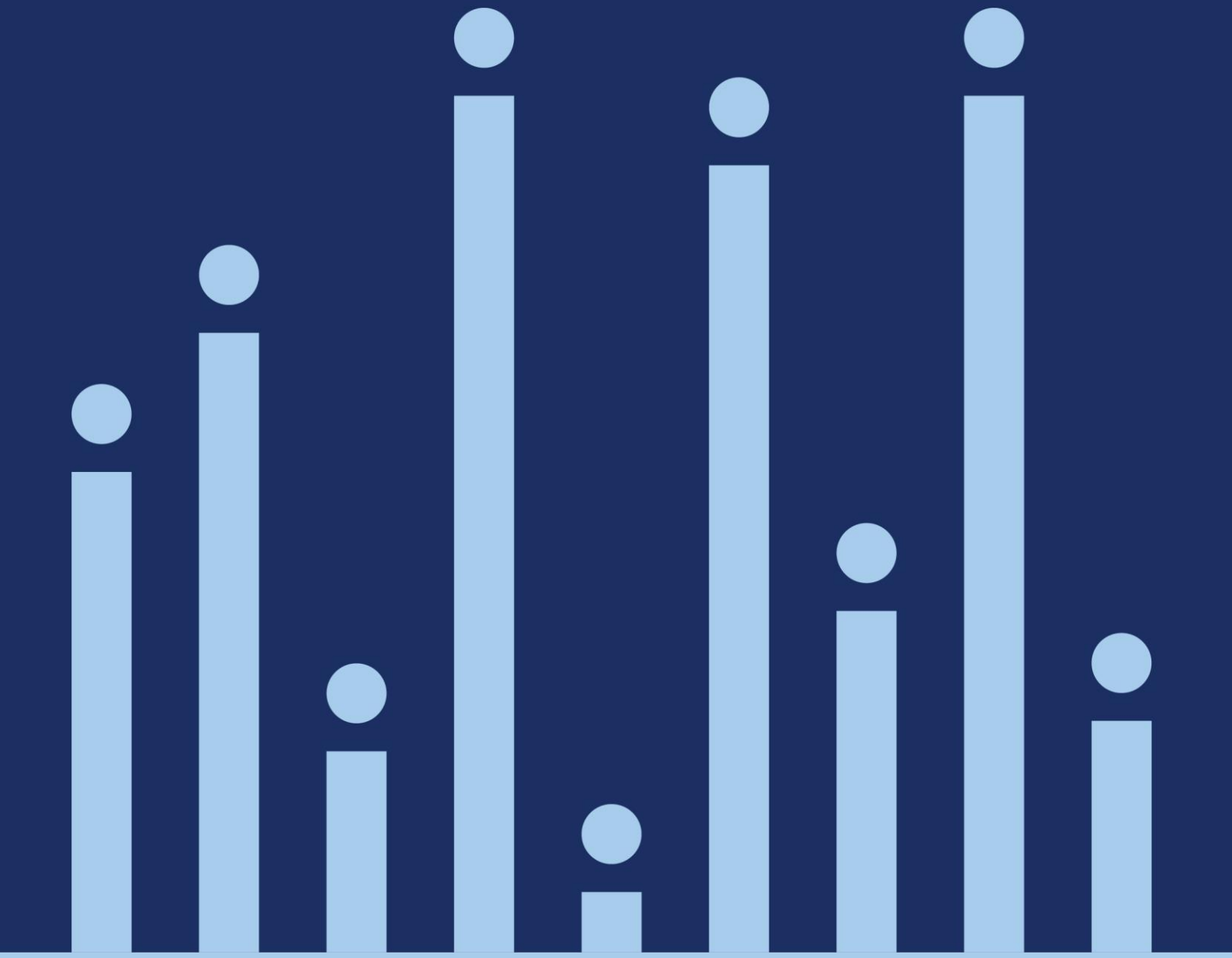


Luminos Program Impact Evaluation

Randomized controlled trial of an accelerated learning program for out-of-school children in Liberia

IDinsight

Luminos Program Impact Evaluation
Randomized controlled trial of an accelerated learning program for out-of-school children in Liberia

November 15, 2023

Authors
Kenna Mokobi: **kenna.mokobi@IDinsight.org**
Serigne Fallou Deme: **serigne.fallou-deme@IDinsight.org**
Ewoud Nijhof: **ewoud.nijhof@IDinsight.org**
Mico Rudasingwa: **mico.rudasingwa@IDinsight.org**
Jeff McManus: **jeffery.mcmanus@IDinsight.org**
Cassandre Pignon: **cassandre.pignon@IDinsight.org**
Leah Mwai: **leah.mwai@IDinsight.org**

Acknowledgements
We would like to thank Luminos, Q&A Inc. and the Liberian Ministry of Education for their commitment to realizing true social impact, based on data and evidence; Luminos for coordinating and implementing the accelerated learning program's curriculum in schools, Q&A for the support provided to collect the data and evidence and the Liberian Ministry of Education for being in support of the initiative and guidance in gaining ethical clearance. We'd also like to thank the implementing partners Restoring Our Children's Hope (ROCH) and Liberia Institute for the Promotion of Academic Excellence (LIPACE) for assisting Luminos by implementing the program, providing support for our team on the ground and sharing data for the evaluation. Finally, we would like to thank Yarvo Jones-Boakai, our external consultant, for providing guidance and support of this project.

About IDinsight
IDinsight uses data and evidence to help leaders combat poverty worldwide. Our collaborations deploy a large analytical toolkit to help clients design better policies, rigorously test what works, and use evidence to implement effectively at scale. We place special emphasis on using the right tool for the right question, and tailor our rigorous methods to the real-world constraints of decision-makers.

IDinsight works with governments, foundations, NGOs, multilaterals and businesses across Africa and Asia.

We work in all major sectors including health, education, agriculture, governance, digital ID, financial access, and sanitation.

We have offices in Dakar, Lusaka, Manila, Nairobi, New Delhi and Rabat. Visit www.IDinsight.org and follow on Twitter @IDinsight to learn more.

Contents

Executive Summary

This report describes an impact evaluation of the Luminos program in Liberia conducted by IDinsight over the 2022-23 school year.

The Luminos program is an accelerated learning program that teaches children basic reading and numeric skills and supports socioemotional development. Luminos focuses on children who have never been to school or have been out of school for several years. The program uses a learning approach that is adaptive to individual learning levels, offers diverse interaction and play-based learning, and is centered around intensive structured instruction. Luminos does not charge any fees and provides free daily lunch.

Our impact evaluation consists of a randomized controlled trial (RCT) that compared out-of-school children (OOSC) in 50 treatment communities where the program was offered in the 2022-23 school year to 50 control communities where the program was not offered. We assessed 1502 OOSC at baseline and endline on literacy (using the Early Grade Reading Assessment, or EGRA) and numeracy (using the Early Grade Math Assessment, or EGMA). We administered an extended survey that included a module on socioemotional learning (SEL) to 324 OOSC per community. We also assessed 348 government schoolchildren (GSC) in grades 1, 2, and 3 from the nearby primary school (and administered the SEL module to 3 GSC) in every community to provide a benchmark for learning gains in the OOSC sample.

The results show large, significant learning gains on all tasks in literacy and numeracy in the treatment communities compared to the control communities. On average, treatment OOSC were able to read 29 words per minute compared to 7 words per minute for control OOSC at endline. Treatment OOSC correctly answered twice as many addition questions and twice as many subtraction questions than control OOSC. 93% of children in our treatment sample who were offered the program attended at least some classes, and so results are similar for children offered the program (intent-to-treat, or ITT estimates) as for children who attended the program (treatment-on-the-treated, or TOT estimates). Effects were similar in size for girls vs boys, younger vs older children, children who were previously enrolled in school vs dropouts, and children who started with lower baseline learning levels vs higher baseline learning levels. We also observe positive learning gains in all treatment communities that exceed learning gains in almost all control communities, indicating that the program is having positive impact in all communities where it is implemented.

Compared with government schoolchildren in the same communities, children in the Luminos program started the school year with much lower literacy and numeracy scores, but ended the school year with similar numeracy scores and substantially higher literacy scores than their peers in school. We find few significant treatment effects or changes over time in the SEL results. However, we expect that these null results are due to the validity of the research instrument in this specific context rather than reflecting a true null effect of the program.

Our results show that the Luminos program has large impacts on foundational literacy and numeracy skills for out-of-school children after ten months. Effect sizes are on the upper end of effects measured in RCTs of other remedial initiatives and structured pedagogical programs. We plan to conduct a follow-up round of data collection in one year to assess the persistence of these large learning effects and to assess whether students in treatment communities have successfully enrolled in government schools.

Introduction

IDinsight is a research organization that provides social sector leaders with evidence to improve their programs. The Luminos Fund is an international NGO that helps out-of-school children (OOSC) catch up to grade level and reintegrate into government schools. The Luminos Liberia program (previously known as Second Chance) provides 10 months of intensive classroom instruction to OOSC to integrate students into government schools at the end of the program. The Luminos program began in Ethiopia and expanded to Liberia in 2016.

Background

The Luminos program teaches children basic reading and numeric skills and supports socioemotional development. Luminos focuses primarily on children who have never attended school or have been out of school for at least the last two years. The Luminos program uses a learning approach that, relative to the government school curriculum, is more adaptive to individual learning levels, offers diverse interaction and play-based learning, and is centered around intensive, structuredlessons. For reading, Luminos uses a phonics-based approach. Within each community, Luminos recruits facilitators from the community to be community teachers. Teachers receive 25 days per year of intensive training, as well as additional coaching. Both students and facilitators are intensively monitored throughout the school year by Luminos' supervisors. Luminos does not charge any fees and provides free daily school lunch in partnership with the national school feeding program, Mary's Meals.

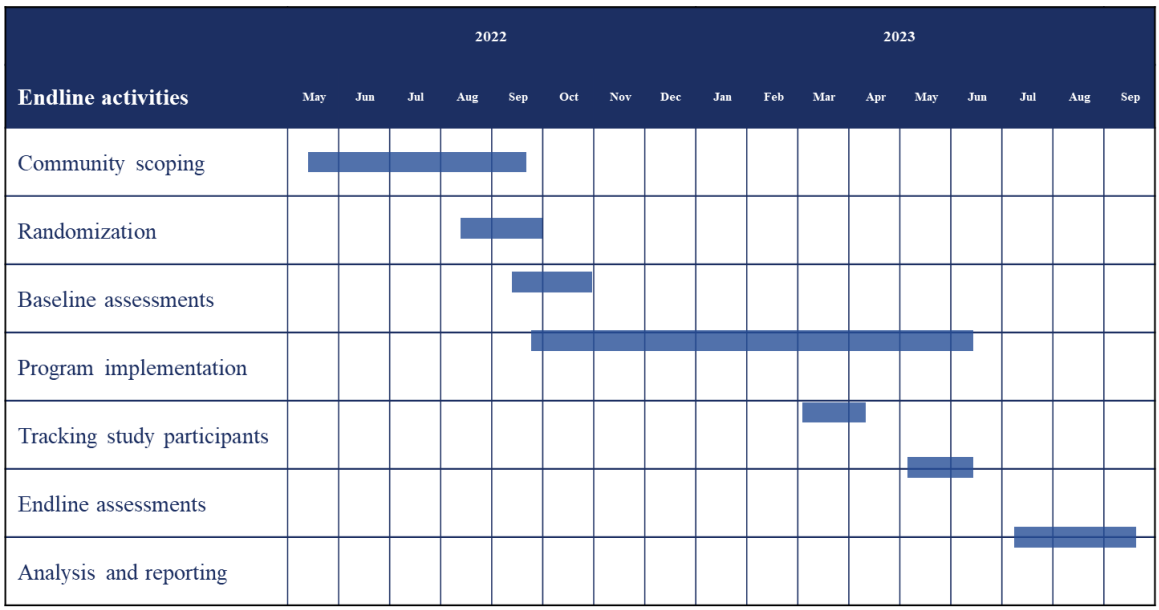
The Luminos program operates in an environment with one of the lowest enrolment rates in the world, where an estimated 22% to 41% of school-age children are out of school.[1] In addition, many students in school are overaged or drop out without acquiring basic skills such as reading and simple arithmetic. The Luminos program aims to help OOSC rapidly acquire foundational literacy and numeracy skills so that children are prepared to (re-)enroll in government schools and effectively engage with the curriculum.

[1] Global Education Monitoring Report, UNESCO Institute for Statistics, 2023, UNESCO.

Evaluation process

IDinsight partnered with Luminos to conduct a randomized controlled trial (RCT) to assess the impact of the Luminos program on children's learning. The RCT consists of a 'treatment group' of 50 communities that received the Luminos program in the 2022-23 school year and a 'control group' of 50 comparable communities that did not receive the Luminos program in the 2022-23 school year. The RCT was implemented in four phases. First, Luminos and two of their largest implementing partners in Liberia, Restoring our Children's Hope (ROCH) and Liberia Institute for the Promotion of Academic Excellence (LIPACE), conducted scoping of suitable communities for the program and created lists of eligible children in all communities between May and September 2022. Second, IDinsight randomized communities into control and treatment groups and conducted baseline data collection, between August and October 2022. Third, as the program was implemented between September 2022 and June 2023, IDinsight and Q&A (the survey firm) conducted regular follow-ups, focusing mainly on control communities, to facilitate tracking students during the endline. Finally, IDinsight conducted endline data collection at the end of the program and analyzed the differences between baseline and endline between May 2023 and September 2023. Table 1 summarizes the evaluation timeline.

Table 1: Timeline of program and evaluation activities



Study Design

To estimate the impact of the Luminos program, the RCT was conducted in 100 communities across 5 Liberian counties, namely, Bomi, Grand Cape Mount, Margibi, Montserrado and Bong. We estimate program impact by comparing outcomes for OOSC in treatment communities, where the Luminos program was implemented in the 2022-23 academic year, with similar OOSC in control communities, where the program was not implemented. We summarize the randomization and sampling processes below.

Research questions

Our primary research question is the following:

A. What is the causal impact of the Luminos program on learning outcomes for out-of-school children in Liberia?

We used the Early Grade Reading Assessment (EGRA) to assess foundational literacy skills, such as recognizing sounds, letters, words, and reading and listening comprehension. We used the Early Grade Mathematics Assessment (EGMA) to assess foundational numeracy skills, such as recognizing numbers, quantity discrimination and arithmetic. Children were assessed at baseline and endline using both EGRA and EGMA. A detailed description of EGRA and EGMA subtasks and assessed competencies can be found in Table 2 below.

The secondary research questions for the evaluation were as follows:

A. What is the causal impact of the Luminos program on social-emotional learning (SEL)?
B. How do treatment effects of the Luminos program vary by subgroups of interest, such as by gender, baseline performance, county, and implementing partner?
C. How do learning gains compare for OOSC in the Luminos program and government school children in the same communities?

We measured several socioemotional learning (SEL) skills, including empathy, open-mindedness, self-concept, and self-awareness, using a battery of questions drawn from the ISELA Self-Concept and Empathy modules, the Dweck Growth Mindset scale, part of the ISELA Learning Environment Safety module, and customized questions regarding attitude and behaviors towards education and learning. A table describing the SEL indicators we measured in our survey is included in the appendix (table A10).

Table 2: Literacy and numeracy competencies assessed

Early Grade Reading Assessment			
	Subtask (# of items in subtask)	**Early Reading skills**	**Description of task - Instruction**
1	Letter name identification (100)	Ability to recognize letters and accurately speak the corresponding name	Children attempt to read the names of 100 letters in a 10x10 grid on a piece of paper within 1 minute (Timed task)
2	Phonemics (10)	Ability to discriminate between initial sounds in spoken words	Children are told three words and must identify the word that starts with a different sound. The exercise is repeated 10 times with different words (Untimed)
3	Familiar word reading (50)	Children read words at the appropriate grade level, likely words they have seen before	Children are asked to read familiar/sight words out loud for 1 minute (Timed task)
4	Non-word reading (50)	Children use their skill to sound out letters and their connections reading out made-up words	Children are asked to decode nonsense words by sounding out letters and their connections for 1 minute (Timed task)
5	Passage reading (60)	Children must use a number of the above skills to read out words and follow sentence structure	Children are asked to read a short passage on a topic that is familiar to them, as quickly and accurately as possible, within 1 minute (Timed task)
5a	Reading comprehension (5)	Children must comprehend the passage they have just read	If the child was able to read the previous passage, then the child is asked 5 questions about the story that they just read. If the child could not read the previous passage, then this subtask is marked as zero (Untimed)
6a	Listening comprehension (3)	This is a measure of oral language skills	The enumerator reads a passage to the child, making sure to hide the text from the child, then asks the child questions about the text (Untimed)
Early Grade Mathematics Assessment			
	Subtask (# of items in subtask)	**Early Math skills**	**Description of task - Instruction**
1	Number identification (30)	Recognition of numerals	Children are asked to identify and name written numerals between 0 and 999 for 1 minute (Timed)
2	Quantity discrimination (10)	Understanding quantity and number sense	Children compare two numbers and identify which number is larger. The exercise is repeated 10 times with different numbers (Untimed)
3	Missing number (10)	Understanding sequential number patterns	Children are asked to fill in the missing number when given two other numbers in the set (Untimed)
4	Addition (15)	Basic understanding of addition	Children are asked to solve addition problems with one-digit and two-digit numbers, within 1 minute (Timed task)

5	Subtraction (15)	Basic understanding of subtraction	Children are asked to solve subtraction problems with one-digit and two-digit numbers, within 1 minute (Timed task)
6	Word problem (5)	Ability to do simple arithmetic in the context of real-world problems	The enumerator reads short scenarios featuring simple addition problems to the child (Untimed)

We also sampled and assessed government-school children (GSC) from the same communities to provide a benchmark to learning gains in the treatment and control groups (see Appendix A1 for full side-by-side results). Although OOSC and GSC are not strictly comparable groups, a comparison of learning gains and SEL skills offers suggestive evidence of how the program compares to learning in schools and whether the program prepares children for enrolment into schools. Finally, Luminos provided attendance data, which we use to determine how much of the program children in the treatment group received.

Randomization

The implementing partners (IPs), LIPACE and ROCH, identified 115 communities that met Luminos' eligibility criteria for the program. A community was considered eligible if it had support for the program from leaders and parents within the community, had at least 30 OOSC (i.e., enough to fill one class) 8-14 years old, was located a maximum of 3 hours from the main road and had an available physical space to run the program.[2] The IPs identified eligible OOSCs in each of the 115 communities. Approximately half of these communities had more than 60 OOSC, making them eligible for two 30-child classes. One community had enough children to run three classes, while the remainder had enough children to run one class.

Luminos and IDinsight selected 100 of the 115 communities for the evaluation, and IDinsight randomized them to treatment and control groups. Due to agreements between Luminos and the IPs about the number of classes each IP had to operate in the 2022-23 academic year, the treatment group plus non-sampled communities had to include 55 anticipated classrooms for LIPACE and 50 anticipated classrooms for ROCH. After the full evaluation is completed, Luminos' intends to also offer the program to communities assigned to the control group.

To balance the allocation to treatment and control across each IP and geographically, we stratified communities according to IP, anticipated number of classes per community, and county.[3] The final community sample consisted of ten strata (IP x # classes x county).[4] The probability of being treated varies across strata but is constant within each stratum. We include

[2] For the Luminos program, children are considered to be out-of-school if they have not attended school in the last 2 years.

[3] We would have ideally stratified on baseline scores as well. However, due to the government school calendar and program calendar, there would not have been sufficient time for the IPs to prepare for program launch after baseline data collection.

[4] There were 2 IPs, communities had one, two or three expected classes, and there were 5 counties, though the counties mostly differed between IPs.

strata fixed effects and probability weights in our analysis to account for the randomization process. Figure 1 shows the location of treatment and control communities.

Figure 1: Treatment and control communities in Liberia

Sampling

For sampling OOSC, Luminos and its implementing partners shared with IDinsight a list of roughly 35 children per expected class per study community.

We aimed to sample 20 OOSC per community, stratified by gender and age (i.e., a binary indicator that indicated whether a child was above/below 10 years using age data collected during scoping), with a target sample size of 2,000 OOSC. Following baseline data collection our sample consisted of 1,756 OOSC. The difference between our target and the final sample was due to the challenges with enforcing eligibility criteria and the limited availability of children, meaning just over 50% of OOSC reached could not be assessed. Unavailability was mostly due to children's participation in ongoing traditional schools,[5] which made them inaccessible to enumerators; and relocation of some children between scoping and the baseline data collection.

[5] Traditional schools, within Liberia often known as "bush schools", are temporary, informal schools in which children receive some introduction about and often initiation into the traditional cultures in an area. Sessions are restricted to leaders and students and researchers are strongly discouraged from interrupting or entering traditional schools, making the children inaccessible.

In addition, 35% of children initially sampled did not fit the program eligibility criteria, mainly because they reported being enrolled in school within the last two years. However, after discussions with Luminos, it was determined that these children were eligible for the program that year. Thus, we assessed 238 of these 'exception' children and included them in our sample.

Whenever a sampled student could not be assessed, we replaced that student with another from the list of replacement students within the same student stratum. Due to the reasons mentioned above, replacement was quite common: 46.5% of students in our final baseline sample were from the replacement list. Replacement did not vary significantly across treatment (47.4%) and control (45.6%) communities at baseline. Our baseline sample consists of 1,745 OOSC (902 treatment and 843 control) from 49 treatment and 49 control communities.[6]

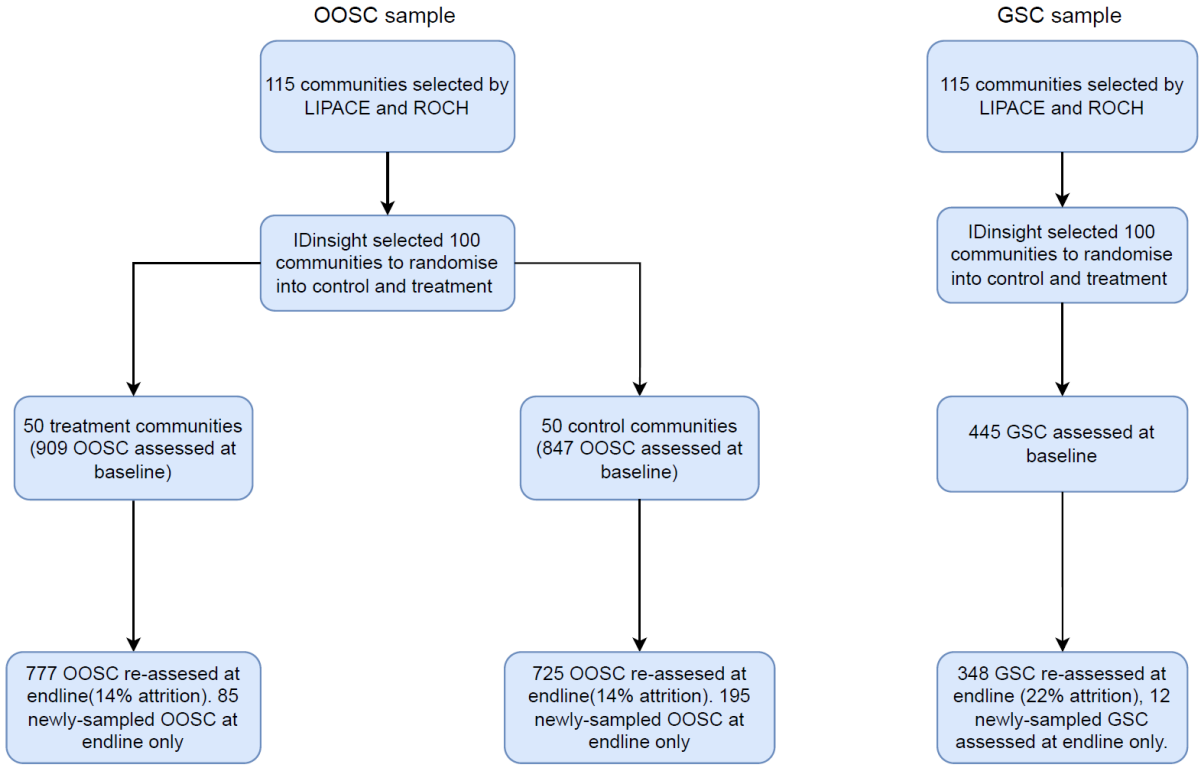
At endline we attempted to follow up with all OOSCs assessed at baseline. We successfully followed up with 1,502 OOSC (777 treatment and 725 control). Study attrition was balanced across treatment (13.9%) and control (14.0%). Given the challenges at baseline with availability, as well as concerns about potentially high attrition rates, we sampled additional children at endline in communities with fewer than 15 OOSC assessed at baseline. These children came from the original sampling lists but do not have baseline scores in our dataset. We assessed 195 of these endline-only children from control communities and 85 from treatment communities. In practice, since attrition rates were balanced and lower than expected, our primary analytical models focus on the 1,502 OOSC who were assessed at baseline and endline. In the appendix we replicate results including endline-only children.

In addition to the OOSC sample, we aimed to assess 5 GSC per community, stratified by grade (1, 2, or 3), for a target sample of 500 GSC. Our GSC sample came from the nearby “link school” in each community, which is the government partner school where children who graduate from the Luminos program would get enrolled. Our final baseline sample consisted of 445 GSC. The difference between our target and the sample was because some link schools had not started classes (despite the school year having started) or did not have sufficient children enrolled in grades 1-3 for sampling. At endline we successfully followed up with 348 GSC. Attrition of GSC in our sample was balanced across treatment (20.9%) and control (22.6%) communities, and slightly higher than OOSC. The higher attrition rates were likely driven by the fact that endline data collection occurred as schools were closing and because we had less contact data on GSC to be able to find and survey them at endline if they were no longer enrolled in or present at school.

Our randomization and sampling procedure are summarized in Figure 2.

[6] One community in treatment (Samie Town) and one community in control (Gohnzoeduah) did not have any eligible OOSC available for assessment at baseline. In the case of Samie Town, traditional schools were active at the onset of the program, making the children unavailable for at least several weeks. Implementors of the program did not operate a program and barred enumerators from entering the community in September. During the endline, enumerators were able to sample OOSC children in both communities.

Figure 2: Randomization and sampling for OOSCs (left) and GSCs (right)



Data collection

Baseline data collection started on September 26 and was completed by October 14. An important precursor to the baseline data collection was the community outreach as done by Luminos and its implementing partners. This ensured that communities were aware that teams of enumerators were coming and that they were welcomed. This community outreach was a pivotal part of making sure data collection could be started, and teams were welcomed into communities, especially in control communities, where the program would not be implemented until the RCT is completed. [7]

Following from baseline, Q&A (the survey firm) conducted in-person follow-ups as well as follow-ups over the phone in control communities for OOSC. This effort was made to reduce potential attrition from the study due to relocation of children. Luminos tracked students' participation within the treatment group through regular recording of children's attendance to the program, so there was no need for additional follow-ups in treatment communities.

Endline data collection started on May 29th, 2023, which was a week earlier than planned since the Ministry of Education announced that government schools were closing one week earlier

[7] Control communities will be offered the program in the year 2024-25 where feasible, following the final round of data collection for this study in May/June 2024

than planned. 34 enumerators and 5 supervisors conducted surveying for 4 weeks following a joint training session held by IDinsight and Q&A.

During data collection, Q&A and IDinsight staff conducted spot checks of field teams. We also ran high-frequency checks to track attrition and survey quality. Since few parents had mobile phones, supervisors reached out to community leaders in advance to inform parents that enumerators would be entering the community to speak to certain children, and request parents to exclude children from farming activities for that day. Teams on the ground also conducted revisits to many of the communities with smaller teams to increase the number of sampled children who were assessed.

Data analysis

Following data collection, we processed and analyzed the data in Stata. We implemented the analytical models that we pre-specified for this RCT on the public AEA RCT registry (https://www.socialscienceregistry.org/trials/10649). We report two sets of results. First, we report the effect of the program on OOSC in communities assigned to the treatment group compared to OOSC in communities assigned to the control group. These 'intent-to-treat' (ITT) effects reflect the impact of the program on the average child who is eligible to join the program. Second, we report the effect of the program on OOSC who participated in the program. Since 93% of our OOSC treatment sample attended at least some Luminos classes, and since no control OOSCs joined Luminos classes, these 'treatment-on-the-treated (TOT) effects are very similar to the ITT effects. Our full analytical models for ITT and TOT estimation are in the appendix.

Evaluation Results

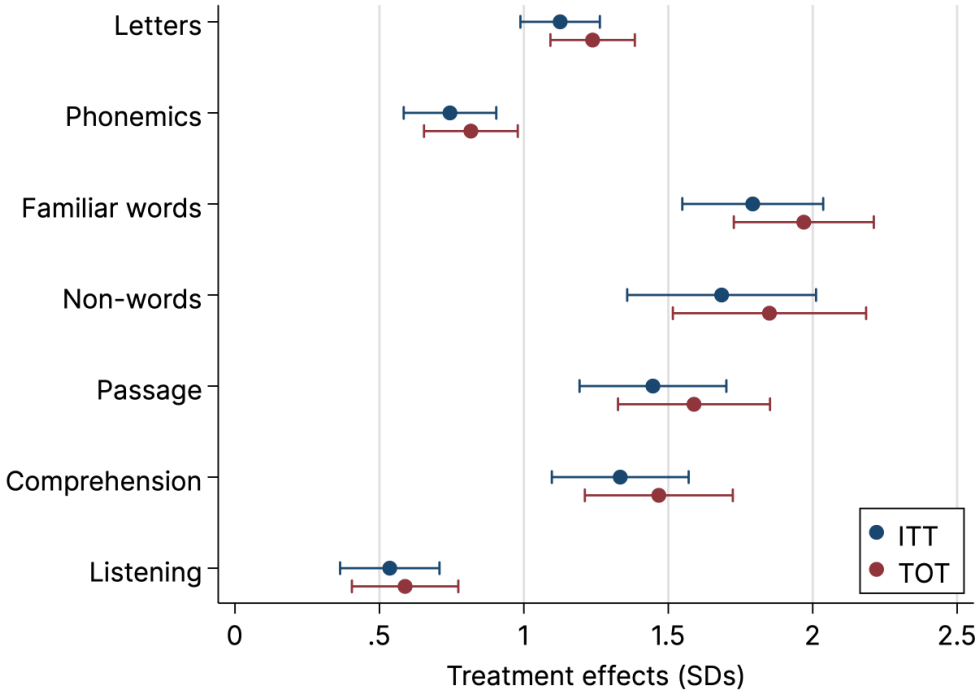
The Luminos program had large, positive effects on literacy and numeracy for OOSC. The results are statistically significant and vary in magnitude, but we see positive effects across all EGRA and EGMA subtasks. We also observe positive effects across all types of students and in all treatment communities. Compared to GSC in the same communities, treatment OOSC started the year with lower reading and math scores but ended the year with similar math scores and higher reading scores. GSC in treatment and control communities had similar learning growth, indicating that there were few spillovers from Luminos classrooms to link schools. We do not find consistent effects on the socioemotional learning modules, but as described below, we believe that this was driven by issues with the SEL instrument in this specific context rather than reflecting a true null effect of the program.

All results are listed in the results table in the appendix tables. We summarize these results below.

Literacy

Figure 3 shows standardized effect sizes (both ITT and TOT) for each of the EGRA subtasks, along with 95% confidence intervals. Standardized effect sizes larger than 0.5 SD are generally considered large for education programs. The Luminos program had large effects across all EGRA subtasks. Treatment effects were between 0.5 SD to 2.1 SD for reading subtasks, and specifically were 1.6 SD for the number of simple words that the child could read per minute. Effects are relatively more modest for reading comprehension and listening comprehension, but still large (> 0.5 SD). All treatment effects (ITT and TOT) are statistically significant at the 1% level. When we measure treatment effects by taking into account the varying difficulty of EGRA sub-tasks using item response theory (IRT), we find that treatment effects in SDs remain positive and significant at the 1% level. The full IRT results can be found in the appendix (Table B1) along with a description of the methodology used.

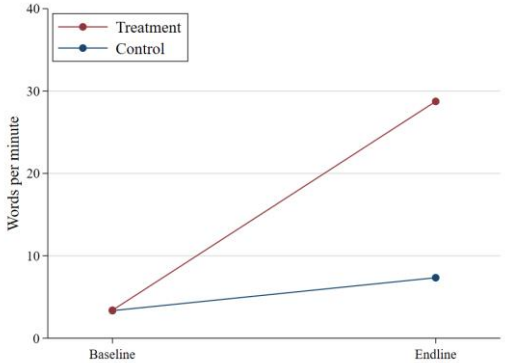
Figure 3: Standardized treatment effects for EGRA subtasks



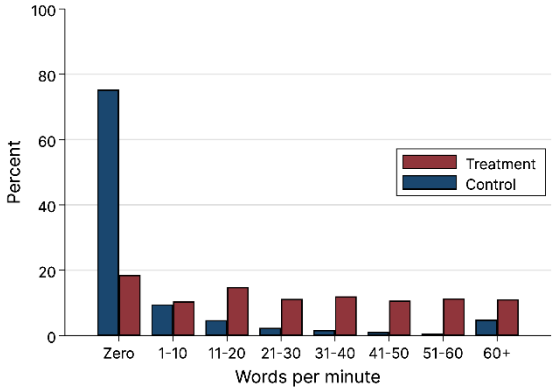
In Figure 4, we illustrate the program's effect on oral reading fluency (ORF) in the passage reading task. Although treatment and control OOSC start the year being able to read only a few words per minute (WPM) on average, by the end of the year treatment OOSC are able to read 29 WPM compared to 7 words per minute in the control group. Panel 4a shows that the ITT of 22 WPM and the TOT of 24 WPM are statistically significant. Panel 4b shows the fraction of OOSC at endline who were able to read varying numbers of WPM. While over 70% of control OOSC were not able to read any words at endline, more than 80% of treatment OOSC were able to read at least some words.

Figure 4: Oral reading fluency

4a: Treatment effects for ORF scores

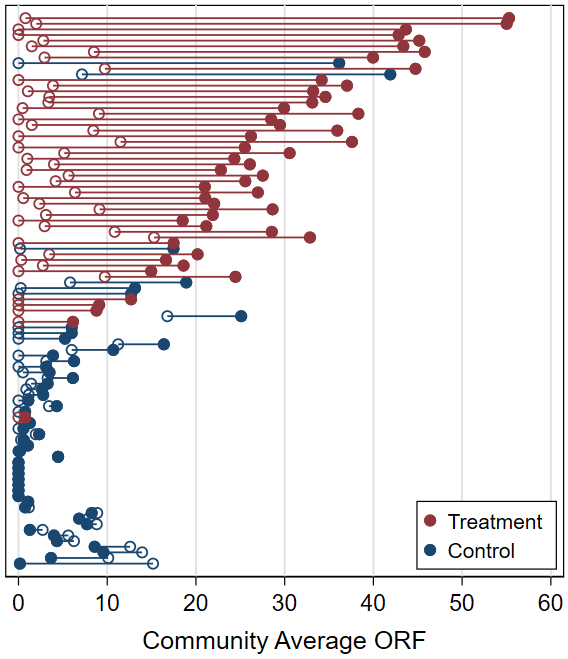


4b: Distribution of ORF scores at endline



As shown in the Appendix Table A3, treatment effects on reading subtasks are similar in size for girls vs boys, younger vs older children, children who were previously enrolled in school vs dropouts, and children who started with lower baseline learning levels vs higher baseline learning levels. In Figure 5 we show average oral reading fluency in each community at baseline and endline, color-coded by treatment arm. We find that learning gains are positive in all treatment communities, and these gains exceed learning gains for almost all control communities, indicating that the program is having positive impact in all communities where it is implemented. Results are similar for other reading subtasks.

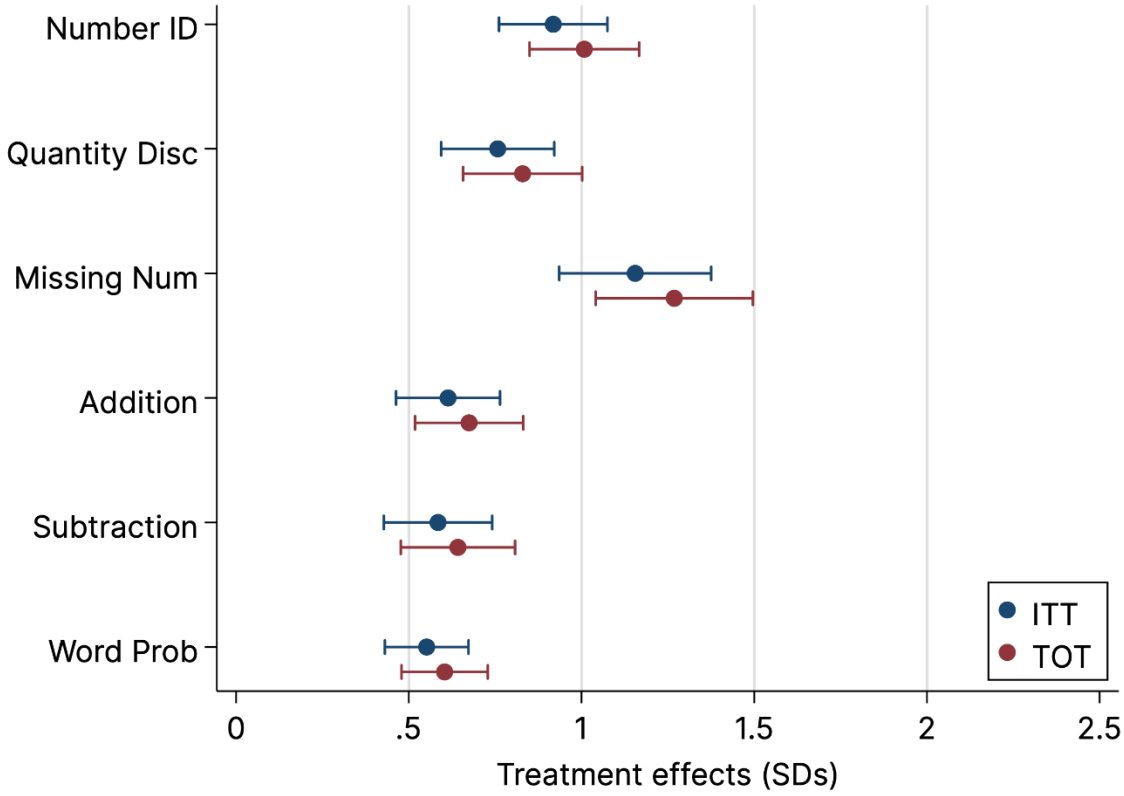
Figure 5: Oral reading fluency, change by community



Numeracy

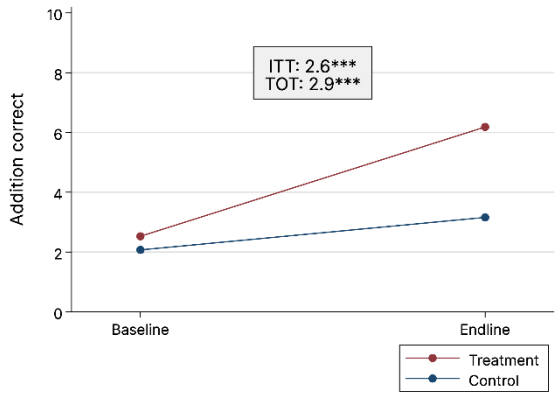
Figure 5 shows standardized effect sizes for each of the EGMA subtasks, along with 95% confidence intervals. While more modest than effects for the EGRA subtasks, treatment effects for EGMA subtasks are still large (> 0.5 SD) and statistically significant at the 1% level.

Figure 5: Standardized treatment effects for EGMA subtasks

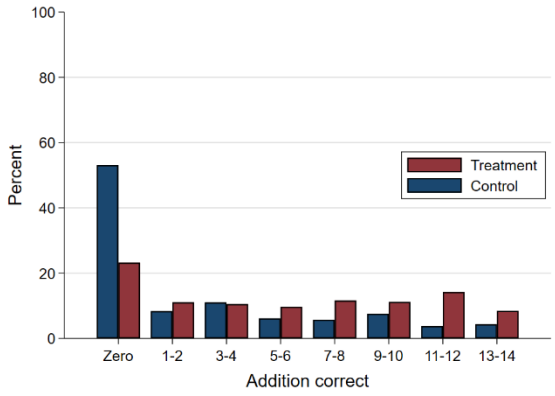


In Figure 6 we deep-dive into the addition and subtraction subtasks. Treatment and control OOSC start the year at similar levels, but by endline treatment OOSC are able to answer roughly twice as many addition and subtraction problems correctly than control OOSC. Panels 6b and 6d show that significantly fewer treatment OOSC received zero scores for these subtasks.

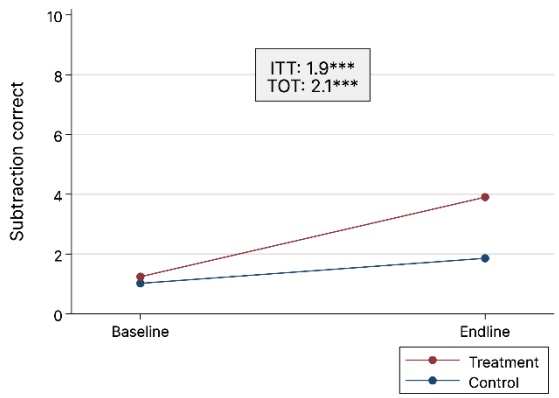
6a: Treatment effects for addition



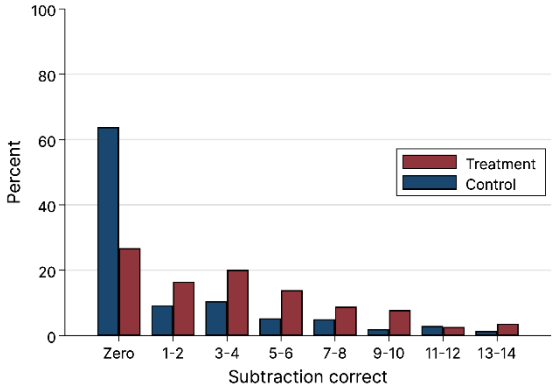
6b: Addition scores at endline



6c: Treatment effects for <u>subtraction</u>



6d: <u>Subtraction</u> scores at endline



When we account for varying subtask difficulty for EGMA using IRT our results remain positive and significant (see appendix B1). Similar to results for reading, we see positive numeracy effects across all subgroups and communities.

Socioemotional Learning

We also attempted to measure socioemotional learning (SEL) skills in a short battery of self-response questions. As this was a secondary outcome and a pilot tool, the sample size was substantially smaller than EGRA and EGMA. From discussions with Luminos, we identified key SEL skills to measure including empathy, open-mindedness, self-concept, and self-awareness. We measured these using a battery of questions drawn from the ISELA Self-Concept and Empathy modules, the Dweck Growth Mindset scale, part of the ISELA Learning Environment Safety module, and customized questions regarding attitude and behaviors towards education and learning.

Across almost all SEL indicators we do not find a statistically significant difference between treatment and control OOSC. However, we expect that these null results are due to the sensitivity of the research instrument rather than reflecting a true null effect of the program. At baseline, for the self-concept and empathy battery, over 90% of children were able to talk about something they hoped for, and correctly identify emotions from a picture and story. Between baseline and endline, however, we see at most a few percentage point changes in these indicators. These ceiling effects are preventing us from seeing how much things would have changed over the course of the program and if there is a difference between treatment and control children.

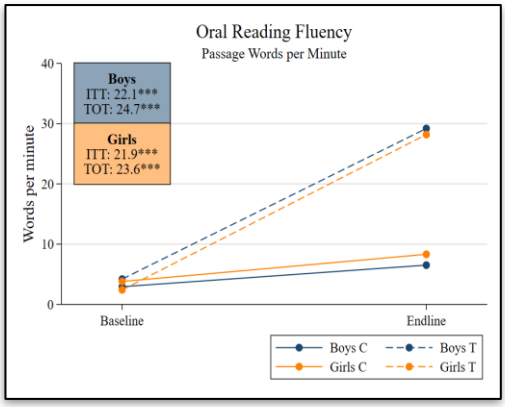
Lastly, for the questions concerning growth mindset, most children agreed with the statement that intelligence was unchangeable. In light of the results we see in other batteries, we believe that for many questions children may have been providing an agreeable response to whatever statement was made as opposed to understanding and responding to the statements being made. All results for the SEL questions are reported in the appendix (Table A1).

Outcomes for Girls

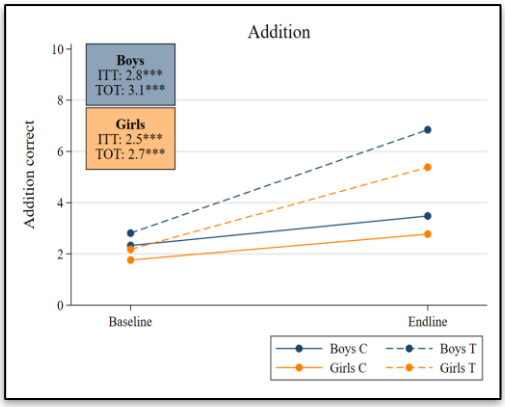
Overall, we find that the program has a similar impact on scores for boys and girls. The differences between treatment effects for boys and girls are not statistically significant for most EGRA and EGMA subtasks. We find that boys outperform girls on the non-word identification task with girls in the treatment group identifying 2.10 fewer non-words (p = 0.02) than boys in the treatment group. Girls also score 0.71 (p = 0.00) and 0.86 (p = 0.00) fewer points on the quantity discrimination and missing number task in EGMA, respectively. But as aforementioned, across all other subtasks differences in scores are not significant. Therefore our conclusion is that the Luminos program generally benefits both boys and girls equally, as hypothesized.

Figure 7 shows a breakdown of the ITT and TOT scores for each of the gender subgroups: boys and girls. Figure 7a illustrates that the Oral Reading Fluency subtask does not have significant differences between boys and girls, while 7b illustrates this for the Addition subtask.

7a: Oral Reading Fluency- Boys vs Girls



7b: Addition- Boys vs Girls



Benchmark: Government Schoolchildren

We collected data from children enrolled in nearby government schools in both treatment and control communities to provide a benchmark for the learning gains in the RCT. These schools were the "link school", meaning the government partner school where children who graduate from the Luminos program would get enrolled. Per community, we tried to sample one child for grade 1 and two children for both grades 2 and 3. After the Luminos program, children will usually enroll in this school, making it a useful benchmark for assessing treatment students' readiness for age-appropriate enrollment.

Generally, we find that GSCs start at higher reading and math levels than OOSCs, but treatment OOSCs have similar or better scores than GSCs at endline. This is particularly clear for reading,

where treatment OOSCs have higher endline scores than GSCs for all EGRA subtasks except for listening comprehension. Endline EGMA scores are similar for GSCs and treatment OOSCs (and much lower for control OOSCs).

Discussion

The study conducted in Liberia adopted a randomized controlled trial (RCT) research design to assess the impact of the Luminos program. The program's objective is to enable children to reach at least the reading and math level of the third year of the government school system, so they can transition to government schools equipped with the necessary foundational skills to integrate into the regular academic curriculum.

The obtained results reflect a significant and remarkable success that aligns with the intentions of the Luminos program. The impact assessment revealed a strong progression among children participating in the program, across all tasks of the reading and numeracy assessments. Out-of-school children recruited for the Luminos program started on average at substantially lower literacy and numeracy levels than the government schoolchildren used for benchmarking. Children in the Luminos program would end at a similar or higher level than the government schoolchildren at the end of the year, with especially clear differences on reading competencies. This indicates that the children learn more in the Luminos program than they would in a typical year in the link schools.

Though we did not gather data on *why* the Luminos program is so successful we believe that three main factors may be playing a role in the success of the program. First, the Luminos program consistently delivered classes 5-days per week for 10 months. From our observations, we believe Luminos program children likely received more hours of instruction over the evaluation period than GSC. Second, the Luminos program was designed to be student-centered and play-based so that it was more engaging. Lastly, anecdotal observations and interviews indicated that monitoring of both children and teachers (facilitators for Luminos) is much more intense in the Luminos program compared to government schools. This monitoring makes it likely that attendance of both students and teachers is higher, as also witnessed by guardians of alumni of the Luminos program – the children and their caretakers would know the Luminos facilitator will always show up, so going to school is never in vain. Luminos' facilitators also receive 25 days/year of intensive training, in addition to frequent classroom-based coaching. In addition, many caretakers mentioned that the Luminos program being free reduces the barrier to entry for new students, and consistently providing free school lunches provides a clear incentive to attend. Beyond this, caretakers would mention the positive attitude and approach of the Luminos program, and the parent support groups that helped them to prepare for enrollment of their children after the program.

The results did not show substantial heterogeneity at the subgroup level on county, gender or age, indicating that the program is effective for all types of students and in all communities where it is implemented (see Appendix A2 and A3). However, slight differences were noted in the EGRA tasks between girls and boys, as well as between children under 10 years old and above 10 years old. In the EGMA tasks, boys achieved slightly higher results on most subtasks with statistically significant differences compared to girls seen for the quantity discrimination

and missing number task, but for all other EGMA subtasks the difference in performance was not statistically significant.

Regarding the results related to socioemotional learning (SEL), the overall observation is that the results are not statistically significant, meaning they are not conclusive. As mentioned above, we believe that the SEL results related to self-concept, empathy, and growth mindset are likely insignificant due to measurement issues, rather than reflecting a true null effect of the program. We also do not see statistically significant differences between treatment and control communities in the likelihood of children reporting safe behavior in their communities, or significant changes over time in treatment children reporting safe behavior within their classrooms. However, we do observe that children in the treatment group were significantly less likely to report violence within their classrooms than within their communities as a whole. Finally, we see that children in treatment communities were significantly more likely to report practicing reading at home than children in control communities.

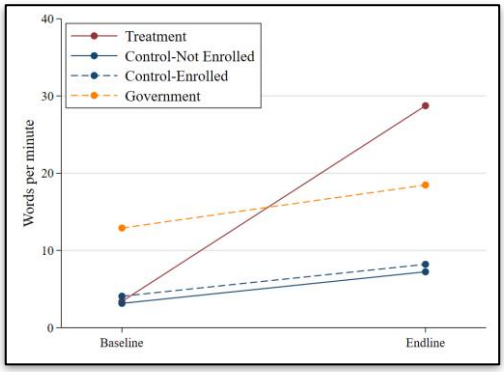
Interestingly, our results also revealed a modest increase in literacy and numeracy scores within the control group. It is not uncommon, within the context of an impact evaluation, to note an improvement in scores among individuals who did not benefit from the program, though it is difficult to identify the factors that led to this improvement. In our evaluation of the Luminos program, we do observe that one-third (32.39%) of children in the control group reported that they were enrolled in government school at endline. However, this does not appear to be driving the modest increase in scores in the control group: children in the control group who reported enrolling in government school had only marginally higher learning gains compared to control children who did not enroll in school (Figure 8).

The proportion of control group children who enrolled in government school is also indicative of one of the challenges Luminos faces: targeting children who are effectively out-of-school, and would not attend school in the absence of the program. During sampling, we often noticed that it is difficult to reliably observe the eligibility criteria that Luminos uses, as several factors (identity, age, prior schooling) are hard to observe and validate. Anecdotal evidence also showed that children or even entire school classes were initially recruited for the Luminos program, only to be filtered out during validation due to ineligibility. Most of all, this is indicative of the high expectations that communities have of the Luminos program once they get to know it: they expect children will improve substantially during the program, and are highly motivated to have their children enrolled in the program.

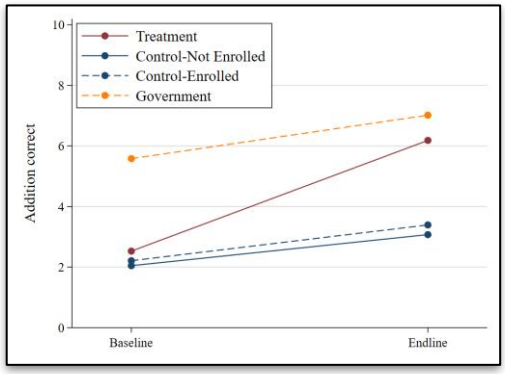
In conclusion, one of the most remarkable aspects of this study lies in the magnitude of improvement observed in the performance of children who participated in the program. The obtained results exhibit high levels of statistical significance, thus confirming the program's effectiveness in enhancing the reading and math skills of participating children. This improvement in performance not only prepares the children for their future educational success but also reflects the effectiveness of the pedagogical strategies implemented.

Figure 8: Treatment effects for OOSC in control enrolled in government school at endline, compared to the other (sub-)populations

8a: Oral Reading Fluency



8b: Addition



Limitations of study

Attrition: As mentioned above the primary hurdles encountered during the data collection process encompassed issues concerning phone reachability, geographical accessibility challenges inherent to Liberia's terrain, and the reservations of influential figures within specific communities. Notably, the absence of phone network coverage compounded with limited access to mobile phones among populations in remote regions emerged as a significant challenge. This predicament was further intensified by the dispersion of target households, situated considerable distances apart from each other. Moreover, data collection coincided with the onset of the rainy season, prompting families in remote locales to engage extensively in agricultural activities. In addition, certain control communities exhibited reluctance to collaborate with the data collection process. This reluctance was, in part, due to misunderstandings between guardians, community leaders, and enumerators regarding the treatment children would receive. On rarer occasions, religious authorities in specific communities denied enumerators entry or access to the children. Nevertheless, we successfully reassessed 86% of OOSC and 78% of GSC who were assessed at baseline, and attrition rates do not differ across treatment and control groups.

Although attrition rates were balanced between treatment and control children, as a robustness check we ran our ITT analysis with inverse-probability weights (IPW) and calculated Lee Bounds on treatment effects to test for the effect of attrition on our estimates. For the IPW analysis and Lee Bounds, treatment effects stay positive and significant at the 1% level across all subtasks. Estimates of treatment effects are very close to the estimate without adjusting for attrition. For example, our IPW estimates indicate treatment children could identify 17.11 more familiar words and our Lee Bounds estimates estimated a lower bound of the treatment effect of 17.29 more familiar words and an upper bound of the treatment effect of 18.10 familiar

words. These estimates are very close to the estimates without adjusting for attrition 18.15 more familiar words. The full set of results are included in appendix A7 and A8.

Eligibility verification and sampling: The process of randomization encountered challenges across both the baseline and endline phases. As mentioned previously, the objective was to sample 20 out-of-school children (OOSC) per community, factoring in gender and age stratification. However, the baseline sample of 1,756 OOSC fell short of the target of 2,000 due to issues regarding the enforcement of eligibility criteria and the limited availability of children. Over 50% of the reached OOSC couldn't be assessed, primarily due to their participation in existing traditional schools or relocations between scoping and baseline data collection. Additionally, 35% of initially reached children didn't meet program eligibility at baseline, but after discussions, 238 of these 'exception' children were included in the sample.

During endline, the challenges of randomization stemmed from the need to increase the number of children sampled at baseline. After baseline data collection was completed, certain communities did not have enough children sampled to reach the minimum threshold of 10 OOSC to include these communities in the analysis. Given the challenges experienced with network connectivity and reachability of children in remote areas at endline, we designated communities below 15 OOSC sampled to add additional children at endline, even if they were not sampled at baseline. Regarding the endline-only sample, we utilized similar stratification policies as at baseline; we stratified across gender and age, starting initially with follow-ups with those we reached at baseline and then following the order set at baseline to replace children who could not be found. The compounding of these challenges led to multiple rounds of mop-ups in control communities. Control communities were especially challenging given children had activities outside of their communities in lieu of schooling. Our final sample of children at the endline was 1,782 OOSC reached, with 1,502 OOSC who were surveyed at endline and baseline. The endline analysis accommodated these cases with a binary variable to address missing baseline scores. Our treatment effect estimates do not change much when we include children assessed only at endline along with children assessed at baseline and endline. Treatment effects remain positive and significant at the 1% level and treatment effects in standard deviation range continue to range from 0.50 SD to 1.75 SD. The full results are available in appendix A6.

Next steps

Our study focused on measuring the impact of the Luminos program, and did not identify specific aspects of the program that contribute to its success. The benchmark with government schoolchildren is also not a perfect comparison, as we can expect GSC to be structurally different from OOSC. Nevertheless, the fact that OOSC start lower at most tasks, and often overtake the benchmark GSC group indicates that the Luminos program is likely more effective in preparing OOSC for an age-appropriate grade compared to link schools. As such, it seems very likely that scaling up the Luminos program across Liberia will not just reduce the number

of OOSC across Liberia, but will also reduce the incidence of over-age children in the government school system.

In the 2022-23 academic year, Luminos nearly doubled the number of students receiving its program in Liberia. Despite this substantial growth, results across the board remain consistent with previous monitoring results from Luminos, and are consistent across treatment communities. Across communities, the children enrolled also initially showed very low numeracy and literacy scores, consistent with Luminos' target population. This shows the scalability of the Luminos approach. At the same time, the challenges at baseline sampling show that observing the eligibility criteria and effectively targeting OOSC can be challenging in the context of Liberia, which could be a challenge for Luminos' goal of focusing their efforts on the most marginalized groups.

Two big questions remain, that this impact evaluation cannot answer: do children indeed enroll and stay enrolled in the government school system after attending the Luminos program? And what makes the Luminos program more effective than "normal" government schools? For the second question, we have mentioned some observations, but these are not supported by statistical evidence. Designing a mixed-methods evaluation that collects targeted quantitative and qualitative evidence on what sets Luminos apart from government schools could be of great relevance to improving Luminos' effectiveness in the future. The first question could be answered by following up with the children sampled in this study at regular intervals in the future. We are planning to conduct another follow-up round of data collection next year to determine if treatment effects persist for children who participated in the Luminos program, and if Luminos program children enrolled and stayed enrolled in government school, to provide some evidence for the first lingering question.

Appendices

Results tables

Appendix A: Results tables

Table A1: Comprehensive Results Table

	Control (N = 725)	Control (N = 725)	Control (N = 725)	Treatment (N = 777)	Treatment (N = 777)	Treatment (N = 777)					GSC (N = 348)	GSC (N = 348)	GSC (N = 348)
	BL	EL	Change	BL	EL	Change	DiD	ITT	ITT (SDs)	TOT	BL	EL	Change
EGRA													
Letter name identification (100)	28.29	38.01	9.72	32.98	80.84	47.86	38.14	37.12	1.13	40.83	63.63	72.99	9.37
Letter names per minute	28.47	38.80	10.33	33.08	84.19	51.11	40.78	39.10	1.13	43.00	63.59	74.85	11.25
Phonemics (10)	2.94	3.46	0.53	3.03	5.13	2.10	1.58	1.67	0.74	1.83	4.03	4.60	0.57
Familiar word reading (50)	2.03	4.43	2.40	1.66	22.24	20.58	18.18	18.15	1.79	19.93	7.34	13.23	5.89
Familiar word per minute	1.90	4.94	3.04	1.64	23.12	21.49	18.45	18.56	1.56	20.38	7.19	14.25	7.05
Non-word reading (50)	0.25	0.85	0.60	0.05	8.19	8.14	7.54	7.90	1.68	8.68	0.89	2.59	1.70
Non-word reading per minute	0.09	0.92	0.83	0.05	8.03	7.99	7.15	7.67	1.91	8.43	0.38	2.62	2.24
Passage reading	3.32	6.05	2.73	3.36	26.62	23.26	20.53	21.28	1.45	23.37	12.87	16.38	3.51
Passage score per minute	3.34	7.34	4.00	3.40	28.74	25.34	21.34	22.04	1.12	24.21	12.90	18.47	5.57

	Control (N = 725)	Control (N = 725)	Control (N = 725)	Treatment (N = 777)	Treatment (N = 777)	Treatment (N = 777)					GSC (N = 348)	GSC (N = 348)	GSC (N = 348)
Reading comprehension (5)	0.18	0.39	0.21	0.15	1.63	1.48	1.27	1.31	1.33	1.44	0.73	1.24	0.51
Listening comprehension (3)	1.47	1.38	-0.10	1.55	1.97	0.42	0.52	0.57	0.54	0.63	1.90	1.97	0.07
EGMA													
Number identification (30)	5.94	8.83	2.89	6.42	17.69	11.27	8.38	8.25	0.92	9.06	14.44	18.65	4.21
Numbers per minute	6.64	9.53	2.98	6.98	18.97	12.21	9.23	8.90	0.83	9.76	15.39	20.40	5.15
Quantity discrimination (10)	2.38	3.56	1.18	2.85	6.38	3.53	2.35	2.47	0.76	2.71	5.05	6.20	1.16
Missing number (10)	0.28	0.68	0.40	0.27	2.49	2.22	1.83	1.84	1.16	2.03	0.97	1.79	0.82
Addition (15)	2.07	3.16	1.09	2.53	6.18	3.66	2.57	2.63	0.61	2.89	5.58	7.01	1.43
Addition per minute	2.11	3.26	1.15	2.58	6.35	3.77	2.63	2.67	0.59	2.93	5.60	7.26	1.66
Subtraction (15)	1.02	1.86	0.84	1.24	3.90	2.66	1.82	1.88	0.58	2.07	2.96	4.48	1.53
Subtraction per minute	1.01	1.92	0.91	1.27	3.97	2.71	1.80	1.84	0.55	2.02	2.90	4.55	1.64
Word problems (5)	1.21	1.54	0.31	1.25	2.35	1.10	0.78	0.75	0.55	0.82	1.99	2.58	0.59
	BL	**EL**	**Change**	**BL**	**EL**	**Change**	**DiD**	**ITT**	**ITT (SDs)**	**TOT**	**BL**	**EL**	**Change**
SEL													
ISELA Self Concept													
% who could describe future events	0.92	0.89	-0.03	0.93	0.93	-0.01	0.03	0.05	0.16	0.05	0.96	0.95	-0.02

	Control (N = 725)	Control (N = 725)	Control (N = 725)	Treatment (N = 777)	Treatment (N = 777)	Treatment (N = 777)					GSC (N = 348)	GSC (N = 348)	GSC (N = 348)
% who could include themselves in the future	0.87	0.87	0.00	0.90	0.91	0.00	0.00	0.04	0.13	0.05	0.94	0.94	-0.02
% who could identify an obstacle	0.74	0.70	0.02	0.75	0.64	-0.13	-0.15	-0.08	-0.16	-0.08	0.79	0.74	-0.06
% who could identify resources that could help them	0.95	0.93	-0.02	0.96	0.93	-0.02	-0.01	0.02	0.08	0.02	0.97	0.96	0.00
% who could mention an additional thing they hope for	0.81	0.78	-0.02	0.91	0.86	-0.03	-0.01	0.09	0.22	0.10	0.92	0.84	-0.10
% who could identify an obstacle to the additional thing they hope for	0.88	0.67	-0.12	0.76	0.70	-0.06	0.06	-0.01	-0.02	-0.01	0.86	0.75	-0.14
% who could identify resources that could help them with the additional thing they hope for	0.95	0.96	0.01	0.97	0.98	0.02	0.01	0.03	0.18	0.04	0.99	0.95	-0.05
ISELA Empathy													
% who correctly identified emotion	0.96	0.97	-0.02	0.96	0.99	0.03	0.05	0.02	0.13	0.02	0.99	0.99	0.01
% who could identify a way to help	0.88	0.94	0.02	0.89	0.97	0.09	0.06	0.05	0.19	0.05	0.89	0.96	0.07
% who could identify a second way to help	0.83	0.87	0.00	0.78	0.89	0.08	0.08	-0.01	-0.04	-0.02	0.81	0.90	0.11
% who could identify emotions in a story	0.83	0.82	-0.07	0.83	0.84	-0.01	0.06	0.03	0.07	0.03	0.87	0.93	0.07

	Control (N = 725)	Control (N = 725)	Control (N = 725)	Treatment (N = 777)	Treatment (N = 777)	Treatment (N = 777)					GSC (N = 348)	GSC (N = 348)	GSC (N = 348)
% who could empathize with a response to someone else's feelings	0.87	0.91	0.00	0.89	0.96	0.06	0.06	0.04	0.15	0.05	0.94	0.97	0.04
Do you ask for support in obstacles?	0.90	0.91	-0.01	0.89	0.90	0.00	0.02	0.02	0.06	0.02	0.94	0.92	-0.03
Can you name some good friends?	0.96	0.95	-0.03	0.96	0.99	0.02	0.05	0.03	0.16	0.04	0.98	0.99	0.00
DWECK Growth Mindset													
It always stays the same how smart you are, and you can't really do much to change it	4.14	3.94	-0.12	3.84	3.97	0.23	0.35	0.14	0.13	0.16	3.86	3.90	0.06
How smart you are is something about you that you can't change very much	4.09	3.76	-0.31	3.88	3.80	0.01	0.31	0.11	0.08	0.11	3.71	3.73	0.07
You can learn new things, but you can't really change how smart you are	4.12	3.87	-0.30	3.90	4.04	0.18	0.48	0.31	0.27	0.34	3.74	3.98	0.24
Safeguarding and Other													
In the last week, did you feel afraid in your community?	0.44	0.38	-0.05	0.31	0.35	0.03	0.09	-0.02	-0.04	-0.02	0.25	0.30	0.05
In the last week, did you feel afraid in your class?[8]				0.18	0.17	0.00		0.00			0.17	0.15	-0.01

[8] All learning environment safety questions relating to classrooms were only asked to treatment and GSC, under the assumption that control OOSC do not attend classes.

	Control (N = 725)	Control (N = 725)	Control (N = 725)	Treatment (N = 777)	Treatment (N = 777)	Treatment (N = 777)					GSC (N = 348)	GSC (N = 348)	GSC (N = 348)
In the last week, did children in your community get into a fight where someone was hit?	0.50	0.48	-0.01	0.52	0.54	-0.00	0.01	0.02	0.04	0.02	0.45	0.43	-0.01
In the last week, did children in your class get into a fight where someone was hit?				0.32	0.32	-0.01		0.00			0.21	0.26	0.02
In the last week, did an adult in your community scream or yell angrily at a child?	0.62	0.53	-0.14	0.55	0.60	0.05	0.18	0.09	0.18	0.10	0.54	0.43	-0.11
In the last week, did an adult in your class scream or yell angrily at a child?				0.27	0.24	0.02		0.00			0.21	0.28	0.07
In the last week, did an adult in your community hit or kick a child?	0.61	0.51	-0.11	0.49	0.48	0.00	0.11	0.02	0.04	0.02	0.45	0.40	-0.06
In the last week, did an adult in your class hit or kick a child?				0.14	0.15	0.01		0.00			0.29	0.17	-0.11
% who want to go to school next year	0.98	0.99	0.00	0.99	0.98	-0.01	-0.01	0.00	0.01	0.00	1.00	0.99	-0.01
% who practice reading at home	0.57	0.64	0.08	0.66	0.85	0.19	0.11	0.19	0.39	0.20	0.86	0.83	-0.03

Table A2: EGRA (Raw Scores)

Subtask	Control Average Difference	ITT	ITT (SDs)	ITT SE	p-value	N	TOT	TOT SE	p-value	N
Letter name identification (100)	9.72	37.12	1.13	2.28	0.00	1502	40.83	2.46	0.00	1502
Phonemics (10)	0.53	1.67	0.74	0.18	0.00	1502	1.83	0.19	0.00	1502
Familiar word reading (50)	2.40	18.15	1.79	1.24	0.00	1502	19.93	1.25	0.00	1502
Non-word reading (50)	0.60	7.90	1.68	0.77	0.00	1502	8.68	0.80	0.00	1502
Passage reading (60)	2.73	21.28	1.45	1.88	0.00	1502	23.37	1.97	0.00	1502
Reading comprehension (5)	0.21	1.31	1.33	0.12	0.00	1481	1.44	0.13	0.00	1481
Listening comprehension (3)	-0.10	0.57	0.54	0.09	0.00	1502	0.63	0.10	0.00	1502

Table A3: EGRA Subgroup Analysis (Raw scores)

Subtask	Control Average Difference	ITT	ITT (SDs)	ITT SE	p-value	N	TOT	TOT SE	p-value	N
Letter name identification (100)	9.72	37.12	1.13	2.28	0.00	1502	40.83	2.46	0.00	1502
Male	8.24	37.62	1.14	2.51	0.00	820	42.17	2.73	0.00	820
Female	11.47	36.64	1.11	2.82	0.00	682	39.39	2.79	0.00	682
Older (10 and above)	9.08	33.27	1.01	2.42	0.00	893	35.13	2.54	0.00	893
Younger (Below 10)	10.56	40.87	1.24	2.89	0.00	609	48.42	3.39	0.00	609
LIPACE	10.84	33.65	1.02	2.76	0.00	846	37.55	2.50	0.00	846
ROCH	8.32	40.93	1.24	3.56	0.00	656	44.11	4.25	0.00	656
Previous School (Yes)	8.55	27.75	0.84	2.53	0.00	603	30.94	2.49	0.00	603
Previous School (No)	10.80	43.87	1.33	3.00	0.00	899	47.78	3.58	0.00	899
Bomi	8.16	35.73	1.08	5.11	0.00	194	37.39	4.84	0.00	194
Bong	20.50	28.32	0.86	3.77	0.00	147	29.01	4.01	0.00	147
Grand Cape Mount	5.30	45.63	1.38	2.98	0.00	691	49.09	3.58	0.00	691
Margibi	15.61	29.62	0.90	5.19	0.00	319	37.84	5.05	0.00	319
Montserrado	10.98	33.42	1.01	4.95	0.00	151	34.16	4.57	0.00	151

Subtask	Control Average Difference	ITT	ITT (SDs)	ITT SE	p-value	N	TOT	TOT SE	p-value	N
Phonemics (10)	**0.53**	**1.67**	**0.74**	**0.18**	**0.00**	**1502**	**1.83**	**0.19**	**0.00**	**1502**
Male	0.54	1.64	0.73	0.21	0.00	820	1.83	0.22	0.00	820
Female	0.52	1.68	0.75	0.22	0.00	682	1.80	0.22	0.00	682
Older (10 and above)	0.37	1.78	0.80	0.21	0.00	893	1.88	0.21	0.00	893
Younger (Below 10)	0.74	1.39	0.62	0.25	0.00	609	1.63	0.27	0.00	609
LIPACE	0.68	1.53	0.69	0.24	0.00	846	1.71	0.24	0.00	846
ROCH	0.34	1.83	0.82	0.26	0.00	656	1.97	0.28	0.00	656
Previous School (Yes)	0.58	1.37	0.61	0.27	0.00	603	1.53	0.27	0.00	603
Previous School (No)	0.48	1.89	0.84	0.20	0.00	899	2.05	0.22	0.00	899
Bomi	-0.24	1.92	0.86	0.57	0.01	194	2.02	0.55	0.00	194
Bong	0.41	1.18	0.53	0.34	0.00	147	1.21	0.33	0.00	147
Grand Cape Mount	0.35	2.22	0.99	0.22	0.00	691	2.37	0.23	0.00	691
Margibi	1.39	0.68	0.30	0.35	0.07	319	0.86	0.41	0.03	319
Montserrado	0.85	1.66	0.74	0.68	0.04	151	1.70	0.64	0.01	151
Familiar word reading (50)	**2.40**	**18.15**	**1.79**	**1.24**	**0.00**	**1502**	**19.93**	**1.25**	**0.00**	**1502**
Male	2.14	18.05	1.78	1.22	0.00	820	20.17	1.18	0.00	820

Female	2.71	18.20	1.80	1.44	0.00	682	19.59	1.48	0.00	682
Subtask	**Control Average Difference**	**ITT**	**ITT (SDs)**	**ITT SE**	**p-value**	**N**	**TOT**	**TOT SE**	**p-value**	**N**
Younger (Below 10)	2.00	17.64	1.74	1.62	0.00	609	20.80	1.78	0.00	609
Older (10 and above)	2.70	18.22	1.80	1.19	0.00	893	19.23	1.17	0.00	893
LIPACE	2.63	20.42	2.02	1.76	0.00	846	22.84	1.55	0.00	846
ROCH	2.11	15.25	1.51	1.64	0.00	656	16.39	1.78	0.00	656
Previous School (Yes)	2.93	18.56	1.83	1.71	0.00	603	20.73	1.53	0.00	603
Previous School (No)	1.91	17.90	1.77	1.48	0.00	899	19.42	1.60	0.00	899
Bomi	0.90	24.95	2.46	2.94	0.00	194	26.15	2.63	0.00	194
Bong	4.11	9.14	0.90	2.05	0.00	147	9.36	2.14	0.00	147
Grand Cape Mount	1.22	20.93	2.07	1.34	0.00	691	22.33	1.22	0.00	691
Margibi	4.86	13.66	1.35	2.74	0.00	319	17.52	3.01	0.00	319
Montserrado	3.59	18.46	1.82	4.12	0.00	151	18.86	3.86	0.00	151
Non-word reading (50)	0.60	7.90	1.68	0.77	0.00	1502	8.68	0.80	0.00	1502
Male	0.75	8.81	1.88	0.94	0.00	820	9.85	1.00	0.00	820
Female	0.42	6.79	1.45	0.83	0.00	682	7.31	0.86	0.00	682
Older (10 and above)	0.42	8.09	1.72	0.79	0.00	893	8.53	0.79	0.00	893

Younger (Below 10)	0.84	7.45	1.59	0.97	0.00	609	8.78	1.08	0.00	609
LIPACE	0.68	8.31	1.77	1.05	0.00	846	9.29	1.02	0.00	846
Subtask	**Control Average Difference**	**ITT**	**ITT (SDs)**	**ITT SE**	**p-value**	**N**	**TOT**	**TOT SE**	**p-value**	**N**
ROCH	0.50	7.44	1.59	1.18	0.00	656	8.00	1.26	0.00	656
Previous School (Yes)	0.96	8.64	1.84	1.12	0.00	603	9.64	1.08	0.00	603
Previous School (No)	0.26	7.59	1.62	0.87	0.00	899	8.24	0.95	0.00	899
Bomi	0.08	11.71	2.50	1.19	0.00	194	12.31	1.52	0.00	194
Bong	0.99	2.06	0.44	0.59	0.00	147	2.10	0.58	0.00	147
Grand Cape Mount	0.36	10.10	2.15	0.98	0.00	691	10.78	0.99	0.00	691
Margibi	1.22	3.59	0.77	1.59	0.04	319	4.60	1.84	0.01	319
Montserrado	0.90	9.93	2.12	2.33	0.00	151	10.15	2.13	0.00	151
Passage reading (60)	2.73	21.28	1.45	1.88	0.00	1502	23.37	1.97	0.00	1502
Male	2.47	21.04	1.43	1.85	0.00	820	23.51	1.95	0.00	820
Female	3.04	21.66	1.47	2.26	0.00	682	23.32	2.33	0.00	682
Older (10 and above)	2.53	21.56	1.47	1.89	0.00	893	22.74	1.89	0.00	893
Younger (Below 10)	3.00	19.85	1.35	2.29	0.00	609	23.42	2.60	0.00	609
LIPACE	4.05	24.11	1.64	2.85	0.00	846	26.93	2.81	0.00	846

ROCH	1.08	17.59	1.20	2.25	0.00	656	18.92	2.49	0.00	656
Previous School (Yes)	2.68	23.23	1.58	2.25	0.00	603	25.91	2.18	0.00	603
Previous School (No)	2.78	20.08	1.37	2.35	0.00	899	21.79	2.58	0.00	899
Subtask	**Control Average Difference**	**ITT**	**ITT (SDs)**	**ITT SE**	**p-value**	**N**	**TOT**	**TOT SE**	**p-value**	**N**
Bomi	0.70	33.15	2.25	4.06	0.00	194	34.76	3.69	0.00	194
Bong	3.61	8.99	0.61	2.41	0.00	147	9.20	2.48	0.00	147
Grand Cape Mount	0.60	24.85	1.69	1.96	0.00	691	26.53	1.97	0.00	691
Margibi	9.77	12.71	0.86	4.63	0.01	319	16.29	5.72	0.00	319
Montserrado	0.22	24.28	1.65	4.88	0.00	151	24.82	4.56	0.00	151
Reading comprehension (5)	0.21	1.31	1.33	0.12	0.00	1481	1.44	0.13	0.00	1481
Male	0.19	1.36	1.38	0.12	0.00	807	1.52	0.13	0.00	807
Female	0.23	1.27	1.30	0.15	0.00	674	1.37	0.16	0.00	674
Older (10 and above)	0.23	1.33	1.36	0.11	0.00	880	1.41	0.12	0.00	880
Younger (Below 10)	0.17	1.22	1.25	0.16	0.00	601	1.45	0.19	0.00	601
LIPACE	0.24	1.49	1.52	0.15	0.00	826	1.67	0.15	0.00	826
ROCH	0.16	1.08	1.10	0.17	0.00	655	1.17	0.19	0.00	655
Previous School (Yes)	0.20	1.47	1.50	0.13	0.00	594	1.64	0.14	0.00	594

Previous School (No)	0.21	1.21	1.23	0.15	0.00	887	1.32	0.17	0.00	887
Bomi	0.17	1.71	1.74	0.21	0.00	194	1.78	0.20	0.00	194
Bong	0.41	0.38	0.39	0.22	0.10	146	0.39	0.22	0.07	146
Grand Cape Mount	0.06	1.59	1.61	0.14	0.00	675	1.69	0.15	0.00	675
Subtask	**Control Average Difference**	**ITT**	**ITT (SDs)**	**ITT SE**	**p-value**	**N**	**TOT**	**TOT SE**	**p-value**	**N**
Margibi	0.43	0.90	0.92	0.25	0.00	315	1.16	0.32	0.00	315
Montserrado	0.28	1.50	1.53	0.27	0.00	151	1.54	0.24	0.00	151
Listening comprehension (3)	-0.10	0.57	0.54	0.09	0.00	1502	0.63	0.10	0.00	1502
Male	-0.07	0.50	0.47	0.12	0.00	820	0.56	0.13	0.00	820
Female	-0.13	0.64	0.60	0.09	0.00	682	0.69	0.09	0.00	682
Older (10 and above)	-0.07	0.52	0.49	0.10	0.00	893	0.55	0.10	0.00	893
Younger (Below 10)	-0.14	0.58	0.55	0.12	0.00	609	0.69	0.14	0.00	609
LIPACE	-0.20	0.61	0.57	0.12	0.00	846	0.68	0.12	0.00	846
ROCH	0.03	0.55	0.51	0.14	0.00	656	0.59	0.15	0.00	656
Previous School (Yes)	-0.12	0.42	0.40	0.11	0.00	603	0.47	0.12	0.00	603
Previous School (No)	-0.08	0.66	0.63	0.11	0.00	899	0.72	0.12	0.00	899
Bomi	-1.10	0.75	0.71	0.22	0.01	194	0.79	0.21	0.00	194

Bong	0.35	0.04	0.04	0.27	0.89	147	0.04	0.26	0.88	147
Grand Cape Mount	0.02	0.78	0.73	0.12	0.00	691	0.83	0.13	0.00	691
Margibi	0.12	0.35	0.33	0.18	0.07	319	0.45	0.23	0.05	319
Montserrado	-0.03	0.49	0.46	0.22	0.05	151	0.50	0.20	0.01	151

Table A4: EGMA (Raw Scores)

Subtask	Control Average Difference	ITT	ITT (SDs)	ITT SE	p-value	N	TOT	TOT SE	p-value	N
Number identification (30)	2.89	8.25	0.92	0.71	0.00	1502	9.06	0.73	0.00	1502
Quantity discrimination (10)	1.18	2.47	0.76	0.27	0.00	1502	2.71	0.29	0.00	1502
Missing number (10)	0.40	1.84	1.16	0.18	0.00	1502	2.03	0.19	0.00	1502
Addition (15)	1.09	2.63	0.61	0.32	0.00	1502	2.89	0.34	0.00	1502
Subtraction (15)	0.84	1.88	0.58	0.25	0.00	1502	2.07	0.27	0.00	1502
Word problems (5)[9]	0.31	0.75	0.55	0.08	0.00	1487	0.82	0.09	0.00	1487

[9] Due to an error in the skip logic during baseline data collection, the Word Problems subtask in EGMA recorded only 4 out of 5 responses. This error was corrected at endline and we were able to record all 5 responses. While there is a discrepancy in data, our treatment effects were calculated using ANCOVA, where we include a control for the baseline score. We are still able to measure treatment effects because even though there is a difference in the number of questions asked, the baseline score should still be predictive of a child's word score at endline.

Table A5: EGMA Subgroup Analysis (Raw scores)

Subtask	Control Average Difference	ITT	ITT (SDs)	ITT SE	p-value	N	TOT	TOT SE	p-value	N
Numbers identification (30)	2.89	8.25	0.92	0.71	0.00	1502	9.06	0.73	0.00	1502
Male	3.13	8.32	0.93	0.79	0.00	820	9.31	0.81	0.00	820
Female	2.60	8.22	0.92	0.79	0.00	682	8.84	0.81	0.00	682
Older (10 and above)	2.79	7.27	0.81	0.71	0.00	893	7.66	0.72	0.00	893
Younger (Below 10)	3.02	8.91	0.99	0.95	0.00	609	10.53	1.04	0.00	609
LIPACE	3.18	8.62	0.96	0.95	0.00	846	9.61	0.87	0.00	846
ROCH	2.53	7.73	0.86	1.08	0.00	656	8.32	1.17	0.00	656
Previous School (Yes)	2.77	7.14	0.79	0.87	0.00	603	7.96	0.83	0.00	603
Previous School (No)	3.00	9.09	1.01	0.87	0.00	899	9.87	0.95	0.00	899
Quantity discrimination (10)	1.18	2.47	0.76	0.27	0.00	1502	2.71	0.29	0.00	1502
Male	0.97	2.74	0.84	0.30	0.00	820	3.05	0.30	0.00	820
Female	1.44	2.14	0.66	0.36	0.00	682	2.30	0.39	0.00	682
Older (10 and above)	0.97	2.29	0.70	0.26	0.00	893	2.40	0.27	0.00	893
Younger (Below 10)	1.47	2.65	0.81	0.40	0.00	609	3.13	0.45	0.00	609
LIPACE	1.21	2.58	0.79	0.38	0.00	846	2.87	0.39	0.00	846

ROCH	1.15	2.34	0.72	0.39	0.00	656	2.52	0.42	0.00	656
Subtask	**Control Average Difference**	**ITT**	**ITT (SDs)**	**ITT SE**	**p-value**	**N**	**TOT**	**TOT SE**	**p-value**	**N**
Previous School (Yes)	1.22	2.11	0.65	0.35	0.00	603	2.33	0.36	0.00	603
Previous School (No)	1.14	2.77	0.85	0.33	0.00	899	3.02	0.36	0.00	899
Missing number (10)	0.40	1.84	1.16	0.18	0.00	1502	2.03	0.19	0.00	1502
Male	0.37	2.18	1.37	0.23	0.00	820	2.43	0.23	0.00	820
Female	0.43	1.41	0.88	0.22	0.00	682	1.52	0.23	0.00	682
Older (10 and above)	0.42	2.07	1.29	0.18	0.00	893	2.18	0.18	0.00	893
Younger (Below 10)	0.37	1.46	0.91	0.25	0.00	609	1.72	0.29	0.00	609
LIPACE	0.46	1.89	1.19	0.24	0.00	846	2.11	0.24	0.00	846
ROCH	0.32	1.79	1.12	0.27	0.00	656	1.92	0.29	0.00	656
Previous School (Yes)	0.37	2.15	1.35	0.26	0.00	603	2.40	0.27	0.00	603
Previous School (No)	0.43	1.69	1.06	0.21	0.00	899	1.83	0.23	0.00	899
Addition (15)	1.09	2.63	0.61	0.32	0.00	1502	2.89	0.34	0.00	1502
Male	1.15	2.78	0.65	0.38	0.00	820	3.11	0.40	0.00	820
Female	1.01	2.49	0.58	0.41	0.00	682	2.68	0.42	0.00	682
Older (10 and above)	1.19	2.47	0.58	0.31	0.00	893	2.60	0.32	0.00	893

Subtask	Control Average Difference	ITT	ITT (SDs)	ITT SE	p-value	N	TOT	TOT SE	p-value	N
Younger (Below 10)	0.96	2.65	0.62	0.48	0.00	609	3.13	0.56	0.00	609
LIPACE	1.17	2.63	0.61	0.38	0.00	846	2.94	0.38	0.00	846
ROCH	0.98	2.62	0.61	0.55	0.00	656	2.82	0.60	0.00	656
Previous School (Yes)	1.17	2.23	0.52	0.41	0.00	603	2.48	0.43	0.00	603
Previous School (No)	1.02	2.85	0.66	0.40	0.00	899	3.10	0.43	0.00	899
Subtraction (15)	0.84	1.88	0.58	0.25	0.00	1502	2.07	0.27	0.00	1502
Male	0.92	1.95	0.61	0.28	0.00	820	2.18	0.31	0.00	820
Female	0.74	1.75	0.55	0.33	0.00	682	1.89	0.34	0.00	682
Older (10 and above)	1.01	1.90	0.59	0.26	0.00	893	2.00	0.27	0.00	893
Younger (Below 10)	0.61	1.81	0.56	0.35	0.00	609	2.14	0.40	0.00	609
LIPACE	0.88	1.82	0.56	0.31	0.00	846	2.03	0.31	0.00	846
ROCH	0.78	1.97	0.61	0.43	0.00	656	2.12	0.47	0.00	656
Previous School (Yes)	1.01	1.78	0.55	0.36	0.00	603	1.98	0.38	0.00	603
Previous School (No)	0.68	1.98	0.62	0.30	0.00	899	2.15	0.33	0.00	899
Word problems (5)	0.31	0.75	0.55	0.08	0.00	1487	0.82	0.09	0.00	1487
Male	0.30	0.77	0.56	0.11	0.00	814	0.86	0.11	0.00	814

Subtask	Control Average Difference	ITT	ITT (SDs)	ITT SE	p-value	N	TOT	TOT SE	p-value	N
Female	0.33	0.73	0.53	0.09	0.00	673	0.78	0.10	0.00	673
Older (10 and above)	0.39	0.69	0.51	0.11	0.00	886	0.73	0.11	0.00	886
Younger (Below 10)	0.22	0.82	0.60	0.13	0.00	601	0.95	0.15	0.00	601
LIPACE	0.19	0.75	0.55	0.10	0.00	840	0.84	0.11	0.00	840
ROCH	0.47	0.74	0.55	0.14	0.00	647	0.80	0.14	0.00	647
Previous School (Yes)	0.32	0.60	0.44	0.10	0.00	599	0.67	0.11	0.00	599
Previous School (No)	0.31	0.89	0.65	0.10	0.00	888	0.96	0.11	0.00	888

Table A6: Regression for Returning Children and Children Only Surveyed at EL

Subtask	Control Mean	ITT	ITT (SDs)	ITT (SE)	ITT (P-Value)	N	TOT	TOT (SE)	TOT (P-Value)	N
Letter name identification	9.72	35.20	1.07	2.24	0.00	1782.00	42.18	2.73	0.00	1782.00
Letter names per minute	10.33	37.12	1.07	2.45	0.00	1782.00	44.47	2.95	0.00	1782.00
Phonemics	0.53	1.55	0.69	0.17	0.00	1782.00	1.85	0.20	0.00	1782.00
Familiar word reading	2.40	17.11	1.69	1.22	0.00	1782.00	20.43	1.38	0.00	1782.00
Familiar word per minute	3.04	17.57	1.48	1.28	0.00	1782.00	20.98	1.47	0.00	1782.00
Non-word reading	0.60	7.28	1.55	0.73	0.00	1782.00	8.68	0.81	0.00	1782.00
Non-word reading per minute	0.83	7.02	1.75	0.70	0.00	1782.00	8.38	0.78	0.00	1782.00
Passage reading	2.73	20.01	1.36	1.79	0.00	1782.00	23.89	2.13	0.00	1782.00
Passage score per minute	4.00	20.84	1.06	2.18	0.00	1781.00	24.88	2.61	0.00	1781.00
Reading comprehension	0.21	1.22	1.24	0.11	0.00	1754.00	1.46	0.14	0.00	1754.00
Listening comprehension	-0.10	0.51	0.48	0.09	0.00	1782.00	0.61	0.10	0.00	1782.00
Number identification	2.89	7.54	0.84	0.70	0.00	1782.00	9.01	0.80	0.00	1782.00
Numbers per minute	2.98	7.98	0.74	0.85	0.00	1594.00	9.59	1.01	0.00	1594.00
Quantity discrimination	1.18	2.30	0.70	0.24	0.00	1782.00	2.75	0.29	0.00	1782.00
Missing number	0.40	1.70	1.06	0.17	0.00	1782.00	2.03	0.19	0.00	1782.00
Addition	1.09	2.38	0.55	0.30	0.00	1782.00	2.84	0.36	0.00	1782.00
Addition per minute	1.15	2.42	0.53	0.32	0.00	1782.00	2.89	0.38	0.00	1782.00
Subtraction	0.84	1.72	0.54	0.24	0.00	1782.00	2.06	0.29	0.00	1782.00
Subtraction per minute	0.91	1.70	0.51	0.25	0.00	1782.00	2.03	0.30	0.00	1782.00
Word problems	0.31	0.68	0.50	0.08	0.00	1767.00	0.80	0.09	0.00	1767.00

Table A7: Lee Bounds

Note*: All the results in the following table show are statistically significant (p<0.05)

Subtask	Lower Bound	Upper Bound
EGRA		
Letter name identification	40.02	41.72
Letter names per minute	42.52	44.67
Phonemics	1.56	1.71
Familiar word reading	17.29	18.10
Familiar word per minute	17.75	18.79
Non-word reading	7.11	7.64
Non-word reading per minute	6.85	7.28
Passage reading	19.85	20.87
Passage score per minute	20.83	22.32
Reading comprehension	1.19	1.34
Listening comprehension	0.55	0.60
EGMA		
Number identification	8.17	8.68
Numbers per minute	7.97	10.55
Quantity discrimination	2.58	2.75
Missing number	1.75	1.88
Addition	2.77	3.02
Addition per minute	2.83	3.13
Subtraction	1.94	2.17
Subtraction per minute	1.95	2.19
Word problems	0.75	0.79

Table A8: IPW

Note*: All the results in the following table show are statistically significant ($p<0.05$)

Subtask	ITT (without IPWs)	ITT (with IPWs)
EGRA		
Letter name identification	37.12	39.22
Letter names per minute	39.10	41.59
Phonemics	1.67	1.75
Familiar word reading	18.15	18.75
Familiar word per minute	18.56	19.12
Non-word reading	7.90	7.94
Non-word reading per minute	7.67	7.64
Passage reading	21.28	21.81
Passage score per minute	22.04	22.74
Reading comprehension	1.31	1.35
Listening comprehension	0.57	0.62
EGMA		
Number identification	8.25	8.68
Numbers per minute	8.90	9.47
Quantity discrimination	2.47	2.74
Missing number	1.84	1.88
Addition	2.63	2.76
Addition per minute	2.67	2.79
Subtraction	1.88	1.98
Subtraction per minute	1.84	1.94
Word problems	0.75	0.79

Table A9: Treatment Children Average Score by Attendance (>50%=High Attendance)

	Low Attendance (<=50% of classes)		High Attendance (>50% of classes)	
Subtask	**Baseline**	**Endline**	**Baseline**	**Endline**
EGRA				
Letter name identification	25.25	56.19	31.09	60.86
Letter names per minute	25.62	57.40	31.53	63.36
Phonemics	3.19	3.38	2.97	4.40
Familiar word reading	2.28	9.99	1.81	14.05
Familiar word per minute	2.49	11.60	1.97	15.03
Non-word reading	0.04	3.74	0.15	4.78
Non-word reading per minute	0.07	4.34	0.22	5.42
Passage reading	3.39	12.41	3.34	17.17
Passage score per minute	3.73	14.62	3.67	19.51
Reading comprehension	0.24	0.95	0.16	1.05
Listening comprehension	1.36	1.71	1.53	1.69
EGMA				
Number identification	5.32	10.44	6.24	13.68
Numbers per minute	6.03	13.90	6.98	14.66
Quantity discrimination	2.43	4.08	2.64	5.10
Missing number	0.25	1.09	0.28	1.67
Addition	2.01	3.59	2.33	4.82
Addition per minute	2.14	8.67	2.40	5.40
Subtraction	1.10	2.15	1.14	2.98
Subtraction per minute	1.16	2.59	1.18	3.32
Word problems	1.01	1.82	1.24	1.98

Table A10: SEL Instrument

Social-Emotional Learning			
Subtask		Indicator	Description of assessment/Question
Dweck Growth Mindset	1	Growth mindset, first question	Participant responds on a Likert scale to statement – "It always stays the same how smart you are, and you can't really do much to change it."
	2	Growth mindset, second question	Participant responds on a Likert scale to the statement – "How smart you are is something about you that you can't change very much."
	3	Differences between learning new things and intelligence	Participant responds on Likert scale to the statement – "You can learn new things, but you can't really change if you are smart or not smart."
Self-Concept Module	1	% who could describe future events	"I want you to think about something you hope will happen in your life in the future. Can you tell me what you would like to happen?"
	2	% who could include themselves in the future	The enumerator interprets if the participant described themselves from the response to the question above
	3	% who could identify an obstacle	"Can you tell me one thing that could happen that would stop you from doing this?"
	4	% who could identify resources that could help them	"Can you tell me who or what will help support you in this?"
	5	% who could mention an additional thing they hope for	"Can you tell me one other thing that you hope will happen in your life in the future?"
	6	% who could identify an obstacle to the additional thing they hope for	"Can you tell me one thing that could happen that would stop you from doing this?"
	7	% who could identify resources that could help them with the additional thing they hope for	"Can you tell me who or what will help support you in this?"
Empathy Module	1	% who correctly identified emotion	Participant is shown image of a child crying, and asked "How do you think this child is feeling right now?"
	2	% who could identify a way to help	"What would you do to make this child feel better?"
	3	% who could identify a second way to help	"Is there anything else you could do to help this child?"
	4	% who could identify emotions in a story	Participant is read a story explaining why the child is crying, and asked to interpret the story.

	5	% who could emphasize with a response to someone else's feelings	"How do you think the other child felt after the girl started crying?"
Other Socio-Emotional learning	1	% who practice reading at home	"Do you practice reading at home?"
	2	Social awareness	Participant is asked to name their good friends.
	3	% who want to go to school next year	"Think about next year: Do you want to go to school next year?"
	4	ISELA Learning environment safety: community safety – perceptions	"In the last week, did you feel afraid in your community?"
	5	ISELA Learning environment safety: classroom safety- perceptions	"In the last week, did you feel afraid in your class?"
	6	ISELA Learning environment safety: community safety – peer safety	"In the last week, did children in your community get into a fight where someone was hit?"
	7	ISELA Learning environment safety: classroom safety – peer safety	"In the last week, did children in your class get into a fight where someone was hit?"
	8	ISELA Learning environment safety: community safety – verbal violence	"In the last week, did an adult in your community scream or yell angrily at a child?"
	9	ISELA Learning environment safety: classroom safety – verbal violence	"In the last week, did an adult in your class scream or yell angrily at a child?"
	10	ISELA Learning environment safety: community safety – physical violence	"In the last week, did an adult in your community hit or kick a child?"
	11	ISELA Learning environment safety: classroom safety – physical violence (only treatment & GSC)	"In the last week, did an adult in your class hit or kick a child?"

Appendix B: Analytical modelling

Appendix B1: Item Response Theory Process and Results

Item response theory (IRT) is a statistical method used to estimate underlying or 'latent' ability and traits that cannot be directly observed. In education research IRT has been used to estimate attributes such as literacy skills and numeracy skills. The main underlying assumption in IRT is that each question in a well-designed test gives us information about a participant's latent trait or ability (Columbia University, 2023). In our context, when we use IRT we assume that a child's performance on the EGRA and EGMA subtasks can be explained by a measure of their latent literacy and numeracy ability, respectively. Before presenting the various steps we took to estimate the Luminos program's effect on numeracy and literacy, we highlight that IRT models typically require test questions to be dichotomous (i.e., 0/1) or ordinal (i.e., Likert scales). In order to use IRT with our data we have transformed all the scores on EGRA and EGMA subtasks to dichotomous variables by defining success (i.e., 1) as when a child scored more than 50% of the possible points on a subtask, and a child failed (i.e., 0) otherwise. For example, for the letter identification task, which has a maximum possible score of 100, if a child was able to identify between 51 – 100 letters they succeeded (i.e., assigned 1) at the task and if they identified 0 – 50 they failed the task (i.e., assigned 0). This 50% cutoff was set to ensure that we had enough variation at baseline and endline. A higher cut-off, such as greater than 80%, leads to some subtasks having no child marked as succeeding at the subtask at baseline or endline.

There are three common dichotomous IRT models. First, a 1-parameter which estimates each test item's difficulty. For example, word identification is a task that was answered correctly by a smaller proportion of our returning OOSC sample (24%) than the letter name identification (63%). The 1-parameter model will use this information to assign a higher difficulty to the word identification subtask than the letter name identification task. A 2-parameter model also assumes that questions can vary in their discrimination. Discrimination is defined as a question's ability to differentiate between participants with larger and smaller values of the latent trait being estimated. For example, a numeracy task with perfect discrimination would be one that is answered incorrectly by all students below a fixed value of numeracy ability but is then answered correctly by all students above the same fixed value. A 1-parameter model still calculates a discrimination parameter but it assumes all items in the test discriminate high-performing and low-performing participants in the same way. In a 2-parameter model the model assumes that discrimination can vary across questions (StataCorp, 2023). A 3-parameter model adds a guessing parameter which assumes that some of the variation in success rates on items can be explained by the fact that a participant might have guessed the right answer. So, if a question is difficult and discriminates well but is answered correctly by a weaker performing student, a 3-parameter model will try to capture how often participants guessed the right answer to a question.

We conducted all our IRT analysis in STATA 18. We use a 1-parameter IRT model for both EGRA and EGMA scores. We do not use a 2-parameter model because for EGRA subtasks a 2-parameter model failed to converge. This means that when we allow EGRA items to vary in

how well they discriminate between stronger and weaker students, the model could not calculate stable estimates of the model's parameters. This might be because high-performing children in our sample failed easier questions in patterns that were hard to explain by the model, or vice-versa. Another explanation could be that the IRT model we use, which is a unidimensional model of latent ability, may be too simple to explain literacy ability in our sample. Our model assumes that literacy can be measured by one latent trait. It may be that more than one ability is responsible for higher performance on literacy tasks and literacy ability should be split into more than one trait. For example, perhaps literacy should be measured using a multi-dimensional IRT, as described by Hartig and Höhler, 2009, which calculates two latent traits such as (1) "listening" and (2) "reading comprehension" for literacy. The fact that 18% (133/725) of control students got the maximum score on the listening comprehension task at endline while only 2% of control students (15/725) got the maximum score on passage reading at endline might indicate that language ability should be modeled as having of several underlying competencies and can't be summarized into one value for literacy ability in our sample. We note that we are able to successfully estimate a 2-parameter model for EGMA subtasks but we model numeracy using a 1-parameter model to be consistent with our calculation of literacy ability.

For both EGRA and EGMA, we run an IRT model on baseline and endline data that produces estimates of item difficulty and each student's ability in each phase. The raw latent ability measure that is calculated by the model typically ranges from -4 to 4 with smaller values indicating lower values of the latent trait and higher values indicating larger values of the latent trait. However, these values are often hard to interpret so we report all results for our regressions using standardized values of the latent ability trait. We include in the EGRA IRT model 7 dichotomous test items for success or failure on each, and for EGMA IRT model we use 6 dichotomous items for success or failure in each EGMA subtask. We run the IRT model and store estimates of baseline and endline numeracy and literacy ability from the IRT model. We then run our ITT and TOT regressions using our measure of literacy and numeracy ability instead of the raw endline and baseline scores. This leads to the following analytical model:

$$\theta_{ij}^{el} = \beta^*_i T_j + \beta^* \theta_{ij}^{bl} + X_{-\theta_{bl}}{}'_{ij} \gamma + \propto'_s + \varepsilon^*_{ij}$$

Where:

- θ_{ij}^{el} denotes the numeracy or literacy for child i in community j at endline calculated using IRT
- θ_{ij}^{bl} denotes the numeracy or literacy for child i in community j at baseline calculated using IRT

IRT Estimates

Ability	ITT (SDs)	TOT (SDs)
Literacy (EGRA)	1.26***	1.38***
Numeracy (EGMA)	0.81***	0.89***

We find that the IRT estimates corroborate our main results and show positive and significant effects of the Luminos program on numeracy and literacy. We find that children assigned to the treatment group increased in their literacy by 1.26 SDs and increased in their numeracy ability by 0.81 SDs. When we account for non-compliance, treatment effects increase to 1.38 SDs gains in literacy ability and 0.89 SDs gains in numeracy ability. These estimates are very close to a raw average of the individual ITT effects the 7 EGRA subtasks (1.22 SDs) and the 6 EGMA subtasks (0.76 SDs).

Appendix B2: Analytical model

Mathematical specification

For the primary research question, as well as the sub-questions, we estimate the effect of the Luminos program using the following Ordinary Least Squares model:

$$Y_{ij} = \beta^*_i T_j + X'_{ij}\gamma + \propto'_s + \varepsilon^*_{ij}$$

Where:

- Y_{ij} denotes the outcome variable (reading/numeracy/SEL score/attendance) for child i in community j
- T_j denotes the treatment status of community j (1 for the treatment group; 0 for the control group)
- β_i is the estimated treatment effect
- $X'_{ij}\gamma$ is a vector of child-level covariates, including the baseline value of the outcome variable, a binary variable for gender, age at baseline, and a binary for whether the child reported attending school prior to the study
- $\propto'_s$ is a vector of dummy variables corresponding to the stratum that the child is found in i.e. implementing partner crossed with county
- ε_{ij} denotes the child error term i, clustered at the community-level j, since random assignment occurred at the community-level
- * denotes the sampling weights applied to each community, which is equal to the inverse probability of being sampled from all eligible communities within the stratum. These weights are necessary to include in all analyses to recover unbiased estimates, since the probability of treatment varied by stratum (this was done intentionally in order to ensure that each implementing partner had the target number of Luminos classrooms in the 2022-23 academic year).

To produce estimates of standardized effects, we run the same model as above, but for each outcome variable we normalize values using the control mean and standard deviation at endline, and for each baseline outcome variable we normalize values using the control mean and standard deviation at baseline.

For any variables that have missing baseline values (because for instance the child refused to answer that question), we include a dummy variable in the regression that is equal to one if the child is missing that baseline value and zero otherwise, and set the missing baseline value to zero. This ensures that we do not drop observations in our regression.

To produce TOT estimates, we define ‘treated’ as ‘having attended any Luminos classes, per attendance records’. Per this definition, 93% of children in the OOSC sample who were assigned to treatment communities are considered ‘treated’, while 0% of children in the OOSC

sample who were assigned to control communities are considered `treated' (i.e. no control children received any of the program). The binary variable for attending any classes that is used for TOT analysis is *attendance_var*. For the TOT estimates, we instrument this attendance variable using assigned treatment (*treatment*) in a 2SLS model with the same covariates, strata fixed effects, weights, and clustered standard errors as in the ITT model.

To produce simple mean estimates of outcomes at baseline or endline (including difference-in-difference estimates) that are consistent with the ITT and TOT models, we include strata weights in all calculations.

Note that the SEL module was administered to a subset of children, per our design document, and so the sample sizes for all SEL analyses are smaller than for the EGRA and EGMA analyses.

To address possible outliers in the calculation of per minute scores, we winsorized per-minute scores across the entire sample of OOSC and GSC at the 2% level. We winsorize by replacing all values above and including the 98^{th} percentile value, and below and including the 2^{nd} percentile value, with the next value below the 98^{th} percentile and above the 2^{nd} percentile, respectively.

www.IDinsight.org
@IDinsight

IDinsight