [0.042]	0.85		
	5	0.097*** [0.033]	0.44				
	Pooled	0.066*** [0.020]	0.25	0.167*** [0.031]	0.60	0.598*** [0.037]	1.84
Total	**1**					0.550*** [0.058]	1.20
	2			0.096*** [0.035]	0.25	0.413*** [0.047]	1.02
	3	0.053** [0.025]	0.16	0.162*** [0.031]	0.47	0.462*** [0.039]	1.20
	4	0.075*** [0.026]	0.24	0.171*** [0.031]	0.55		
	5	0.105*** [0.026]	0.43				
	Pooled	0.074*** [0.016]	0.27	0.141*** [0.020]	0.43	0.440*** [0.029]	1.14

Notes: Each cell represents the coefficient on treatment in a regression specification for that cohort and subject at the end of each year. Standard errors clustered at the village-level are in brackets below coefficients. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Results from regressions without sampling weights are presented in **Appendix Table B.1**.

After one year of the program, treatment effects were modest but statistically significant: pooling across all cohorts and subjects, students in Educate Girls schools had 0.07 SD higher test scores than students in control schools, representing 0.27 additional years of schooling relative to students in the control group. Gains were largest in Math and English, and positively correlated with grade level.

After two years of the program, treatment effects were similar in size and, for students who were in the second year of the program, added to first-year gains. For each cohort in Year 2, the pooled coefficient on treatment was comparable or slightly larger than the sum of their treatment effect in Year 1 and the treatment effect of their peers in the next grade up in Year 1. For instance, the treatment effect for Cohort 4 students, who were completing grade 5 at the end of Year 2, was 0.17 SD, comparable to their treatment effect in Year 1 when they were in grade 4 plus the treatment effect of their Cohort 5 peers in Year 1. The treatment effect for Cohort 2 students at the end of Year 2, after completing their first year in the program, was 0.10 SD, larger than their Cohort 3 or 4 peers at the end of Year 1 but comparable to their Cohort 5 peers at the end of Year 1. Treatment effects were slightly larger in English and, for the first time, statistically significant in Hindi. Given increasing variance in test scores, and possible convergence between treatment and control during summer breaks, this suggests that programming in Year 2 may have been slightly more effective than in Year 1.

Year 3, on the other hand, witnessed a dramatic increase in the effectiveness of the program. Across all subjects and cohorts, students in treatment schools outpaced their peers in control schools more than in the previous two years. Students who experienced the program in Year 3 gained on average 0.44 SD relative to control students by the end of the three-year evaluation, representing 1.1 additional years of schooling. Despite having only participated in the program for one year, Cohort 1 students (who were in grade 3 during the final year) gained 0.55 SD relative to control students, representing 1.2 additional years of schooling, and 1.6 and 2.0 additional years of schooling in Math and English respectively.

While impressive, these results may understate the true impact of the program each year, especially Year 3, for two reasons. First, test scores probably decayed between school years, and probably more for students in Educate Girls schools who had higher end-of-year scores. Treatment effects in Years 2 and 3 are therefore likely more than the change in treatment effects year-to-year. Second, despite the additional difficulty level in Hindi, some students hit the max score each year, and over time more treatment students hit this ceiling, both in absolute numbers and relative to the control group. **Table 7** lists the percent of students obtaining the highest score each year in each subject.

Table 7: Percent of students achieving top score on each assessment

	Baseline		**Year 1 Endline**		**Year 2 Endline**		**Year 3 Endline**	
Subject	**C**	**T**	**C**	**T**	**C**	**T**	**C**	**T**
Hindi	10%	9%	23%	21%	27%	28%	31%	32%

Math	4%	2%	11%	11%	11%	12%	11%	31%
English	3%	2%	7%	5%	8%	12%	10%	27%

By the end of Year 3, nearly three times as many treatment students as control students were achieving the maximum score on the Math and English sections. Assuming that some of these students would have scored even higher if the assessment had included additional difficulty levels for these competencies, treatment students would have gained even more relative to the control group.

A simple example illustrates the approximate scale of the downward bias caused by ceiling effects on treatment estimates. Suppose that each student who hit the ceiling, regardless of whether they were in treatment schools or control schools, would have a 50% chance of moving up one level, and each one of them a 50% chance of moving up another level, and so on. **Appendix Table B.2** shows the average treatment effects estimates that would result from this simulation. Across all subjects, estimated EYOS at the end of Year 3 would be ~12% larger. Estimates of EYOS for Math and English would be approximately 2.1 years for students at the end of Year 3 instead of 1.8.

4.2 Heterogeneous effects by baseline performance

To examine the effectiveness of the new curriculum's pivot toward lower-performing students in Years 2 and 3, we estimate treatment effects by a student's starting competency level, combining all cohorts. Results are presented in **Table 8**. Since students who start at a higher level have less room to improve, we would not obtain valid results if we defined outcomes in the same way as above – as SD gains on the ASER assessment – for this analysis. In fact, defining outcomes as SD gains results in monotonic decreases in treatment effects from the bottom to the top of the baseline distribution, though this is driven largely by the mechanics of ceiling effects rather than true heterogeneity.

Instead, to examine heterogeneous effects by baseline level, we define learning gains as a binary for whether the student moved up at least one level from baseline to endline. Although this outcome is coarser than SD gains, it does not suffer from ceiling effects for any levels except for students who start at the top level; these students are consequently excluded from the analysis.

Table 8: ATEs by baseline performance

Subject	**Baseline learning level**	**% sample**	**% moved up at least ONE level by…**					
			Year 1		**Year 2**		**Year 3**	
			Control	**ATE**	**Control**	**ATE**	**Control**	**ATE**
Hindi	**Beginner**	34.7%	0.417	0.029 [0.048]	0.581	0.081** [0.033]	0.778	0.086*** [0.018]
	Letter recognition	28.5%	0.472	0.076*** [0.028]	0.688	0.047* [0.025]	0.873	0.036** [0.018]
	Word recognition	4.2%	0.738	0.186** [0.090]	0.770	0.081 [0.093]	0.938	0.003 [0.059]

	Paragraph fluency	15.0%	0.470	0.059 [0.039]	0.670	0.085** [0.043]	0.896	0.065* [0.035]
	Story fluency	8.0%	0.457	0.007 [0.044]	0.787	-0.038 [0.067]	0.913	0.100 [0.116]
	Story+ fluency	9.6%						
Math	**Beginner**	20.7%	0.729	-0.042 [0.110]	0.829	0.092** [0.036]	0.928	0.023** [0.011]
	Numbers 1-9 recognition	39.0%	0.338	0.002 [0.025]	0.555	0.019 [0.023]	0.755	0.140*** [0.017]
	Numbers 10-99 recognition	27.9%	0.251	0.126*** [0.028]	0.476	0.199*** [0.031]	0.633	0.321*** [0.031]
	Subtraction	9.2%	0.255	0.041 [0.048]	0.453	0.136 [0.098]	0.686	0.104 [0.222]
	Division	3.2%						
English	**Beginner**	52.3%	0.362	0.027 [0.027]	0.569	0.059** [0.024]	0.730	0.122*** [0.017]
	Capital letter recognition	14.3%	0.670	0.037 [0.034]	0.702	0.056 [0.038]	0.885	0.062** [0.024]
	Small letter recognition	29.2%	0.182	0.034 [0.024]	0.277	0.187*** [0.037]	0.487	0.307*** [0.038]
	Word recognition	1.9%	0.338	-0.023 [0.137]	0.560	1.184 [1.603]	0.625	0.000 [0.000]
	Sentence fluency	2.3%						

Notes: Each cell in the ATE (average treatment effect) columns represents the coefficient on treatment in a regression specification for that baseline competency at the end of each year. Standard errors clustered at the village-level are in brackets below coefficients. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

In the first year of the program, students who were in the middle of the distribution in Hindi and Math at baseline experience the most statistically significant gains from treatment, whereas treatment effects in English are spread across baseline levels and are not statistically significant for any one level. By the end of Year 2, students who were at the lower end of the distribution in all subjects begin to experience statistically significant treatment effects. By the end of Year 3 these gains were cemented for students who initially started in the three lower levels in Math and English, though students who started in the middle of the distribution were still by far the most likely to move up due to treatment. Students in the second-highest level were not significantly more likely to move up a level compared to control students for any subject. A similar pattern emerges when assessing the likelihood of moving up at least two levels, though this analysis is more restrictive since it omits the top two competencies from each subject (see results in **Appendix Table B.3**).

Overall, these results suggest that the program was most effective at reaching students toward the middle of the distribution, though over time the program generated learning gains for students at the bottom of the distribution as well.

4.3 Heterogeneous effects by student attendance

Educate Girls added home visits in Year 3 of the program to reach students who were habitually absent. To explore the effectiveness of these visits, we estimate treatment effects by the number of times that a student was absent from school on the day of an endline (and was tested at home) in **Table 9**. While few in number, our unannounced visits for assessing students provide an unbiased measurement of a student's propensity to miss school. We restrict this analysis to students who were assessed in at least two endlines (i.e. Cohorts 2, 3, and 4) and therefore on whom we have multiple observations of attendance.

Across these cohorts, 66% of students were never absent for an endline, 24% were absent once, 9% were absent twice, and 1% were absent three times. We combine students who were absent twice or three times into one group of 'habitual absentee students' since there are too few 3-absence students for heterogeneity analysis. While in theory Educate Girls' program could affect student absenteeism, in practice we see little difference in rates of absenteeism between treatment and control: The average number of absences is 0.47 among treatment students and 0.43 among control students (p=0.28). Moreover, absentee students appear similar on observables across treatment and control. There are no significant differences across treatment and control groups for 1-absence or 2-absences students in terms of baseline test scores, caste category, gender, or age: The p-value on the F-statistic from a joint orthogonality test with these variables is 0.677 for 1-absence students and 0.929 for 2-absence students.

Table 9: ATEs by surveyed at home vs at school

Absences	**ATEs by the end of…**		
	Year1	**Year2**	**Year3**
No absences	0.068*** [0.022]	0.159*** [0.024]	0.461*** [0.035]
1 absence	0.097** [0.038]	0.140*** [0.039]	0.524*** [0.068]
2+ absences	-0.114 [0.115]	0.084 [0.081]	0.449*** [0.142]

Notes: Each cell represents the coefficient on treatment in a regression specification for that subgroup at the end of each. Standard errors clustered at the village-level are in brackets below coefficients. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Treatment effects were similar for no-absence and 1-absence students across all three years; a single absence may not be indicative of a student's general propensity to miss school. On the other hand, students who were absent two or three times did not benefit at all from treatment in the first two years, but by the end of Year 3 had almost benefited as much as their no-absence peers. While attendance during our spot checks may not be representative of attendance throughout the school year, and absenteeism may be correlated with other characteristics that affect how much a student benefits from treatment, these findings are consistent with the narrative that the addition of home tutoring sessions in Year 3 effectively reached serially absent students.

In the appendix we present further heterogeneity analysis by gender (**Appendix Table B.4**) and caste (**Appendix Table B.5**). Treatment appears to benefit different types of students similarly across years, with treatment effects slightly larger for female students in Year 3 and slightly smaller for SC/ST students in Year 1; both results are significant at the 10% level but not the 5% level. The lack of further heterogeneity results are consistent with the lack of subgroup targeting by Educate Girls.

5. Discussion

Our results provide further evidence that remedial education programs can be a highly effective way of improving student test scores, but that implementation details matter. At the end of Year 1, treatment effects were modest and comparable to experimental estimates of less-effective remedial education interventions in India (e.g. summer camps in Bihar and Uttarakhand in Banerjee, Banerji, Berry, et al., 2016). By the end of Year 3, treatment effects were large and comparable to the most effective interventions (e.g. the treatment-on-the-treated estimates of the Balsakhi program in Banerjee, Cole, Duflo, & Linden, 2007). Students who only participated in the program during the final year experienced gains that were comparable to their peers who were in their second and third years of the program. These gains were equivalent to an additional year of business-as-usual schooling. Even more impressive, these results likely understate the magnitude of Year 3 treatment effects due to ceiling effects and increasing variance in test scores.

A similar narrative emerged from our annual reports to the DIB Working Group, though that analysis included a second student population (those who were absent at baseline and newly enrolled students) and a different analytical model and estimator. At the end of the first year of programming, Educate Girls had achieved 26% of the three-year learning target, and only 52% of the target by the end of the second year. Despite uncertainty at the start of Year 3 about whether Educate Girls could meet the DIB targets, due to exceptionally strong performance in the final year Educate Girls not only met the three-year learning target but exceeded it by 60%. Educate Girls also exceeded the enrollment target by 16%, triggering the maximum return on investment to the UBS Optimus Foundation and maximum incentive payments to themselves.

The experience of the DIB suggests that flexibility in programming, together with incentives and access to high-quality outcome data early in the evaluation, may provide the enabling factors necessary for implementers to learn what works in their specific context. But questions remain around which specific conditions created by DIBs spur greater program impact.

For instance, to what extent were improvements in Educate Girls' program driven by conditions created by the DIB – financial and reputational incentives, flexible program funding, and the availability of rigorous data on performance – versus simply having more time to solve implementation issues? Some changes, such as the curricular overhaul and the addition of home visits, were plausibly the result of learnings from data supplied to Educate Girls in early years.

Other changes, such as lengthening the implementation calendar and increasing the number of volunteer days per week, may have also occurred in a non-DIB setting as Educate Girls worked out implementation obstacles and allocated more resources to the learning program. Financial and reputational incentives could have put pressure on Educate Girls to solve all of these problems more quickly. Further research, ideally experimental, is needed to assess the causal effects of the overall DIB model and its individual components.

Even if the DIB model enables program innovation and generates impact, other questions remain around general equilibrium effects and scalability. For instance, what effect does participating in a DIB have on an implementer's non-DIB funded activities? There could be positive knowledge spillovers if implementers adopt the lessons from DIB-funded activities to non-DIB activities. Or there could be negative spillovers if implementers divert resources and focus away from non-DIB activities to meet the high-stakes targets of the DIB. Programmatic improvements over the course of the DIB might be sustained after the DIB ends and financial incentives are removed. Or implementers might return to business-as-usual programming. It is unclear if DIBs are even the optimal pay-for-results financing instrument. Simpler arrangements between donors and service providers may generate the same conditions with lower transaction costs and no risk premiums. While the results of the first DIB are encouraging, many questions remain before we recommend widespread adoption of this new financing instrument.

CRediT authorship contribution statement

Lucas Kitzmüller: Software, Formal analysis, Investigation, Writing – Original Draft. **Jeffery McManus:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – Original Draft, Writing – Review & Editing. **Neil Buddy Shah:** Conceptualization, Methodology, Supervision, Funding acquisition. **Kate Sturla:** Methodology, Investigation, Writing – Original Draft, Project administration.

Acknowledgements

We thank Ryan Fauber, Girish Tripathi, Nadeen Hamza, Elizabeth Bennett, and Soni Jha for excellent research assistance and field management. James Berry, Marc Shotland, and Pieter Serneels provided helpful comments. We thank Instiglio for excellent management of the Educate Girls Development Impact Bond.

References

Andrews, M., Pritchett, L., & Woolcock, M. (2013). Escaping Capability Traps Through Problem Driven Iterative Adaptation (PDIA). *World Development*, *51*, 234–244. https://doi.org/10.1016/J.WORLDDEV.2013.05.011

ASER Centre (2019). Annual Status of Education Report (Rural) 2018 Provisional, January 15, 2019, Full Report. http://www.asercentre.org/

ASER Centre (2021). Annual Status of Education Report (Rural) 2021, November 17, 2021, Full Report. http://www.asercentre.org/

Banerjee, A. V., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2016). Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India. *NBER Working Paper Series*, No. 22746. http://www.nber.org/papers/w22746

Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India. *American Economic Journal: Economic Policy*, *2*(1), 1–30. https://doi.org/10.1257/pol.2.1.1

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *Quarterly Journal of Economics*. https://doi.org/10.1162/qjec.122.3.1235

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2013). *Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education*. *SSRN*. https://doi.org/10.2139/ssrn.2241240

Business Insider (2018). Schools in India can once again detail students if they fail their exams. Published July 19, 2018. https://www.businessinsider.in/schools-in-india-can-once-again-detain-students-if-they-fail-their-exams/articleshow/65052599.cms

Glewwe, P., & Muralidharan, K. (2016). Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications. In *Handbook of the Economics of Education*. https://doi.org/10.1016/B978-0-444-63459-7.00010-5

Gustafsson-Wright, E., Boggild-Jones, I., Segell, D., Durland, J. (2017). Impact Bonds in Developing Countries: Early Learnings from the Field. Center for Universal Education at Brookings. September 2017.

Instiglio (2022). Educate Girls Development Impact Bond Website. Accessed on 8 September 2022. http://instiglio.org/educategirlsdib/

Kremer, M., Chaudhury, N., Halsey Rogers, F., Muralidharan, K., & Hammer, J. (2005). Teacher Absence in India: A Snapshot. *Journal of the European Economic Association*, 3 (2-3). https://doi.org/10.1162/jeea.2005.3.2-3.658.

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). https://doi.org/10.1371/journal.pone.0065775

Pritchett, L., Samji, S., & Hammer, J. S. (2013). *It's All About MeE: Using Structured Experiential Learning ('e') to Crawl the Design Space*. *SSRN*. https://doi.org/10.2139/ssrn.2248785

Pritchett, L., & Sandefur, J. (2013). *Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix*. *SSRN*. https://doi.org/10.2139/ssrn.2364580

Slobig, Z. (2018). Listen: Safeena Husain on Breaking the Cycle of Illiteracy. Skoll Foundation. http://skoll.org/2018/08/03/development-impact-bond-educate-girls-results/

Vivalt, E. (2015). Heterogeneous treatment effects in impact evaluation. *American Economic Review: Papers & Proceedings*, *105*(5), 467–470. https://doi.org/10.1257/aer.p20151015

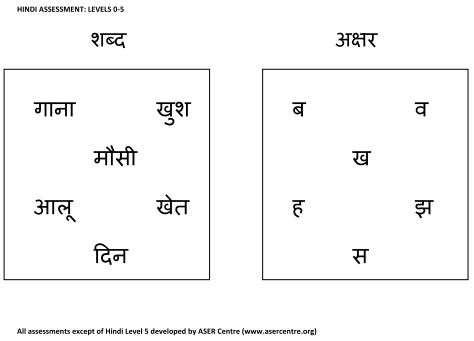
Vivalt, E. (2017). *How Much Can We Generalize from Impact Evaluations?* Retrieved from

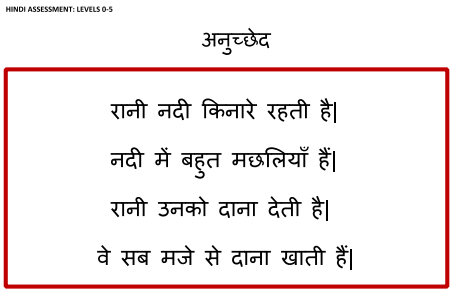
http://evavivalt.com/wp-content/uploads/How-Much-Can-We-Generalize.pdf

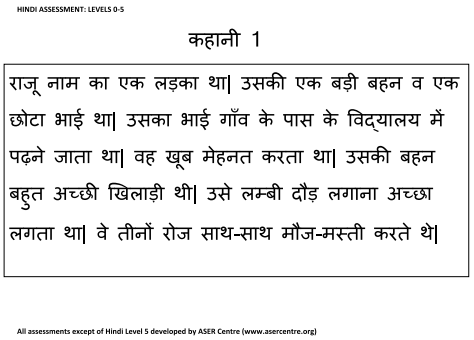
Wild, L., & Ramalingam, B. (2018). *Building a global learning alliance on adaptive management*.

Appendix A

Figure A.1: ASER Testing Tool for Hindi in Year 3 Endline







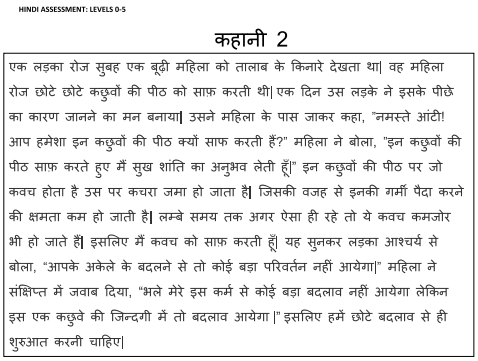
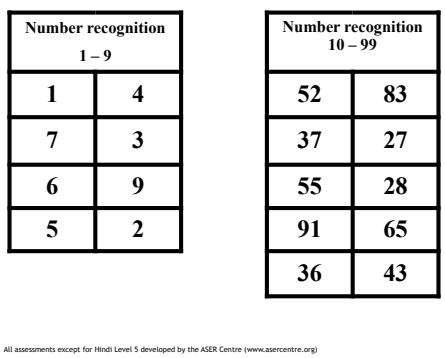


Figure A.2: ASER Testing Tool for Math in Year 3 Endline



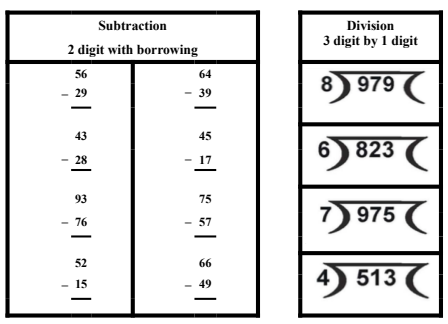
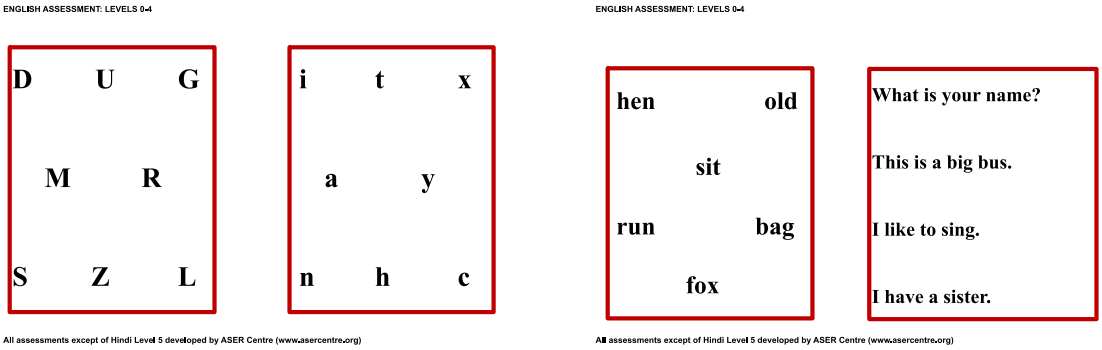


Figure A.3: ASER Testing Tool for English in Year 3 Endline



Appendix B

Table B.1: Main effects without sampling weights

Subject	*Cohort*	*ATEs by the end of…*					
		Year 1		*Year 2*		*Year 3*	
		SD	*EYOS*	*SD*	*EYOS*	*SD*	*EYOS*
Hindi	**1**					0.158*** [0.054]	0.38
	2			0.054 [0.041]	0.15	0.036 [0.045]	0.09
	3	0.026 [0.032]	0.08	0.099*** [0.037]	0.30	0.127*** [0.038]	0.33
	4	0.034 [0.027]	0.11	0.074** [0.033]	0.26		
	5	0.007 [0.027]	0.03				
	Pooled	0.023 [0.017]	0.08	0.071*** [0.023]	0.23	0.098*** [0.029]	0.25
Math	**1**					0.731*** [0.072]	1.47
	2			0.084* [0.045]	0.22	0.717*** [0.051]	1.90
	3	0.044 [0.037]	0.16	0.161*** [0.039]	0.49	0.630*** [0.041]	1.85
	4	0.102*** [0.037]	0.46	0.183*** [0.046]	0.59		

	5	0.184*** [0.034]	1.03				
	Pooled	0.118*** [0.023]	0.57	0.148*** [0.028]	0.47	0.655*** [0.032]	1.78
English	**1**					0.715*** [0.070]	1.81
	2			0.114** [0.052]	0.35	0.596*** [0.054]	1.66
	3	0.049 [0.034]	0.16	0.200*** [0.046]	0.64	0.643*** [0.052]	1.77
	4	0.092** [0.036]	0.30	0.242*** [0.044]	0.82		
	5	0.094*** [0.034]	0.39				
	Pooled	0.080*** [0.021]	0.29	0.192*** [0.031]	0.65	0.632*** [0.038]	1.78
Total	**1**					0.514*** [0.059]	1.04
	2			0.099*** [0.035]	0.24	0.424*** [0.047]	0.96
	3	0.058** [0.026]	0.16	0.185*** [0.032]	0.50	0.498*** [0.039]	1.21
	4	0.078*** [0.025]	0.24	0.176*** [0.032]	0.53		
	5	0.110*** [0.026]	0.42				
	Pooled	0.080*** [0.016]	0.27	0.147*** [0.021]	0.43	0.450*** [0.030]	1.08

Notes: Each cell represents the coefficient on treatment in a regression specification for that cohort and subject at the end of each year. Standard errors clustered at the village-level are in brackets below coefficients. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B.2: Main effects with ceiling simulation

Subject	***Cohort***	***ATEs by the end of…***					
		Year 1		***Year 2***		***Year 3***	
		SD	***EYOS***	***SD***	***EYOS***	***SD***	***EYOS***
Hindi	**1**					0.244*** [0.050]	0.66
	2			0.091* [0.047]	0.29	0.044 [0.047]	0.13
	3	0.009 [0.035]	0.03	0.081** [0.036]	0.27	0.087** [0.039]	0.27

	4	-0.001 [0.029]	0.00	0.057* [0.033]	0.20		
	5	0.012 [0.035]	0.04				
	Pooled	0.011 [0.020]	0.04	0.070*** [0.022]	0.24	0.108*** [0.027]	0.33
Math	**1**					0.868*** [0.089]	1.98
	2			0.039 [0.044]	0.12	0.702*** [0.055]	2.34
	3	0.038 [0.036]	0.15	0.082** [0.034]	0.32	0.511*** [0.043]	2.04
	4	0.091** [0.043]	0.38	0.107** [0.044]	0.41		
	5	0.130*** [0.033]	0.51				
	Pooled	0.098*** [0.023]	0.43	0.091*** [0.026]	0.35	0.600*** [0.033]	2.14
English	**1**					0.775*** [0.079]	2.17
	2			0.120** [0.055]	0.38	0.622*** [0.055]	2.06
	3	0.019 [0.032]	0.06	0.164*** [0.050]	0.59	0.535*** [0.053]	1.95
	4	0.063* [0.035]	0.22	0.214*** [0.045]	0.85		
	5	0.017 [0.032]	0.07				
	Pooled	0.035* [0.020]	0.13	0.176*** [0.032]	0.68	0.598*** [0.039]	2.10
Total	**1**					0.618*** [0.062]	1.40
	2			0.110*** [0.039]	0.30	0.440*** [0.049]	1.16
	3	0.036 [0.026]	0.10	0.140*** [0.031]	0.42	0.446*** [0.038]	1.27
	4	0.051* [0.027]	0.14	0.145*** [0.031]	0.45		
	5	0.070** [0.029]	0.20				
	Pooled	0.055*** [0.017]	0.17	0.129*** [0.019]	0.40	0.453*** [0.029]	1.28

Notes: Each cell represents the coefficient on treatment in a regression specification for that cohort and subject at the end of each year. Standard errors clustered at the village-level are in brackets below coefficients. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table B.3: ATEs by baseline performance (likelihood of moving up at least TWO levels)

Subject	Baseline learning level	% moved up at least TWO levels by…					
		Year 1		Year 2		Year 3	
		Control	ATE	Control	ATE	Control	ATE
Hindi	Beginner	0.098	-0.002 [0.029]	0.245	0.029 [0.031]	0.502	0.106*** [0.028]
	Letter recognition	0.354	0.070*** [0.025]	0.529	0.058** [0.028]	0.782	0.018 [0.024]
	Word recognition	0.123	0.009 [0.054]	0.410	-0.170* [0.098]	0.738	0.196 [0.138]
	Paragraph fluency	0.172	0.015 [0.025]	0.469	0.055 [0.046]	0.732	0.022 [0.076]
	Story fluency						
	Story+ fluency						
Math	Beginner	0.188	-0.066 [0.113]	0.166	0.011 [0.050]	0.380	0.196*** [0.033]
	Numbers 1-9 recognition	0.035	0.037*** [0.013]	0.135	0.093*** [0.021]	0.282	0.404*** [0.025]
	Numbers 10-99 recognition	0.045	0.020* [0.012]	0.149	0.034 [0.029]	0.262	0.417*** [0.046]
	Subtraction						
	Division						
English	Beginner	0.199	0.065*** [0.023]	0.293	0.035* [0.021]	0.527	0.190*** [0.023]
	Capital letter recognition	0.030	0.012 [0.015]	0.079	0.059* [0.036]	0.230	0.488*** [0.047]
	Small letter recognition	0.043	0.002 [0.010]	0.163	0.087*** [0.031]	0.320	0.257*** [0.048]
	Word recognition						
	Sentence fluency						

Notes: Each cell in the ATE (average treatment effect) columns represents the coefficient on treatment in a regression specification for that baseline competency at the end of each year. Standard errors clustered at the village-level are in brackets below coefficients. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table B.4: Treatment effects by gender

	ATEs by the end of…		
	Year1	Year2	Year3

Treatment	0.083***	0.122***	0.405***
	[0.021]	[0.029]	[0.039]
Female	-0.000	-0.002	-0.035
	[0.021]	[0.028]	[0.035]
Treat*Female	-0.019	0.039	0.072
	[0.031]	[0.039]	[0.057]
Observations	4069	3871	3571

Notes: Each column represents a regression of normalized test scores at the end of each year on the list of variables in the first column and the controls described in the regression specification above. Standard errors clustered at the village-level are in brackets below coefficients. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table B.5: Treatment effects by caste

	ATEs by the end of…		
	Year1	**Year2**	**Year3**
Treatment	0.103***	0.146***	0.450***
	[0.022]	[0.031]	[0.045]
SC_ST	-0.011	-0.028	-0.053
	[0.020]	[0.032]	[0.043]
Treat*SC_ST	-0.063*	-0.011	-0.021
	[0.035]	[0.048]	[0.069]
Observations	4069	3871	3571

Notes: Each column represents a regression of normalized test scores at the end of each year on the list of variables in the first column and the controls described in the regression specification above. Standard errors clustered at the village-level are in brackets below coefficients. * p < 0.1, ** p < 0.05, *** p < 0.01.