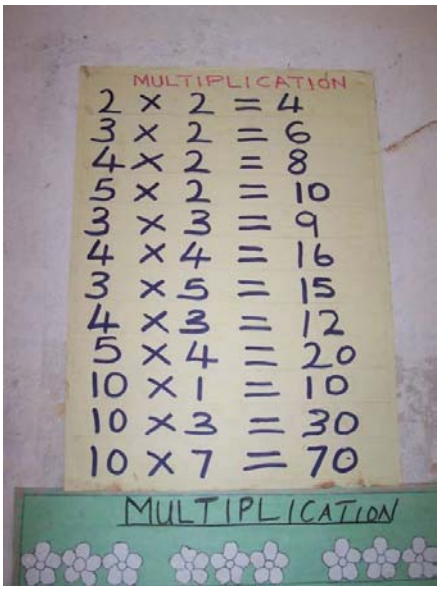
THE WORLD BANK

Information for Accountability: Impact Evaluation of EGRA and teacher training



RTI INTERNATIONAL

Table of Contents

Figures

Tables

Abbreviations

CESLY	Core Education Skills for Liberian Youth [USAID program]
CIASES	Centro de Investigación y Acción Educativa Social [Nicaraguan nongovernmenal organization]
DEO	District Education Officer
DIBELS	Dynamic Indicators of Basic Early Literacy Skills
EGMA	Early Grade Mathematics Assessment
EGRA	Early Grade Reading Assessment
GLH	General Linear Hypothesis
LC	listening comprehension
LTTP2	Liberia Teacher Training Program
MOE	Ministry of Education
NC	North Carolina
ORF	oral reading fluency
PMP	Performance Management Plan
PTA	parent-teacher association
RTI	RTI International [trade name of Research Triangle Institute]
SD	standard deviation
US	United States
USAID	United States Agency for International Development

1. Executive Summary

1. This document reports on the results of an Early Grade Mathematics Assessment (EGRA) study done in Liberia in May 2010 as part of process of collaboration between the Ministry of Education, the World Bank, and USAID, with technical assistance provided under contract by RTI and its consultants and collaborators. The EGMA assessment was conducted as a component of the overall EGRA Plus: Liberia final assessment, and it will be analyzed and reported on within this context.

2. Building on the success of the Early Grade Reading Assessment (EGRA) as a measurement tool, many countries have begun to show interest in moving away from assessments alone and toward interventions focused on changing teacher pedagogy, and as a result, increasing student reading achievement. Liberia's path toward improved standards in reading and now mathematics started with a World Bank-funded pilot assessment using EGRA in 2008, which was used as a system-level diagnosis. Based on the pilot results that showed that reading levels of Liberian children are low, the Ministry of Education (MOE) and USAID/Liberia decided to fund a two-year intervention program, EGRA Plus: Liberia, to improve student reading skills by implementing an evidence-based reading instruction program. EGRA Plus: Liberia was designed as a randomized controlled trial. The WB led the process of designing the trial, which resulted in three groups of 60 schools that were randomly selected into full treatment, light treatment, and control groups.

3. These groups were clustered within districts, such that several nearby schools were organized together. The intervention was targeted at grades 2 and 3. The design was as follows: The control group did not receive any interventions. In the "full" treatment group, reading levels were assessed; teachers were trained on how to continually assess student performance; teachers were provided frequent school-based pedagogic support, resource materials, and books; and, in addition, parents and communities were informed of student performance. In the "light" treatment group, the community was informed about reading achievement using school report cards based on EGRA assessment results or findings and student reading report cards prepared by teachers. The results from the final EGRA Plus: Liberia evaluation indicate that students in full treatment schools are learning to read three times faster than their counterparts in control schools. As a result of these efforts, the Liberia's Ministry of Education, supported by USAID through Liberia Teacher Training Program Phase 2, is unfolding reading and math interventions to more than 2,000 schools in 10 out of 15 Liberia's counties.

4. Through the Spanish Impact Evaluation Fund, the World Bank (WB) tasked RTI in April 2010 to conduct the Early Grade Mathematics Assessment with the main purpose to determine if the EGRA Plus: Liberia project and its intervention has had any secondary impact on improving student performance in mathematics. The study and this report will describe the current level of student learning in basic math for 2nd and 3rd graders, obtain a measure of progress achieved by students between 2nd and 3rd grade, as well as discern a difference in math skills

between students in EGRA Plus control, treatment, and light treatment groups. The report will also discuss the improvements in reading given the early engagement of the World Bank in funding the onset of EGRA+ activities and most importantly designing this randomized controlled trial.

5. As is the case with reading, a strong foundation in mathematics during the early grades is crucial for success in mathematics in later years. Mathematics is a skill very much in demand in today's economy as has been demonstrated by various economists. Most competitive jobs require some level of mathematics skill. It has also been noted that the problem-solving skills and mental agility and flexibility that children develop through mathematics transfer to other areas of life and work.

6. We found that for full treatment, EGRA Plus increased math scores in number identification, quantity discrimination, addition, subtraction, multiplication and fraction knowledge. For light treatment, scores increased for multiplication and fractions, but decreased in number identification. More research is necessary to determine whether the full treatment effects were due to the close relationship between reading skills and outcomes in other subjects, or whether the pedagogical techniques that the teachers obtained in EGRA Plus were also effective in other subjects. However, it does buttress the point of view that reading can be a starting place for quality improvements in other subjects, and even at higher levels in the education sector.

7. Based on the results from the EGRA Plus Project in the context of both reading and math, the following are some of the most important recommendations for the future efforst in improving reading and mathematics in early grades in Liberia:

- **Scale up the EGRA Plus program.** Given the remarkable success of EGRA Plus, there appears to be an opportunity for the Liberian Ministry of Education to scale up and expand the intervention with the focus on both reading and mathematics. The Liberia Teacher Training Program (LTTP2) is a potential incubator for further interventions and offers an opportunity to determine whether the remarkable impacts of this program can be replicated at scale. For the last calendar quarter of 2010, at USAID's direction and with remaining EdData II Task 6 funds, RTI expanded the EGRA Plus: Liberia intervention to all schools—control, light, and full—for another semester.
- **Develop benchmarks for mathematics and reading.** The wealth of data obtained in the three waves of assessment from EGRA Plus provide enough evidence for the Liberian Ministry of Education to determine what rates of fluency, comprehension, and word skills are necessary at each level. Such a benchmark development process will help to target resources and efforts, to invigorate the efforts to improve educational outcomes. It is advised that at the same time, the benchmarks for performance in mathematics are developed as well.
- **Target reading and mathematics techniques using professional development.** Liberian teachers have been proven to be receptive to new pedagogical techniques and strategies. With targeted efforts, teachers can improve how well children read and do mathematics, quite quickly. We recommend that the evidence from this program

be included in pre-service and in-service teacher professional development programs of the Liberian Ministry of Education going forward. Through LTTP2, the MOE will develop the intervention in mathematics in Grades 1-3. The experience from the reading intervention will be used in order to ensure that the teaching package that will be developed is easily absorbed by Liberian teachers.

- **Use reading improvements to increase learning in other subjects.** The findings showed that reading improvements have the potential for carryover effects in other subjects, in this case mathematics. This suggests that reading is a ripe subject for interventions, since other subjects might be improved by the simple method of increasing reading outcomes.
- **Improve girls' reading (and mathematics) achievement.** The findings from the EGRA Plus Project showed that while boys outperformed girls at the baseline, with instruction and investment, girls could narrow and even close the sex gap. Therefore, education officials can and should demand high achievement for girls in the classrooms under their jurisdiction, and efforts should be made to encourage teachers to have high expectations for girls.
- **Expand the use of scripted programs for lesson delivery.** The experience of EGRA Plus makes clear that scripted lesson plans can be a part of an effective program for reading improvement. The increased rates of learning between the midterm and final assessment show that while there was some initial resistance to such methods, the creation of and support for lesson plans for teachers has a high likelihood of continuing to be effective in Liberia.

2. Introduction

8. The Early Grade Reading Assessment (EGRA) Plus: Liberia program (2008–2010) was an experimental intervention. The intervention was part of a joint collaboration among the Liberian Ministry of Education, World Bank Liberia, and USAID/Liberia. The baseline assessment was conducted in November 2008, the midterm assessment was conducted in June 2009, and the final assessment took place in June 2010. As part of the June 2010 final assessment, the WB funded an implementation of the Early Grade Mathematics Assessment with the two main goals: 1. Establish a baseline in mathematics that can be used for future efforts, and 2. Determine if EGRA Plus intervention that was focused on improvements of reading had any secondary impact on improvement of student performance in mathematics.

9. The EGRA Plus: Liberia program used empirical data from reading assessments in grades 2 and 3 to track progress toward quality improvements in early grade reading instruction. In this report, we would like to thank Nathalie Lahire and Muna Meky from the World Bank, and Amber Gove from RTI for developing the EGRA Plus: Liberia Project design.

10. The research and intervention design allowed for the comparison of three different groups. The first was a control group that received no program interventions, but whose performance was measured (without alerting them to the fact there would be repeated measurement). The second group, the "light" intervention, was a set of schools where parents and community members were provided student achievement data in the area of literacy; they were made aware that there would be testing again. In addition, light intervention teachers were trained in the development of a student reading report card, which they issued four times a year. The final group, the "full" intervention, provided an intensive teacher-training program targeting reading instructional strategies, in addition to the same type of information on student achievement that was provided to parents and communities in light treatment schools. Note that the assignment of schools into treatment groups was random and proportional to enrollment in public schools, while accounting for geographic clustering.

11. In this report, the data will be presented in two ways. First we will look at the reading results at project completion by comparison with baseline and midterm assessment results. We will then discuss the mathematics performance at the time of the final assessment and compare it across different school types. We briefly describe the methodology used to conduct these assessments. During November 2008, a national baseline assessment of early grade literacy skills was performed in 176 schools with 2,988 students.[1] The target (and the assessment) was targeted at 60 control, 60 light, and 60 full treatment schools.[2] In each

[1] The sample size was to have been 180 schools; the four missing schools were assessed in January and February 2009, but were not included in the baseline data analysis.

[2] The sampling procedure used in this study and in the intervention was one means of identifying the true impact of the program. Without having a counterfactual or comparison group, it would have been impossible to know whether any impacts we saw were the result of program effects, typical growth over the course of the

school, either 10 or 20 students were assessed, depending on the size of the school and number of teachers. The assessment itself had several sections, all of which had been tested in a variety of other low-income countries, as well as in the June 2008 pilot assessment in Liberia.

12. The June 2009 midterm assessment was conducted in the same EGRA schools. A total of 175 schools and 2,882 students were included in this survey. The June 2010 final assessment was conducted in 175 schools and with 2,688 children. As was the case with the baseline and midterm assessment, either 10 or 20 students were assessed, with the target to have at minimum 10 students from grade 2 and 10 students from grade 3, depending on the size of the school. For all three assessments, students were randomly selected using a systematic sampling procedure implemented by assessors, rather than teachers, in order to prevent teachers from selecting only the best students. The only addition to the final assessment in June 2010, was the EGRA assessment tool.

13. Analysis of the EGRA itself showed that the assessment was reliable and that its various sections assessed different parts of the underlying early grade reading skills, in addition to tying together well as a reliable test. In fact, the final Cronbach's alpha results showed reliability of 0.87, which is quite good, and similar to what was found at the baseline and midterm. The analysis of the EGMA instrument was analyzed for its reliability using Cronbach Alpha analysis and was found to be reliable.

14. The beginning portions of this analytical report lay out the various sections of the assessment, and point out how they are related to important characteristics of early reading skills and proficiency. Note that the purpose of this report is to examine the outcomes from the three rounds of EGRA assessments to determine whether there was a program impact that could be identified.

15. This analytical report is organized as follows:

- First, we present the results of an Early Grade Mathematics Assessment (EGMA), compared by treatment group.
- Second, we present Early Grade Reading Assessment analysis, compared by treatment group and by assessment (baseline, mid-term, final assessment)
- Third, we present recommendations for the sustainability and scale-up of the program.

school year, or changes that applied to all students equally. Having a control group allowed us to differentiate among those possibilities. As noted, in this case, there was one control group and two experimental groups (one having a full intervention and one a light intervention).

3. Early Grade Intervention in Reading

16. The EGRA Plus: Liberia intervention, designed based on the findings of the World Bank pilot assessment of reading in 2008, was itself based on a three-stage intervention strategy. First, a baseline reading assessment was implemented in a nationally representative set of Liberian primary schools. This assessment not only served as the baseline for all the impact evaluations, but also informed the intervention itself, taking student achievement evidence as the first step in assessing teacher training needs, and developing teacher professional development courses to respond to the critical learning areas for improving student achievement.

17. Second, RTI, in collaboration the Ministry of Education and supported by Liberian Education Trust, implemented a teacher professional development program that included intensive, week-long capacity-building workshops. These workshops gave teachers an opportunity to learn techniques for high-quality instruction in early grade reading. Teachers also received ongoing professional development support and regular feedback regarding their teaching. The intervention was buttressed with activities designed to foster community action and stakeholder participation, particularly around the production and dissemination of EGRA findings reports at various stages in the EGRA Plus intervention. The project also encouraged meetings between school managers and community members. Light intervention schools received primarily this set of school and community action activities, while full intervention schools also received onsite professional development and supervision support for teachers in grades 2 and 3. Activities related to teacher professional development and community participation went on for the full duration of the project.

18. The third major intervention activity was an additional two rounds of EGRA, which allowed for a longitudinal research design. This design allowed researchers and the Ministry of Education to identify whether and how the interventions had a significant impact on student achievement, as well as which causal mechanisms were responsible for the project's success. Coupled with this was an Early Grade Mathematics Assessment that was conducted in order to establish a baseline and determine if improvement in reading had any impact on mathematics scores.

3.1 Lessons That EGRA Plus: Liberia Offers for Future Education Projects

19. EGRA Plus: Liberia was designed to pilot an effective model of teacher support that would lead to improved learning outcomes. As such, it pulled all of the levels together—from the national-level staff to the strong involvement of parents. It demonstrated how teachers are best supported by Coaches and District Education Officers (DEOs), and how Coaches and DEOs in turn are supported by the project management and the MOE. This would have not been possible had the project been focused on too many different goals. In the case of EGRA Plus, the sole focus was reading, and all resources and attention were channeled toward

improving student reading outcomes. However, these lessons learned can be use for improvements of mathematics and other subjects in Liberia.

20. Currently in Liberia, all efforts to improve the delivery of education services stall at the District Education Office level; thus, little to no further support is provided to teachers by subject-matter specialists. This is because DEOs are responsible for all of their assigned schools, and they do it all—from payroll to school management and teaching, leaving little or no time for pedagogical support to schools. What's missing is the extension of the DEOs' office to the school level. EGRA Plus: Liberia introduced this bridge. The EGRA Plus project used a one-step-only cascade whereby teachers were trained by Coaches at a cluster level for several weeks and then supported through in-school visits per year that included coaching and supervision. In other words, Coaches were trained first, and then they in turn trained, supervised, and mentored the teachers in the classroom. Liberia needs to move in the direction of instituting a role of a pedagogical advisor (Coach) in order to ensure timely and effective support to its teachers.

21. Apart from this important component, the following notes suggest effective implementation tips, organized around the key inputs that seem to have made a difference.

3.1.1 Teacher and Student Learning Resources

- **Time on task.** Specific lesson plans were provided for EGRA Plus, but there had to be a realistic number. There are numerous holidays and interruptions of teaching in Liberia. It is important to make sure that the scope and sequence designed for a reading intervention are exactly in line with the number of realistically available days for teaching. Yet the lesson plans must, at the same time, ideally be able to produce beginning literacy at the end of their sequence, in one year, if they are to be implemented with fidelity.
- **Lessons need to be tightly scripted.** This is particularly important when teachers do not have necessary lesson-planning skills, or skills in teaching reading, as is most often the case in poor countries. When the lessons are scripted, teachers learn both content and pedagogy as they go. Their application of the scripts will not be perfect in the beginning, but by the end of the first semester, they will have a good sense of the instructional model and how to learn the content. Eventually, good teachers can and will depart from the script. But a tight script is a vital foundation and starting place.
- **Packaging of materials.** The teacher manual needs to be in one book and needs to be durable. If it is too large, it needs to be split into two volumes, one for each semester. This was done in Liberia. But the key is that all resources need to be in one place, and sequentially available, with not much multi-sourcing of alternative techniques and resources. Providing teachers with lots of options often seems good to donors, but can actually be crippling.
- **Curriculum-based assessment.** It is important that teachers assess student performance on a regular basis and issue student report cards to parents about their

children's performance. Teachers need to be shown how this is done and be supported while doing so. By the time they do it two or three times, they will have learned how.

- **Periodic assessment and reporting to parents.** The teacher manual needs to contain step-by-step instructions for assessing students and creating separate student report cards for parents (this would be an individual student card) and parent-teacher association meetings (this would be a card that represents averages for the school).
- **Decodable books.** These books are important for teaching sounds, and must be provided, but they will be used only if they are tied to the lesson plans in the (above-mentioned) manual. No such books exist for Liberia, and they need to be developed.
- **Library books.** The more students have to read the better. The challenge is to secure books for the schools and to enforce their use. Parental involvement is important, as one of the requirements by teachers is that children read at home every day for at least 20 minutes. However, this arrangement must be agreed upon between school authorities and parents.
- **Pocket charts.** Teachers received pocket charts that they could use for arranging letter cards when teaching sounds and spelling, as well as constructing sentences using word cards. Again, the use of pocket charts needs to be required, taught, and checked upon. There may be other techniques that work well, but selecting *one* such technique makes the logistics easy and reduces teacher confusion.
- **Various trackers/logs.** For EGRA Plus these included a library log, log to track students' reading at home, and trackers for assessing students. These trackers were used regularly to introduce and enforce accountability.

3.1.2 Teacher Training and School-Based Support

Teacher training

- **Cluster-based training.** Training in EGRA Plus was organized once per semester at the cluster level. Teachers from intervention schools were invited for training that was one week long. One week really is not enough, especially when teachers completely lack skills, but when coupled with the monthly school-based support that supplements this training, it works. Since this was a cascade—meaning that we first trained Coaches, who then trained teachers in turn—it was important that Coaches were trained in the same way the teachers were going to be trained; i.e., much as if they were themselves going to teach children to read. This way, as Coaches were being trained, they would know exactly what to do with the teachers. A one-stop cascade works under these circumstances, but it is unlikely that more than one stop would work.

School-based support

- **Purpose and frequency of visits.** Visits by the Coaches to the school level were organized for two purposes. The first was to support teachers once per month. Sometimes teachers received more than one visit depending on the need, but one visit per month was a minimum. The second was to work with PTAs and teachers on

student report cards, as well as other aspects of the intervention (e.g., request parents to make sure that children read at home every day).

- **Fidelity of implementation.** These visits had to be systemized so that all Coaches were doing exactly the same thing every month. Such systematization was written out specifically for EGRA Plus, and as such it provided clear guidance for both project management and Coaches as to what needed to be done.
- **Accountability.** Coaches were equipped with various logs that tracked teacher performance, and in turn their own performance. One such tracker looked at how far teachers had come with the intervention and, if there was a need, Coaches paid an extra visit to teachers to catch them up. There wass a classroom observation tool that Coaches used to observe a teacher teaching a particular lesson. This was not generic, but was tied very specifically to reading. The feedback was then used to speak about perfecting the skills of teachers.

Coaches

- **Training.** Coaches were trained by a reading expert, either international or local during a week-long training event. Training of coaches was organized once per semester (thus twice per year).
- **Hiring.** The key is to hire committed master trainers who care about what they do. Paper qualifications matter much less than care, intelligence, drive, and willingness to learn.
- **Supervision.** The work of the Coaches was verified through EGRA assessments (both formal and informal) and this was the best indicator of their commitment. If the data from their schools showed no improvement, we knew that they were not doing their job well. So hiring of good master trainers is key, but without strong supervision, hiring is only half the work. Using this approach, out of 15 Coaches, we needed to replace only one.
- **Support to Coaches.** EGRA Plus ensured sufficient funding to Coaches for the use of cell phones. This way they could communicate at any time with our reading expert, who resided in Monrovia. In addition, the reading expert conducted regular weekly or bi-weekly discussions with Coaches in order to determine progress and challenges. Also, the reading expert visited each coach once per semester. During this visit, at least one of the schools (picked by the reading expert and not by the Coach) was visited to determine the uptake by teachers. Finally, district-level competitions were organized through which Coaches and their respective District Education Officers wrote their success stories and submitted them, with the opportunity to win prizes.

3.1.3 EGRA Assessments

- **Formal assessments.** All schools (control, light, and full) received a baseline, midterm, and final assessment. This was the best way to know if the intervention was working over time.

- **Informal assessments.** Project management conducted informal assessments halfway through each semester in a subsample of intervention schools. This was a good mechanism to determine if Coaches were doing their job and if adjustments needed to be made. At the same time, it served as a good tool to keep the project management working hard.

3.1.4 Community Outreach

- **Reading competition:** It is very important to have cluster schools compete in reading. Coaches organized these with PTAs. Key drivers behind this at the beginning were the Coach, teachers, and principal, and then parents were invited to the competition.
- **District-level competition.** Coaches and DEOs (who were representing their schools) competed among each other. This was organized by the project management during the semiannual refresher training for Coaches.
- **Radio shows.** In each of the target districts, four radio shows were aired, one per month. These radio shows talked about the importance of reading, current reading levels of students in Liberia, and tips for parents and teachers on what they could do to help children learn how to read.
- **PTA meetings:** Student performance and progress were discussed with parents during the PTA meetings. This was the time when the school reading report card was discussed, parents were given tips on what and how to support at home, and the schools told about their efforts to help children learn how to read.

3.1.5 System Improvements

- **Policy improvements.** The commitment to the revival of reading in Liberia is best illustrated by the Ministry of Education's issuance of a letter to all EGRA target schools requiring teachers to teach reading every day for 45 minutes. EGRA is currently included in the MOE's Education Plan as a result of this commitment. Our hope is that the explicit teaching of reading will be brought back into the official curriculum. Apart from reading, it will be important to ensure a similar policy focus in the area of early grade mathematics.
- **Transfer of reading skills to MOE staff.** MOE staff attended each of the EGRA reading workshops. The key was to train District Education Officers at the central level. More needs to be done in order to fully strengthen the MOE capacity. For a project that was small in size and also a pilot, EGRA Plus: Liberia made sure to do as much as could be done in a short period of time.
- **Transfer of assessment skills to MOE staff.** Dozens of MOE staff were trained on how to assess student performance, using EGRA, to the point that they could teach it as well.
- **Transfer of data entry and analysis skills to MOE staff.** Data entry was performed and supervised by the MOE. We transferred skills in building the EGRA database

(the first built by MOE after the war), as well as in conducting simple statistical analysis. More support is needed, especially in the context of data analysis.

4. Sustainability and Scale-Up

22. EGRA Plus provided an opportunity to scale up the project and work to ensure sustainability. One component of the EGRA Plus: Liberia project was to assist in building the capacity of MOE staff. By the end of Year 1, EGRA Plus had conducted six capacity-building workshops at which MOE staff were trained, including two EGRA assessment workshops, three EGRA reading workshops, and one workshop on data analysis and reporting.

23. One of these reading workshops marked the beginning of more in-depth involvement of District Education Officers from the EGRA target districts. While during Year 1 they were engaged in supporting the project at the district level, from August 2009 onward they were fully involved in the training activities and in the support to EGRA target schools. They were all trained in instructional methods for reading during the project's refresher course that took place in August 2009. Between September and December 2009, and then in January and June 2010, each DEO, along with Coaches, visited at least eight schools. This gave the DEOs an opportunity to practice some of their skills in teaching reading as well as to provide pedagogic support to teachers. At the end of the first semester, they attended a refresher training in December 2009 together with the Coaches. Finally, DEOs will be invited to attend the final reading policy workshop planned for the end of the project in December 2010.

24. At the national level, the capacity building of MOE staff was further deepened to allow more opportunities for turning newly acquired knowledge into practice. Dozens of MOE staff learned how to assess student reading, and most of them were also deployed for data collection. In Year 2 of the project, they were paired with the project staff to learn how to calibrate (equate) instruments, co-facilitate assessor training, supervise data collection, enter and analyze data, supervise the implementation of reading intervention, and assist with the training and support provided to teachers.

25. The goal of these capacity-building efforts was to lay the foundation for expansion of reading support to all of the schools in the current EGRA districts, as a first step. It is our hope that the donors and MOE will recognize these efforts and start planning soon for ways to ensure that all children in Liberia can experience the same increases in their early reading skills.

26. As a result of these efforts, it was agreed by the MOE and USAID that the EGRA Plus schools would be integrated into LTTP2 as of January 2011. Via this integration, EGRA Plus will have demonstrated to the communities and schools, especially the control schools, that hard work and success are rewarded: These schools will receive further support through LTTP2.

27. As equally important, these schools will receive support in both reading and mathematics for Grades 1, 2, and 3. Mathematics intervention will require more attention than reading given that at this stage given the math intervention has not yet been piloted.

5. EGRA and EGMA Assessments

28. This section briefly introduces the various EGRA and EGMA sections, so that the analysis below will be meaningful. The EGRA tool consists of a variety of sections, and they have been somewhat differentially applied in various countries in order to ensure context-specific relevance. The EGRA Plus: Liberia tool assessed the following set of skills:

1. *Orientation to print*: awareness of the direction of text, and the knowledge that a reader should read down the page. Note that this section is not addressed in the analyses because all the assessed children always answered correctly.
2. *Letter naming fluency*: ability to read the letters of the alphabet without hesitation and naturally. This is a timed test that assesses automaticity and fluency of letter recognition. It is timed to 1 minute, which shortens the overall assessment and also prevents children from having to spend time on something they find very difficult.
3. *Phonemic awareness*: awareness of how sounds work with words. This is generally considered a prereading skill, and it can be assessed in a variety of ways. In the case of Liberia this was assessed by asking the student which word, out of three, started with a different sound (e.g., *ball*, in "mouse, ball, moon").
4. *Familiar word fluency*: ability to read high-frequency words. This assesses whether children can process words quickly. It is timed to 1 minute.
5. *Unfamiliar (or nonsense) word fluency*: ability to process words that could exist in the language in question, but do not, or are likely to be very unfamiliar. The nonwords used for EGRA are truly made-up words. This section assesses the child's ability to "decode" words fluently. It is timed to 1 minute.
6. *Oral reading (connected text) fluency*: ability to read a passage, about 60 words long, that tells a story. It is timed to 1 minute.
7. *Reading comprehension*: ability to answer up to five questions based on whatever portion of the passage the child could read.
8. *Listening comprehension*: ability to follow and understand a simple oral story. This section assesses the child's ability to concentrate and focus to understand a very simple story of three sentences with simple noninferential (factual) questions. It is considered a prereading skill.

29. In order to prevent "teaching to the test," or memorization, the three assessments (baseline, midterm, final) used different passages and reading comprehension questions.

30. In addition to the three rounds of EGRA assessments implemented in EGRA Plus, separate funding from the World Bank and collaboration with the Ministry of Education and USAID/Liberia allowed us to evaluate whether the EGRA Plus program had any impact on mathematics outcomes. Therefore, we developed and applied **an Early Grade Mathematics Assessment** in Liberia, as explained below. The purpose was to evaluate whether the EGRA Plus program had an impact on student achievement in mathematics, although no portion of the EGRA Plus program was developed to target mathematics teaching or learning.

31. As with reading, a strong foundation in mathematics during the early grades is crucial for success in mathematics in later years. Mathematics is a skill very much in demand in today's knowledge economy. Most competitive jobs require some level of mathematics skill, and the problem-solving skills and mental agility and flexibility that children develop through mathematics transfer to other areas of life and work. The EGMA is an individually administered oral assessment of foundation mathematics skills. It can be used to bring awareness to policy makers and educational authorities as to levels of foundational mathematics learning in their systems.

32. As noted above, an EGMA tool for Liberian context was developed through the CESLY project. This tool, which was based on the curricula for grades 2 and 3, was piloted in two schools, and the results were used to improve the assessment. The CESLY EGMA tool was the starting point for development of the EGMA for the EGRA Plus project, which was then improved upon further by RTI math experts and EGRA Plus staff. Once the draft assessment had been developed, it was reviewed during a stakeholder training workshop held May 3–7, 2010. The draft EGMA tool was piloted in one school and feedback from the pilot and the workshop participants was taken into account while the EGMA tool was finalized. The following sections were included in the mathematics assessment:

1. *Number identification* – Learners were asked to identify particular numbers of varying difficulty levels but appropriate for grade 1–3 learners vis-à-vis the curriculum.
2. *Quantity discrimination* – Learners were asked which of two numbers was bigger, testing place value and number sense. This section was timed.
3. *Missing number* – Given a list of three or four numbers, one of which was missing, the child was asked to identify the missing number.
4. *Addition* – A list of common and simple addition facts was presented to the learners, who were asked to solve them as quickly as possible. There were two subsections within this addition section, with the second presenting slightly more computational problems. The first subsection was timed, while the second was not.
5. *Subtraction* – Similar to the addition section above, learners were presented with simple subtraction problems and asked to solve them. There were two subsections within this subtraction section, with the second one slightly more difficult. The first version was timed, while the second was not.

6. *Multiplication* – Learners were presented with a set of multiplication problems and asked to solve them. This was not timed.
7. *Fractions* – Given several items, the learners were asked to identify fractions, add them, and distinguish which fraction was bigger or smaller. This was untimed.

5.1 EGRA Assessor Training

33. The training occurred May 3–7, 2010, and it was facilitated by the Task Coordinator (Medina Korda), EGRA Technical Coordinator (Ollie White), and RTI's Reading Expert (Marcia Davidson). The MOE coordination committee assigned to work with EGRA from the beginning of the project also attended the training. For any application, EGRA teams always train more assessors than needed, in order to ensure that the assessors who are chosen at the end to be deployed are the best possible performers. The total number of trainees in Liberia was 45, from which the 28 best assessors were selected. The total number of MOE staff trained at this training was 17, and five of them were deployed (note that total number of MoE staff selected for deployment was 10, but the managers of the Core Education Skills for Liberian Youth [CESLY] project[3] and the EGRA Plus: Liberia task split the MOE staff to be deployed evenly through these two projects). Note that formal interrater reliability assessments were used for training and selection.

5.2 EGRA Data Collection

34. Data collection for the final assessment commenced on May 17, 2010, and it ended on June 11, 2010. This allowed for four weeks of data collection in 179 schools. There were nine teams, each team consisting of three members to account for the increased work due to EGMA. In total, 176 schools were assessed. Four schools were not assessed. Two of these were affected by a car accident (a bridge collapsed under one car carrying several enumerators). One control school refused to be assessed because it was being denied the treatment. One light school also refused for the same reason.

5.3 EGRA Data Entry

35. For the final assessment, RTI developed a data entry application using Visual Basic that reduced the time for data entry to a third of what was needed on the baseline and midterm assessments. RTI has been working with a Nicaraguan firm—Centro de Investigación y Acción Educativa Social, or CIASES—for the past several years to develop and improve an efficient and user-friendly data entry system. The EGRA data entry system developed by CIASES offers a low-cost, sustainable solution for minimizing errors.

[3] RTI is a subcontractor to Education Development Center, Inc. (EDC) on the USAID CESLY project. RTI's scope of work is to carry out assessments; also the Liberia-specific EGMA was developed under CESLY.

6. Research Design

36. Table 1 below shows the achieved sample for the baseline, midterm, and final assessments. This table shows that two schools that were in the baseline and midterm full treatment set of schools were not included in the final assessment. Note that the sample of children in the three assessments is presented vis-à-vis treatment status—that is, whether a child was in a control, full treatment, or light treatment school. For the midterm assessment, slightly fewer children were found in control schools and light treatment schools, whereas the numbers of children in full treatment schools were very similar.

37. The achieved sample at the final assessment was smaller for control and full intervention schools, but near the target for the light intervention schools. Note that the sampling procedures for each assessment were done randomly and independently of each other. In other words, no attempt was made to resample children assessed in a previous assessment. Children from the baseline assessment may also have taken part in the midterm assessment, but because children's names were not used, it is impossible to tell with any certainty. Note also that the sampling was done from the students in attendance during the day, therefore using systematic random sampling. Table 1 also shows that the impact analysis contained in this report was based on 2,998 baseline, 2,882 midterm, and 2,688 final assessment participants, for a total of 8,568 children, a substantial sample size for this type of analysis.

Table 1: Achieved EGRA Sample for Baseline, Midterm, and Final Assessments, by Treatment Group, for Schools and Students

		Treatment			
		Control	**Full**	**Light**	**Total**
Schools	Baseline	57	59	60	176
	Midterm	56	59	60	175
	Final	59	58	60	177
Students	Baseline	989	934	1065	2988
	Midterm	944	924	994	2882
	Final	808	916	964	2688

38. More details about the sample used in this analysis can be found in Table 2 below. Disaggregating by the three assessments, the sex, grade, and treatment status of all of the children can be found. Interestingly, while there were more boys than girls in the baseline sample (1,623 and 1,327, respectively), there were more girls than boys in the midterm assessment (1,470 and 1,345, respectively) and the final assessment (1,363 and 1,231, respectively). This suggests that analyses should be done with control variables for sex, such that the differential sampling by sex does not skew the results. This is particularly true for considering the treatment status of children's schools. Where light and control schools were more heavily male than female at baseline, these same schools were more female than male in the midterm and final assessments. It is important to note that these variations are logical given the sampling method, and are not of concern as long as sex and treatment status control variables are included in latter analyses.

39. The columns to the right in Table 2 indicate grade level. Note that in all three assessments, there were more grade 2 than grade 3 children. This seems to be indicative of higher enrollment in grade 2, which is plausibly a result of dropout and/or class size in these randomly selected Liberian schools. In any case, this again is not of particular concern giving the sampling strategy, although it suggests that grade level should be a part of future analyses.

Table 2: Achieved Sample, by Assessment, Grade, and Treatment Group

		Treatment				Level		
	Sex	Control	Full	Light	Total	Grade 2	Grade 3	Total
Baseline	Boys	525	530	577	1632	820	803	1623
	Girls	453	400	482	1335	724	603	1327
	Total	978	930	1059	2967	1,544	1406	2950
Midterm	Boys	424	471	456	1351	733	612	1345
	Girls	502	456	530	1488	757	713	1470
	Total	926	927	986	2839	1,490	1325	2815
Final	Boys	354	433	461	1248	639	592	1231
	Girls	433	470	478	1381	711	652	1363
	Total	787	903	939	2629	1,350	1244	2594

40. Table 3 contains basic descriptive statistics for the baseline study (columns to the left), midterm assessment (middle columns), and final assessment (columns to the right). There is a consistent pattern across this table, with children in the midterm assessment outscoring those in the baseline, and children in the final assessment outscoring those in the midterm. For example, children who participated in the final assessment could read more letters (90.4 per minute) than could those at midterm (80.2) or those at baseline (61.1). Given the fact that the midterm and the final assessments occurred at the end of the year and that the baseline was at the beginning of the academic year, it is expected that children would learn skills that would be identified on the EGRA assessment during the academic year. This would cause higher scores at midterm and baseline.

41. However, the consistent improvement between midterm and final assessment scores suggests that there is more to it than that, since scores also increased between 2009 (midterm) and 2010 (final). The pattern holds for every section: letters, phonemic awareness, familiar word fluency, unfamiliar word fluency, oral reading fluency, reading comprehension, and listening comprehension. In each of these sections, average scores were higher at midterm than at baseline (the intergrade learning effect) and at final than at midterm (additional program impact, including secular trend).

42. The magnitude of the differences seems to have been larger between midterm and final than between baseline and midterm. For example, for reading comprehension, baseline (25.1%) and midterm (25.7%) scores were much closer to each other than to final assessment scores (42.4%). The analysis below shows the results of our investigation of whether the differences identified in this analysis occurred because of the EGRA Plus program, because

of a continued secular trend of improving literacy scores in Liberia, or because of the learning effect identified between the baseline and the midterm and final assessments.

Table 3: Descriptive Statistics for Baseline, Midterm, and Final Assessment

Section	Baseline, November 2008			Midterm, June 2009			Final, June 2010		
	N	**Mean**	**Standard Deviation**	***N***	**Mean**	**Standard Deviation**	***N***	**Mean**	**Standard Deviation**
Letter naming fluency	2,982	61.11	25.30	2,789	80.18	26.69	2,502	90.40	25.26
Phonemic awareness	2,982	3.48	2.29	2,882	4.19	2.62	2,688	5.07	2.86
Familiar word fluency	2,957	9.24	13.89	2,771	14.87	16.30	2,464	24.71	21.63
Unfamiliar word fluency	2,961	2.24	6.01	2,773	2.45	5.88	2,494	7.00	13.30
Oral reading fluency	2,963	19.55	20.03	2,725	25.98	25.21	2,345	34.79	31.98
Reading comprehension	2,963	25.08	24.23	2,725	25.72	28.38	2,345	42.35	37.11
Listening comprehension	2,996	33.56	20.54	2,790	74.69	30.23	2,616	75.20	30.01

7. EGRA and EGMA Reliability Analysis

43. In order to examine whether and how the sections in the Liberian EGRA at the final assessment were reliable, and—critically—whether it could be argued that they tested an underlying skill, we carried out the reliability tests for both reading and math described below using the Cronbach Alpha Reliability Coefficient (See Tables 4 and 5). The Cronbach's alpha score for the overall assessment was at 0.87. These scores are well within the "accepted" range of at least 0.6 to 0.7 for a low-stakes assessment such as EGRA and are in line with what was found at the baseline.[4]

Table 4: Cronbach's Alpha Statistics for EGRA Final Assessment

Item[a]	Item-test correlation	Item-rest Correlation	Average inter-item correlation	Alpha
Letter naming fluency	0.72	0.60	0.49	0.85
Phonemic awareness	0.67	0.49	0.52	0.87
Familiar word fluency	0.87	0.81	0.44	0.82
Unfamiliar word fluency	0.70	0.57	0.50	0.86
Oral reading fluency	0.88	0.83	0.44	0.82
Reading comprehension	0.86	0.79	0.45	0.83
Listening comprehension	0.58	0.41	0.55	0.88
Overall assessment				0.87

[a]The term "item" in this context refers to the EGRA sections. In other words, letter naming fluency, for example, is an item as well as a section.

Table 5: Cronbach's Alpha Statistics for EGMA

Item	Item-test correlation	Item-rest Correlation	Average inter-item correlation	Alpha
Number identification	0.60	0.43	0.14	0.56
Quantity Discrimination	0.28	0.05	0.18	0.65
Missing Number	0.30	0.08	0.19	0.65
Additions level 1	0.69	0.54	0.12	0.53
Additions Level 2	0.38	0.17	0.17	0.63
Subtraction Level 1	0.72	0.59	0.12	0.51
Subtraction Level 2	0.62	0.45	0.13	0.55
Multiplications	0.51	0.30	0.15	0.59
Fractions	0.36	0.14	0.18	0.63
Overall Assessment				0.62

[4] Nunnally, J. & Bernstein, I. (1994). *Psychometric theory* (3rd edition). New York: McGraw-Hill.

44. For EGMA, the Cronbach's alpha is also reliable with the overall of 0.6 coefficient. The lowest alpha for EGMA can be observed on the subtraction item (0.52), while the highest is on quantity discrimination test.

8. EGRA Plus Impact on Early Grade Mathematics Assessment

45. Although mathematics was not a part of the intervention assessed in this report, RTI and the World Bank felt that it would be of interest to investigate whether EGRA Plus had knock-on effects, such as increased pedagogical prowess across subjects on the part of teachers; or whether increased facility with reading would allow children to better understand the mathematics content that they were taught. This section presents a snapshot of mathematics achievement across the three groups of schools (full treatment, light treatment, and control).

46. Table 6 below presents the results of an Early Grade Mathematics Assessment disaggregated by treatment status (intervention group). A substantive (rather than statistical) investigation of the results shows that children in the full treatment group scored higher than both control and light treatment children on all of the EGMA sections: specifically number identification, quantity discrimination, missing numbers, addition fact fluency subsection 1, addition fluency subsection 2, subtraction fluency subsection 1, subtraction fluency subsection 2, multiplication scores, and fractions problem-solving. Light treatment schools outperformed control schools on quantity discrimination, addition fluency subsection 1, addition fluency subsection 2, multiplication, and fractions problem-solving, so they outperformed control schools on only five of the nine sections.

Table 6: Early Grade Mathematics Assessment Results, by Treatment Group

Section	Control			Full Treatment			Light Treatment		
	N	Mean	Standard Deviation	*N*	Mean	Standard Deviation	*N*	Mean	Standard Deviation
Number identification	749	15.38	4.66	889	15.98	4.31	944	14.88	5.14
Quantity discrimination (per minute)	612	0.86	1.51	720	1.59	2.44	778	1.15	1.92
Missing number (raw)	753	3.22	1.01	891	3.24	1.09	945	3.20	1.15
Addition 1 per minute	751	6.68	4.36	891	7.54	4.93	943	6.90	4.41
Addition 2 per minute	750	4.39	16.25	891	5.79	19.53	942	4.31	9.75
Subtraction 1 per minute	748	4.91	3.25	891	5.38	3.47	943	4.87	3.40
Subtraction 2 per minute	746	2.16	3.58	889	2.39	2.60	941	2.02	2.32
Multiplication (number correct)	745	0.65	1.40	889	0.89	1.52	942	0.78	1.42
Fractions (number correct of 6 items, %)	808	4.52	14.68	891	10.19	21.86	943	9.31	21.26

47. A simple comparison such as the one in Table 6 above does not have the statistical power to determine whether there were systematic differences between EGMA scores by treatment group, since it does not indicate whether the differences between groups were small enough to be due to chance. Therefore, we fit multiple regression models, controlling for grade and sex, to determine (1) whether the differences in means between the treatment group and control were statistically significant, (2) the magnitude of those differences, and (3) the effect size of the differences (if any). The results of this analysis are in Table 7 below:

- For number identification, full treatment children outscored control children by 0.64 items correct (*p* value .003) for a small effect size of 0.14 SD. Light treatment children, on the other hand, scored 0.39 items lower than control, although at the .10 level of significance.
- For quantity discrimination, both full (*p* value <.001) and light treatment groups (*p* value .004) had higher fluency scores than control children by 0.72 and 0.28 numbers correct, with effect sizes of 0.34 and 0.16 SD.
- There was no statistically significant difference for missing numbers for either full treatment (*p* value .58) or light treatment (0.94) children.
- In the first addition subsection, full treatment children were more fluent (*p* value <.001) by 0.89 items per minute (effect size 0.19 SD), while there was no difference for light treatment children. The differences were not significant on the second addition subsection either.

- For the simpler subtraction subsection, full treatment children did better by 0.5 problems per minute (*p* value <.001) for an effect size of 0.15 SD. The treatment groups made no difference for the second, more complex subtraction problems.
- For multiplication, both treatment groups had higher achievement by 0.25 and 0.14 problems correct (0.17 and 0.10 SD for full treatment and light treatment, respectively).
- Fractions felt a moderate impact from full treatment (0.31 SD) and a small impact from light treatment (0.26 SD), with children in full treatment schools scoring 6.0% higher while light treatment children did 5.0% better than control children (*p* values <.001).

48. In short, it appears that EGRA Plus had inconsistent and small impacts on mathematics achievement for light treatment children, but consistent although still small impacts on full treatment children. It must be noted that without a pre and post analysis, using these EGMA outcome measures, it would be difficult to say whether the changes that we have identified in this analysis were due to the reading intervention, or to differences in mathematics achievement that occurred prior to the EGRA Plus program administration, or to a combination of the two.

49. However, given that full treatment schools scored lower than the light treatment and control schools on most EGRA sections at baseline, it is less likely that the achievement of full treatment schools was lower in reading but higher in math before EGRA Plus commenced in 2008. Thus, it seems probable that the EGRA intervention had a noticeable effect on children's mathematics achievement. Whether this was due to an overall accountability effect, to an overall time-on-task effect, or to the fact that cognitive skills all tend to work together and help each other, is impossible to say without further analysis.

Table 7: Early Grade Mathematics Assessment Regression Results, Controlling for Grade and Sex

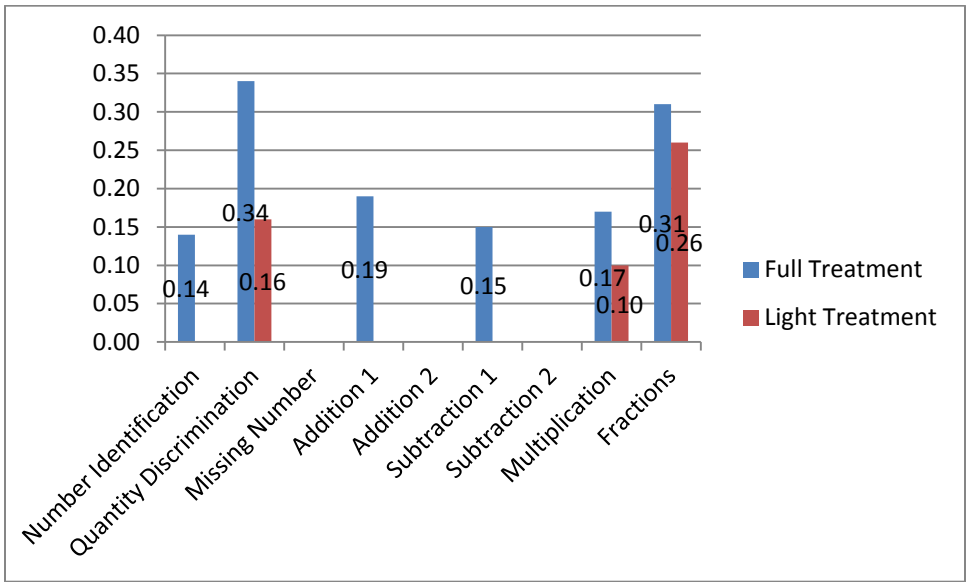
Section	Treatment Group	Coeffi-cient	Std. Error	*T*	Sig.	Confidence Interval		Effect Size (SD)	R^2
						Lower	Upper		
Number identification	Full	0.64	0.22	2.98	.003	0.22	1.07	0.14	.09
Number identification	Light	-0.39	0.23	-1.69	.09	-0.86	0.06	-0.08	.09
Quantity discrimination (per minute)	Full	0.72	0.12	6.19	<.001	0.49	0.94	0.34	.04
Quantity discrimination (per minute)	Light	0.28	0.10	2.90	0.004	0.09	0.47	0.16	.01
Missing number	Full	0.03	0.05	0.64	.52	-0.07	0.14	0.03	.00
Missing number	Light	-0.01	0.05	-0.18	.86	-0.12	0.10	-0.01	.00
Addition 1 (per minute)	Full	0.89	0.23	3.90	<.001	0.44	1.33	0.19	.06
Addition 1 (per minute)	Light	0.26	0.21	1.24	.21	-0.15	0.68	0.06	.05
Addition 2 (per minute)	Full	1.42	0.93	1.53	.13	-0.40	3.24	0.08	.00
Addition 2 (per minute)	Light	-0.09	0.66	-0.13	.90	-1.38	1.21	0.00	.00
Subtraction 1 (per minute)	Full	0.50	0.16	3.03	<.01	0.18	0.82	0.15	.06

Section	Treatment Group	Coeffi-cient	Std. Error	*T*	Sig.	Confidence Interval		Effect Size (SD)	R^2
						Lower	Upper		
Subtraction 1 (per minute)	Light	0.01	0.16	0.08	.94	-.31	0.33	0.00	.03
Subtraction 2 (per minute)	Full	0.21	0.16	1.37	.17	-.09	0.52	0.07	.01
Subtraction 2 (per minute)	Light	-0.15	0.15	-1.00	.32	-.44	0.14	-0.05	.01
Multiplication	Full	0.25	0.07	3.33	<.01	0.10	0.39	0.17	.01
Multiplication	Light	0.14	0.07	1.93	.05	-0.00	0.27	0.10	.01
Fractions (%)	Full	5.95	0.96	6.19	<.001	4.07	7.84	0.31	.03
Fractions (%)	Light	4.98	0.94	5.32	<.001	3.15	6.82	0.26	.02

50. To investigate these relationships further, we produced the following effect sizes. Figure 1 shows the magnitude of the relationship the EGRA Plus program had with student achievement in mathematics. Note that the relationships were strongest in quantity discrimination and fractions, and modest in number identification, addition 1, subtraction 1, and multiplication. In social science research, particularly education research, effect sizes in the range of 0.20 and above are non-negligible and are evidence of a quite successful program on student achievement.

51. The question for further research, then (as already noted), is what occurred in the EGRA Plus program to increase mathematics achievement without any intervention whatsoever on the subject. Moreover, it would be useful to investigate the implications of these findings for the mechanisms by which the EGRA Plus program had such large impacts on reading achievement.

Figure 1: Effect Sizes on Early Grade Mathematics Assessment Outcomes



52. It is a useful (although speculative) thought exercise to examine what mechanisms could have increased the mathematics achievement in this study (that is, if one accepts that the results above are evidence of a program effect of EGRA Plus). Mathematics achievement in general is related, clearly, to reading and comprehension skills. One mechanism might be internal to the children: that is, if children were more skilled in reading (decoding) and understanding what they read, they might do better in mathematics as well since they would now able to understand mathematics text. If that is the case, than we would expect that reading outcomes would predict student achievement in mathematics.

53. On the other hand, the program might have knock-on effects. It is conceivable that if EGRA Plus trained teachers to teach better, or if they acquired mastery of particular pedagogical techniques as a result of EGRA Plus, or if the frequent visits from the literacy Coaches encouraged better teaching across subjects, then it would be the teachers' improved pedagogy (as a result of EGRA Plus) that was responsible for the improvements in mathematics outcomes. The question of interest, then, is whether the increased reading ability of students increased outcomes, or whether it was teachers' improved pedagogy in mathematics. Of course, it is surely not as dichotomous as this example, in that the causal mechanisms likely emanated from some combination of those two (and a myriad of other) factors.

54. The mechanism of change could have been subtly different at the teacher level, of course, since it might have been the motivational aspects of having Coaches, District Education Officers, and directors more heavily involved in the pedagogical process that encouraged teachers to teach better. That assumes, however, that the teachers already had the skills to teach better, but motivation caused them not to do so. This is again a matter of further researchjh, but for this report suffice it to say that this analysis examined all of the factors (skills, motivation, attendance, etc.) internal to or impinging on teachers.

55. The quantitative data available allow a simple exploration of these issues, however. Cognizant that this must be supplemented by further in-depth analysis, Table 8 presents R^2 scores from a variety of multiple regression models. The first column presents the outcome variable, the second the portion of variation explained by scores on oral reading fluency, the third the variation explained by models with both oral reading fluency (ORF) and listening comprehension (LC), and the fourth simple models that include the variables indicating the treatment groups.

56. The findings show that models with oral reading fluency predict more of the variation than do models with listening comprehension. Oral reading fluency is indicative of reading skills (and in many principal components analyses loads heavily as the main predictor of underlying reading achievement), while listening comprehension is more related to oral vocabulary. It appears that children's skills in reading were more predictive of their skills in mathematics than were their oral vocabulary skills, which makes sense given the domains that mathematics skills depend on. The fourth column makes the same point another way: In combination with with oral reading fluency, listening comprehension did not predict much more of mathematics skills than did an oral reading fluency model only (comparing column 3

to column 1). The fifth column includes models with variables indicating full and light treatment groups. This predicted very little of the variation in mathematics outcomes. The sixth column portrays the outcomes of regression models with treatment groups as well as oral reading fluency. Compared to models with just ORF (column 2) the models did not add much to the predictive power except for the quantity discrimination and fraction sections.

57. Interpretation of this table must proceed cautiously since the study was not designed to determine the causal mechanisms for increased mathematics achievement, but only to examine whether there were differences in achievement by group. That said, it appears that the models do not do a particularly good job predicting mathematics outcomes. Where the models are predictive, they depend on oral reading fluency (a child's skill with reading) slightly. Number identification, addition, and subtraction seem to have been related to reading skills, at least somewhat. Note that the treatment groups only increased the predictive power of the models for the fraction and quantity discrimination sections. These might have been more dependent on the improved methods that the teachers gained as a result of the EGRA Plus program.

58. One interpretation of these findings is that if an increase in student skills was the mechanism by which the mathematics achievement increased, then it likely was not restricted to that which could be measured by oral reading fluency. It appears, then, that EGRA Plus was able to increase student skills beyond the areas that the program intended.

59. On the other hand, the evidence suggests that at least some of the impact of EGRA Plus on mathematics achievement was as a result of unmeasured (read: nonreading) factors. That is because the predictive power of models with treatment group predicted almost none of the variance. In fact, the two sections whose variation was somewhat predicted by treatment group (quantity discrimination and fractions) might have been the two sections that depended on student comprehension and reading the most.

Table 8: Multiple Regression R^2 Results by Model

Section	Oral Reading Fluency (ORF)	Listening Compre-hension (LC)	ORF + LC	Treatment Groups	Treatment Groups + ORF
Number identification	0.147	0.056	0.162	0.010	0.149
Quantity discrimination (per minute)	0.004	0.000	0.006	0.020	0.031
Missing number	0.008	0.000	0.009	0.000	0.010
Addition 1 (per minute)	0.148	0.031	0.151	0.007	0.150
Addition 2 (per minute)	0.018	0.004	0.018	0.002	0.018
Subtraction 1 (per minute)	0.128	0.038	0.136	0.006	0.131
Subtraction 2 (per minute)	0.076	0.025	0.082	0.002	0.078
Multiplication	0.046	0.007	0.046	0.004	0.048
Fractions (%)	0.007	0.005	0.010	0.018	0.020

60. In summary, the quantitative data do not allow for a clear understanding of the mechanism by which EGRA Plus increased mathematics achievement. It appears that some sections, particularly quantity discrimination and fractions, were somewhat sensitive to the types of pedagogical improvements engendered by EGRA Plus. The rest of the sections improved when children could read more successfully, but it remains unclear how and why EGRA Plus increased the scores of the other sections.

61. Regardless of mechanism, the fact that EGRA Plus increased mathematics achievement, even moderately, is an important finding. Whether it was through increased reading skills or by improved pedagogy or accountability is unclear but also not necessarily relevant. What matters is that the program increased children's ability to learn new skills and teachers' ability to teach new subjects, even if those skills or topics were never explicitly addressed by the program. In other words, increased facility with reading helped children in other topics. This is a very exciting finding from the perspective that reading skills are foundational to other skill sets—that is, learning basic reading skills can transfer across subject area. Nonetheless, the fact remains that the impact from focused pedagogy, as made evident by the specific focus on reading in the full treatment schools, swamps any generalized effects or approaches.

62. Thus, while these small to moderate increases to mathematics skills might have been due to general pedagogical improvements, clearly the rest of the reading improvements were due to specific skill improvements among teachers; that is, it was not just a general improvement in teaching. It is, instead, evidence that teachers now know how to better teach particular and specific skills in reading. This skill improvement in teaching of reading is also an important finding: It is possible to use modest investments in pedagogical improvement to make trained and untrained teachers more capable pedagogically, with evidence in particular student outcomes. This is a remarkably different approach than much of the recent emphasis on what is normally taken as student-centered and/or learner-centered pedagogy in many reform or improvement projects (but is in most cases only a superficial application of these concepts). These programs often argue that increasing a teacher's general pedagogic skill set (in the areas of classroom management, student-centered pedagogy, etc.) will improve student outcomes across subject areas. That is likely true, to some extent, but EGRA Plus was effective because it taught teachers *particular skills* and other topics (such as learner assessment focused on those skills) and the application of those particular skills increased achievement, quite dramatically.

9. EGRA Impact Analysis

63. Impact studies take a variety of forms and use different strategies to assess the impact of a program on student outcomes. While using simple tabulations is useful (these tabulations be found in Annex C, regression models have a variety of benefits over the more simple comparison techniques, and these are presented in this section. For example, the tests inherent in the models allow for an estimate of whether or not an individual predictor (sex or grade, for example) has a statistically significant impact on a particular outcome.

64. In addition, the research design of this particular study lent itself to an analytic method called differences-in-differences analysis. This type of analysis falls into the category of causal analytic methods, which use statistical techniques to estimate the actual causal impact of a program of interest. This technique uses the longitudinal and the treatment-and-control aspects of a research design to determine two things: (1) whether there are differences between the scores of treatment students before and after the intervention, and (2) whether those differences are distinct from the differences for control students before and after the intervention. It is also possible to determine whether the effects of the interventions are smaller or larger at the midterm or final assessment.

65. Performing this type of analysis requires creating a combined data set with the baseline, midterm, and final assessments. Children are identified either as baseline or midterm and as treatment or control. In this case, the analysis was slightly more complicated because there were two treatment groups. However, using a system of dummy variables in the regression analysis, one can estimate the effect of being in the midterm assessment, being in the light treatment or full treatment group, and then, critically, being in a treatment group *and* in the midterm assessment.

66. Finally, post-hoc General Linear Hypothesis (GLH) tests can compare whether the impact of the two treatment groups was equivalent; or, to put it another way, whether the full treatment program worked better than the light treatment program. The models below have several parameters or variables, which are defined here.

- Midterm – represents a child in the midterm baseline, as distinguished from the baseline or final.
- Final – represents a child in the final data set.
- Light treatment – represents a child in the light treatment group.
- Light Treatment*Midterm – identifies children who were in both the midterm and light treatment groups.
- Light Treatment*Final – identifies children who were in both the final and light treatment groups.
- Full treatment – represents children in the full treatment group.
- Full Treatment*Midterm – identifies children who were in the midterm and full treatment groups.

- Full Treat*Final – represents children who were in the final and full treatment groups.
- Sex (girl) – shows the effect of being a girl, compared to boys.
- Grade (3) – shows the effect of grade 3, compared to grade 2.
- Control Group – in this design, a constant variable that is the average score of a boy in grade 2 in the control group at the baseline.

9.1 General Findings

67. This set of models shows that the full treatment program had a statistically significant impact on student achievement on all of the sections at both the midterm and final assessment. Light treatment had an impact on letter naming fluency, oral reading fluency, reading comprehension, and listening comprehension at the midterm, and on letter naming fluency at the final assessment. The models also show that there were no statistically significant sex differences except for familiar words and unfamiliar words (favoring girls), and grade 3 children outperforming grade 2, as one would expect.

9.1.1 Letter Naming Fluency

68. This model (see Table 9)[5] shows that both the full and light treatment programs had an effect on achievement in letter naming fluency, and at both midterm and the final assessment. Children in the control group scored 54.7 letters, with children at the midterm (rather than baseline) assessment scoring 10.5 letters higher, and children in the final assessment reading 21.7 letters higher. This shows a quite marked increase in letter reading among control schools, and the fact that the final assessment was so much higher than the midterm suggests that the secular trend was quite substantial. More research is necessary to determine the cause and whether experimental leakage contributed to it. The main effect of being in a full treatment group (regardless of baseline or midterm) was 2.0 letters more (full) and no difference for light treatment. Critically, the causal effect of being a child in a light treatment group was an additional 12.5 letters per minute at midterm and 6.0 letters at the final assessment. The effect of being a child in a full treatment group was 13.1 letters per minute at midterm and 14.8 letters at the final assessment. In other words, both treatment groups increased student achievement in letters, and at both the midterm and final assessment. The GLH tests (combined with the standardized coefficients) show that the midterm impacts were larger for light treatment (0.44 SD) than final impacts were (0.21 SD), but that there was no difference between midterm (0.46 SD) and final (0.52 SD) for full treatment.[6] The impact was larger at the midterm for full treatment than for light, and the same was true at the final

[5] Note that these analyses were performed using a differences-in-differences model using reg command in Stata. This allowed us to use the beta coefficients ooption, using the listcoef command. The outcomes are very similar whether xtreg is used (to account for the clustering in schools) or reg with a cluster option. Similarly, the findings are very similar when the sample is limited to schools (175) that were in each of the three rounds of data collection. The findings presented here, therefore, are very robust to model specification and sampling decisions.

[6] Using standardized coefficients, this regression analysis is able to determine the effect size of light and treatment groups at both midterm and final. This is found in the effect size (SD) column of Table 16.

assessment. The model does a reasonably good job of predicting achievement on letter naming fluency, since the R^2 is .26. If one notes that the grade impact was 12.1 letters per minute, then being in the full treatment group had an impact larger than the grade effect; namely, the full treatment "bumped children up" 1.2 grade levels in performance (assuming the grade 2 to grade 3 difference was linear).

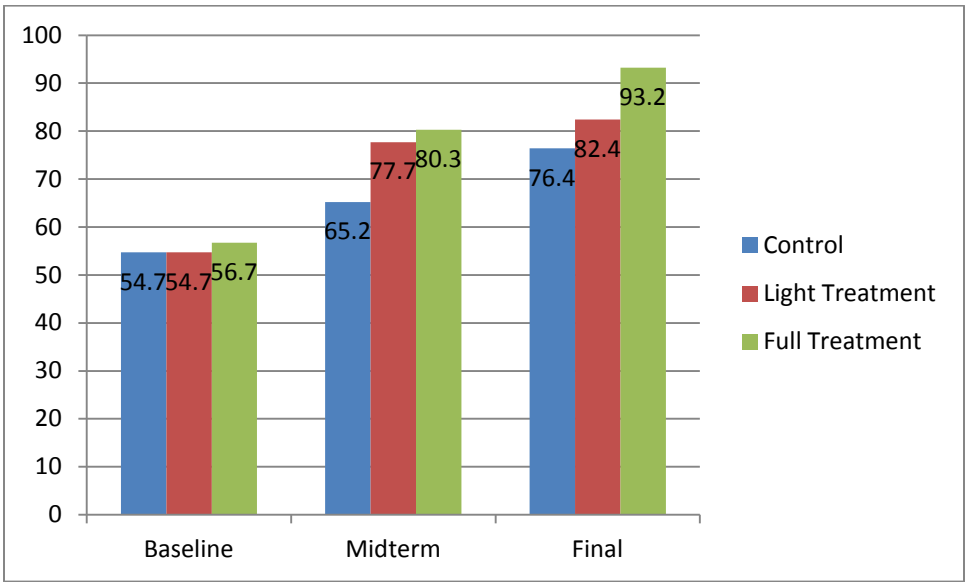
Table 9: Differences-in-Differences Regression Analysis for Letter Naming Fluency

Section	Predictor	Coef-ficient	Std. Error	*T*	Sig.	Effect Size (SD)	Obser-vations	*F*	Sig.	R^2
Letters naming fluency	Midterm	10.5	1.1	9.2	*<.001*					
	Final	21.7	1.2	17.8	*<.001*					
	Full Treatment	2.0	1.1	1.7	.08					
	Light Treatment	.0	1.1	0.0	.98					
	Full Treat*Mid	13.1	1.6	8.1	<.001	0.46				
	Full Treat*Final	14.8	1.7	8.7	<.001	0.52				
	Light Treat*Mid	12.5	1.6	7.9	<.001	0.44				
	Light Treat*Final	6.0	1.7	3.6	<.001	0.21				
	Grade (3)	12.1	0.5	22.1	*<.001*					
	Sex (Boy)	-0.1	0.5	-0.1	.93					
	Control Group	54.7	0.9	62.2	*<.001*		8096	287.85	<.001	.26

GLH Test (Full Treat*Mid - Full Treat*Final = 0): *F* 0.99, *p* value = .34. Therefore, there was no difference in the magnitude of the impact of full treatment between baseline and midterm and between midterm and final assessment.
GLH Test (LightTreat*Mid - Light Treat*Final = 0): *F* 14.58, *p* value <.001. Light treatment had a larger impact at the midterm than at the final.
GLH Test (Full Treat*Mid - Light Treat*Mid = 0): *F* .16, *p* value <.001. Full treatment had the same impact as light treatment at mid-term.
GLH Test (Full Treat*Final - Light Treat*Final = 0): *F* 29.24, *p* value <.001. Full treatment had a larger post-intervention impact than light treatment.

69. Figure 2 below shows the impact of the treatment groups graphically. Note that at baseline, children in full treatment schools read within 2 letters per minute of the baseline, and at both midterm and final assessments they were the highest scoring by a significant margin. The fact that the control scores were higher at the final assessment than at the midterm shows that there remained a secular trend of improvement that our analyses must account for (and do). At both midterm and final, the light treatment and full treatment programs increased children's letter naming fluency with the impact of full treatment slightly more than that of light treatment.

Figure 2: Histograms Comparing Impact of Light Treatment (red) and Full Treatment (green) Programs on Letter Naming Fluency



9.1.2 Phonemic Awareness

70. Table 10 below identifies the impact of the full and light treatments on student achievement in phonemic awareness. The main effects for midterm were that children identified 0.42 sounds more at midterm than at baseline, and an additional 0.91 sounds at the final. The midterm effect suggests a grade learning curve, but the final effect suggests a secular trend in improving phonemic awareness across the sample. In this section, the model shows that the light treatment group had modest impacts on phonemic awareness, 0.36 sounds at mid-term (*p* value <.01) and 0.44 sounds at the final (*p* value <.001). For the full treatment group, on the other hand, the program increased student achievement by 0.47 sounds at the midterm (*p* value <.01) and 1.47 sounds at the final assessment (*p* value <.001). The pattern is the same as for the letter naming fluency section, with no difference by gender (*p* value .58) and grade 3 more than grade 2 (0.79 sounds). The effect sizes for full treatment were small at the midterm (0.18 SD) and moderately large at the final assessment (0.55 SD), and the entire model explains 10% of the variation in phonemic awareness. Note that being in the full treatment group meant an effect of 2.9 times the grade effect; the project "bumped up" the children nearly 2 grades (assuming the grade 2 to grade 3 difference was linear).

Table 10: Differences-in-Differences Regression Analysis for Phonemic Awareness

Section	Predictor	Coef-ficient	Std. Error	*T*	Sig.	Effect Size (SD)	Obser-vations	*F*	Sig.	R^2
Phonemic awareness	Midterm	0.42	.12	3.63	*<.001*					
	Final	0.91	0.12	7.46	*<.001*					
	Full Treatment	0.16	0.12	1.39	.17					
	Light Treatment	0.10	0.11	0.86	.39					
	Full Treat * Mid	0.47	0.17	2.83	<.01	0.18				
	Full Treat*Final	1.47	0.17	8.64	<.001	0.55				
	Light Treat * Mid	0.36	0.16	2.24	.03	0.14				
	Light Treat*Final	0.44	0.17	2.64	<.01	0.17				
	Grade (3)	0.79	0.06	14.34	*<.001*					
	Sex (Boy)	-0.03	0.06	-0.56	.58					
	Control Group	3.04	0.09	33.69	*<.001*		8351	96.64	<.001	.10

GLH Test (Full Treat*Mid - Full Treat*Final = 0): *F* 34.14, *p* value <.001. The impact of full treatment was larger at final than at midterm.
GLH Test (Light Treat*Mid – Light Treat*Final = 0): *F* 0.21, *p* value .65. There is no difference in the impact of light treatment between mid-term and final assessment.
GLH Test (Full Treat*Mid - Light Treat*Mid = 0): *F* 0.42, *p* value .51. There is no difference in the impact of full and light treatment at mid-term.
GLH Test (Full Treat*Final - Light Treat*Final = 0): *F* 39.13, *p* value <.001. Full treatment had a larger impact at final than did light treatment.

9.1.3 Familiar Word Fluency

71. For familiar words, the main effects at both midterm and final (Table 11) were that all children in the entire sample increased their fluency at midterm (4.8 wpm) and at final (10.2 wpm). Girls outperformed boys by 0.7 wpm and grade 3 children read better than grade 2 (7.9 wpm). The differences-in-differences analysis shows that light treatment had no statistically significant impact on achievement at either midterm (*p* value .18) or final (*p* value .81). Full treatment schools did not increase achievement at the midterm (*p* value .58), but increased by 14.3 words per minute at the final assessment (*p* value <.001). The effect size for full treatment at the final assessment was 0.78 SD. The R^2 for the final model was .21, which is larger than for phonemic awareness. The project "bumped up" the children by 1.8 school years in familiar word fluency.

Table 11: Differences-in-Differences Regression Analysis for Familiar Word Fluency

Section	Predictor	Coef-ficient	Std. Error	*T*	Sig.	Effect Size (SD)	Obser-vations	*F*	Sig.	R^2
Familiar word fluency	Midterm	4.8	0.8	6.3	*<.001*					
	Final	10.2	0.8	12.4	*<.001*					
	Full Treatment	1.6	0.8	2.1	*.03*					
	Light Treatment	0.9	0.7	1.2	*.24*					
	Full Treat * Mid	0.6	1.1	0.6	.58	0.03				
	Full Treat*Final	14.3	1.1	12.6	<.001	0.78				
	Light Treat * Mid	1.4	1.1	1.3	.18	0.08				
	Light Treat*Final	0.3	1.1	0.3	.81	0.01				
	Grade (3)	7.9	0.4	21.5	<.001					
	Sex (Boy)	-0.7	0.4	-1.8	.07					
	Control Group	5.0	0.6	8.6	*<.001*		8022	214.54	<.001	.21

GLH Test (Full Treat*Mid - Full Treat*Final = 0): *F* 142.44, *p* value <.001. Therefore, the magnitude of the impact of full treatment between midterm and final assessment was larger than between baseline and midterm.
GLH Test (Full Treat*Final - Light Treat*Final = 0) F 167.14, p-value <.001. The impact of full treatment was larger than the impact of light treatment at final.

9.1.4 Unfamiliar Word Fluency

72. Table 12 presents the relationships between the predictors and unfamiliar word fluency. It shows that there was no difference between baseline and midterm on this variable, and children in the final assessment read 0.9 words per minute more than those at the baseline. The analysis shows that full treatment increased unfamiliar words read per minute by 0.4 words at midterm (*p* value .07) and 11.2 words at the final assessment (*p* value <.001). Effect sizes were 0.11 SD and 1.23 SD, respectively. For light treatment, there was no impact at midterm (*p* value .54) or final (*p* value .92). The entire model has an R^2 of .17, a bit less than for familiar word fluency. GLH testing shows that the full treatment program had a larger impact at final than at midterm. The program impact was a massive eight times larger than the impact of a year's worth of schooling (11.2 over 1.4). Since this impact was so huge, one hesitates to say how many grades it is equivalent to, since it is risky to say that the grade effect would be linear or nearly linear over such a large gain.

Table 12: Differences-in-Differences Regression Analysis for Unfamiliar Word Fluency

Section	Predictor	Coef-ficient	Std. Error	*T*	Sig.	Effect Size (SD)	Obser-vations	*F*	Sig.	R^2
Unfamiliar word fluency	Midterm	-0.3	0.4	-0.7	*.49*					
	Final	0.9	0.4	2.2	*.03*					
	Full Treatment	0.6	0.4	1.7	.10					
	Light Treatment	0.4	0.4	1.2	.25					
	Full Treat * Mid	0.4	0.5	1.8	.07	0.11				
	Full Treat*Final	11.2	0.6	19.5	<.001	1.23				
	Light Treat * Mid	0.3	0.5	0.6	.54	0.04				
	Light Treat*Final	0.1	0.6	0.1	.92	0.01				
	Grade (3)	1.4	0.2	7.8	*<.001*					
	Sex (Boy)	-0.5	0.2	-2.8	<.01					
	Control Group	1.5	0.3	5.0	*<.001*		8057	169.40	<.001	.17

GLH Test (Full Treat*Mid - Full Treat*Final = 0): *F* 308.39, *p* value <.001. Therefore, the magnitude of the impact of full treatment between mid and final assessment is larger than between baseline and midterm.
GLH Test (Full Treat*Mid - Light Treat*Mid = 0): *F* 1.48, *p* value .22. There is no difference in the impact of full and light treatment at mid-term.
GLH Test (Full Treat*Final - Light Treat*Final = 0): *F* 410.07, *p* value <.001. The impact of full treatment is larger than the impact of light treatment at final.

9.1.5 Oral Reading Fluency

73. The differences-in-differences analysis for oral reading fluency (Table 13) shows that there was, once again, a main effect for the midterm (3.4 words per minute) and final assessments (7.1 words per minute). There was no difference by sex (*p* value .94), and grade 3 children read 11.3 words per minute more than grade 2 children. The model shows that the light treatment program increased oral reading fluency by 3.9 words per minute (0.15 SD) at the midterm, but it had no impact at the final assessment (*p* value .52). The full treatment had a small effect at the midterm (5.0 words per minute, 0.19 SD) and a large effect at the final assessment (21.1 words per minute, 0.80 SD). These are impressive results, particularly at the final assessment. The post-hoc GLH test shows that the impact of full treatment was bigger at the final assessment than at the midterm. The tests also show that the full treatment had a larger impact than light treatment at the midterm. The model had an R^2 of .16. As with the other EGRA sections, being in full treatment was equivalent to roughly two years of schooling, "bumping up" the children about two grade-equivalents in reading fluency.

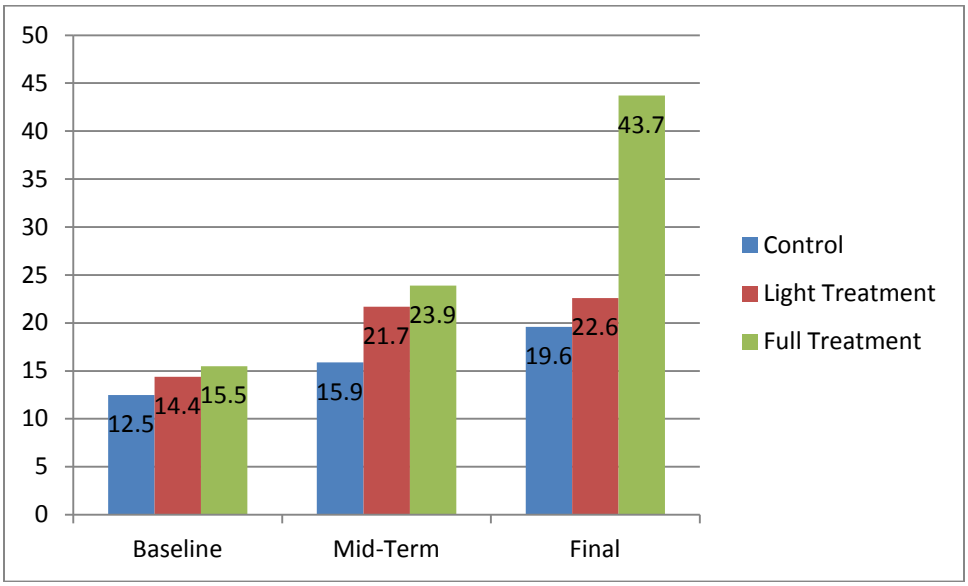
Table 13: Differences-in-Differences Regression Analysis for Oral Reading Fluency

Section	Predictor	Coef-ficient	Std. Error	*T*	Sig.	Effect Size (SD)	Obser-vations	*F*	Sig.	R^2
Oral reading fluency	Midterm	3.4	1.1	3.0	*<.01*					
	Final	7.1	1.2	5.8	*<.001*					
	Full Treatment	3.0	1.1	2.7	<.01					
	Light Treatment	1.9	1.1	1.8	.08					
	Full Treat * Mid	5.0	1.6	3.1	<.01	0.19				
	Full Treat*Final	21.1	1.7	12.4	<.001	0.80				
	Light Treat * Mid	3.9	1.6	2.5	.01	0.15				
	Light Treat*Final	1.1	1.7	0.7	.52	0.04				
	Grade (3)	11.3	0.5	20.7	*<.001*					
	Sex (Boy)	-0.1	0.5	-0.1	.91					
	Control Group	12.5	0.9	14.3	*<.001*		7867	144.59	<.001	.16

GLH Test (Full Treat*Mid - Full Treat*Final = 0): *F* 86.83, *p* value <.001. Therefore, the impact of full treatment at final was larger than at midterm.
GLH Test (Light Treat*Mid - Light Treat*Final = 0): *F* 2.75, *p* value .10. The impact of light treatment was larger at the mid-term than at the post assessment.

74. Figure 3 below shows graphically the impact of full and light treatment on oral reading fluency. When we examine the midterm scores, first it is clear that both light treatment (red bars) and full treatment (green bars) increased oral reading fluency by a significant margin when compared to control (blue bars). When we compare the final assessment, the light treatment had a non-significant impact on oral reading fluency. This suggests that the secular trend increases on oral reading fluency were significant. The full treatment program had an enormous impact on oral reading fluency, causing scores that were more than twice as high as those for control and nearly twice as high as light treatment.

Figure 3: Bar Chart Showing the Impact of Full (green) and Light (red) Treatment on Oral Reading Fluency



9.1.6 Reading Comprehension

75. For the reading comprehension sections, the main effects for midterm and final assessment (Table 14) show that children performed 1.4% worse on the midterm than at the baseline (though insignificant statistically) and better by 7.9% at the final (*p* value <.001). There were no differences by sex (*p* value .95), and children in grade 3 understood better by 12.3% than grade 2 children (*p* value <.001). The full treatment model increased comprehension by 4.7% at the midterm (0.15 SD) and 25.2% at the final assessment (0.82 SD). Light treatment had no impact at midterm (*p* value .34) or at final assessment (*p* value .74). The GLH tests show that full treatment had a larger impact at final than at midterm. Similar to the oral reading fluency model, the R^2 for the reading comprehension was .15. A child who was in the EGRA Plus program benefited from more than two years of typical grade progression in oral reading fluency.

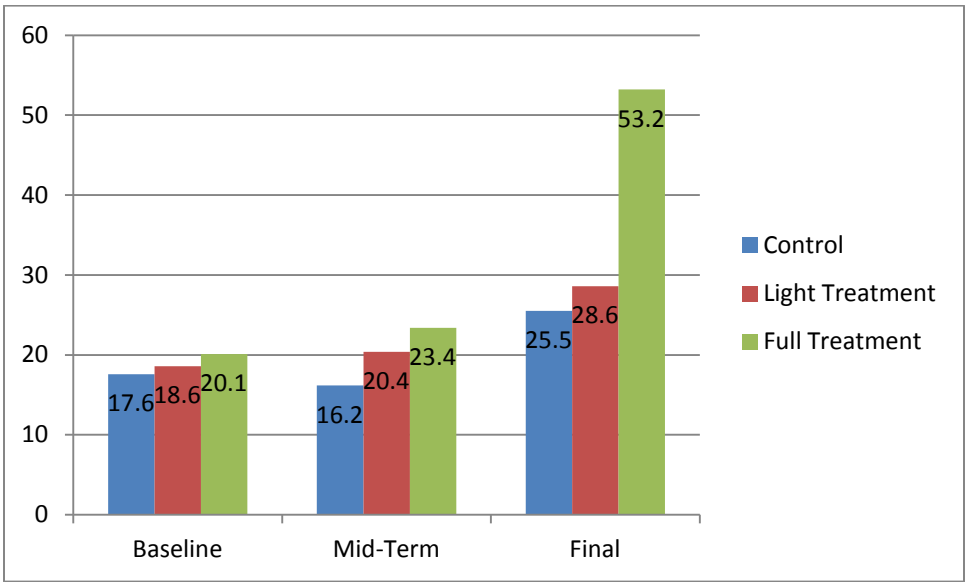
Table 14: Differences-in-Differences Regression Analysis for Reading Comprehension

Section	Predictor	Coef-ficient	Std. Error	*T*	Sig.	Effect Size (SD)	Obser-vations	*F*	Sig.	R^2
Reading Comprehension	Midterm	-1.4	1.3	-1.1	*.29*					
	Final	7.9	1.4	5.5	*<.001*					
	Full Treatment	2.5	1.3	1.9	.06					
	Light Treatment	2.4	1.3	1.9	.06					
	Full Treat * Mid	4.7	1.9	2.5	.01	0.15				
	Full Treat*Final	25.2	2.0	12.7	<.001	0.82				
	Light Treat * Mid	1.8	1.8	1.0	.34	0.06				
	Light Treat*Final	0.7	2.0	0.3	.74	0.02				
	Grade (3)	12.3	0.6	19.2	*<.001*					
	Sex (Boy)	-0.0	0.6	-0.1	.95					
	Control Group	17.6	10	17.2	*<.001*		7867	142.52	<.001	.15

GLH Test (Full Treat*Mid - Full Treat*Final = 0): *F* 103.20, *p* value <.001. Therefore, the magnitude of the impact of full treatment between mid and final assessment is larger than between baseline and midterm.

76. Figure 4 below investigates the impact of full and light treatment on reading comprehension. When we examine the midterm scores, first it is clear that both light treatment and full treatment increased reading comprehension by a modest amount (larger for full treatment). When we compare the final assessment, the light treatment had a non-significant impact on reading comprehension. This suggests that the secular trend increases on reading were significant, just as the trend was for oral reading fluency. The full treatment program had an enormous impact on reading comprehension, causing scores that were more than twice as high as those for control and nearly twice as high as light treatment.

Figure 4: Bar Chart Showing the Impact of Full (green) and Light (red) Treatment on Oral Reading Fluency



9.1.7 Listening Comprehension

77. Finally, for listening comprehension (Table 15), the model shows that full treatment increased scores by 9.8% at midterm and 13.1% at the final assessment, and light treatment increased the scores at midterm by 8.1% and had no effect at the final assessment. The model explains a large percentage of the variation, with an R^2 of .38.

Table 15: Differences-in-Differences Regression Analysis for Listening Comprehension

Section	Predictor	Coef-ficient	Std. Error	*T*	Sig.	Effect Size (SD)	Obser-vations	*F*	Sig.	R^2
Listening comprehension	Midterm	35.1	1.2	28.5	<.*001*					
	Final	36.5	1.3	28.2	<.*001*					
	Full Treatment	1.0	1.2	0.8	.42					
	Light Treatment	1.9	1.2	1.6	.11					
	Full Treat * Mid	9.8	1.7	5.6	<.001	0.29				
	Full Treat*Final	13.1	1.8	7.3	<.001	0.39				
	Light Treat * Mid	8.1	1.7	4.8	<.001	0.24				
	Light Treat*Final	0.8	1.8	0.4	.66	0.02				
	Grade (3)	7.4	0.6	12.6	<.*001*					
	Sex (Boy)	-0.0	0.6	-0.1	.93					
	Control Group	29.2	0.9	30.9	<.*001*		8215	501.23	<.001	.38

GLH Test (Full Treat*Mid - Full Treat*Final = 0): *F* 3.43, *p* value .06. Therefore, there was no difference in the magnitude of the impact between baseline and midterm and between midterm and final assessment at the .05 level.
GLH Test (Full Treat*Mid - Light Treat*Mid = 0): *F* .94, *p* value .33. Therefore, there was no difference in the impact of full and light treatment at midterm.

78. This section of the report shows quite clearly that EGRA Plus: Liberia had a remarkably large impact on student achievement, particularly for the full treatment group. This impact was large enough to overcome the secular trend identified at the midterm (probably the grade learning effect) and the larger trend at the final assessment (which will require more research to fully understand). These impacts were consistently large, nearing one standard deviation for many of the critical areas. As explained earlier, note that the design of the differences-in-differences models allows for an investigation of the effect size of the program's impact as measured by final (or midterm) against baseline, and removes the gains in the control groups. The results are quite similar to what was identified by the simpler Cohen's *d* effect size analysis presented above.

9.2 Interacting EGRA Plus with Sex, Age, and Grade

79. In order to determine whether the sex, age, or grade of the children had a differential effect on student outcomes, we fit additional multiple regression models.[7] First, models were fit to determine whether there was a main effect for age when we controlled for grade. This would answer the question of whether the grade effect would differ for children who were at different ages. Accounting for age is particularly important for a country like Liberia, which has a significant portion of the student population entering school late, due to unrest; or having delays in their schooling, due to the civil war.

[7] The models are not presented here due to space constraints.

80. We tested this in four ways. First, we used the child's absolute age as a predictor. These models show that, controlling for grade, older children scored statistically significantly lower on all sections assessed except letter naming fluency, phonemic awareness and unfamiliar word fluency. This was less than ideal, since the regression model did not manage the wide variation in ages well (ages 5 through 27). Second, we created a variable that converted the child's age to age in relation to the expected age at that particular grade. That is, we used a variable that was a 1 for a child who was 10 in grade 2 (the expected age was 8 or 9), for example. The fits for these models were better than those using the absolute age. The findings were similar: Every year older than the expected age was statistically significantly negatively correlated with every section except letter naming fluency and unfamiliar word fluency. Third, we created a dummy variable that combined all of the children who were overage for their grade into one group, and compared those to students who were at the expected age or below. This was our preferred specification since there was no reason to think that there should be a substantive difference between a learner who was 20 and one who was 25, for example. These models show that overage children actually were more fluent with letter reading, by 1.8 letters per minute (*p* value .04). They read 1.7 fewer familiar words per minute (*p* value <.01), read aloud 2.7 fewer words per minute of connected text (*p* value <.01), and scored 2.1% lower on reading comprehension (*p* value .03). There was no relationship for phonemic awareness or listening comprehension.[8] The fourth and final way we assessed the relationship between age and reading outcomes was to examine whether EGRA Plus: Liberia had a differential effect for overage and non-overage children. There were not many statistically significant relationships, which shows that the program was equally effective across ages.[9] Note that all of these models control for grade, as well. This shows that within a grade, or classroom, children who were overage know the alphabet better, but perform less well on the other tasks. The EGRA Plus program did not discriminate with respect to its impact on student achievement.

81. The models presented above showed little sex differentiation as a main effect. Boys did worse than girls on familiar and unfamiliar word fluency. Another issue is relevant, of course: whether EGRA Plus had a differential effect for boys and girls. Recall that boys did worse at the baseline on many assessments. We found that the program did have a differential effect by sex for a few sections. Girls benefited more on letter fluency at mid-term in light schools by 5.5 letters per minute. Boys benefited more in full treatment schools at the final assessment in phonemic awareness, increasing their scores by 1.9 words rather than 1.0 words for girls (*p* value <.05). For all of the midterm and the rest of the final assessment sections, and for all of the light treatment effects, there were no differences in

[8] We also fit models that compared children who were underage against the rest of the sample. The relationships were insignificant, except that underage children scored lower on listening comprehension. This makes sense, since younger children would have had less exposure to spoken language.

[9] Two models did have statistically significant interactions between program effects and overage children. Specifically, at midterm, EGRA Plus increased the scores of overage children by 13.5 more letters per minute. At the final assessment, EGRA Plus increased the scores of overage children by 3.0 unfamiliar words per minute less.

program impact by sex. This sex differential likely was related to the underachievement of boys at the baseline, but does merit further analysis.

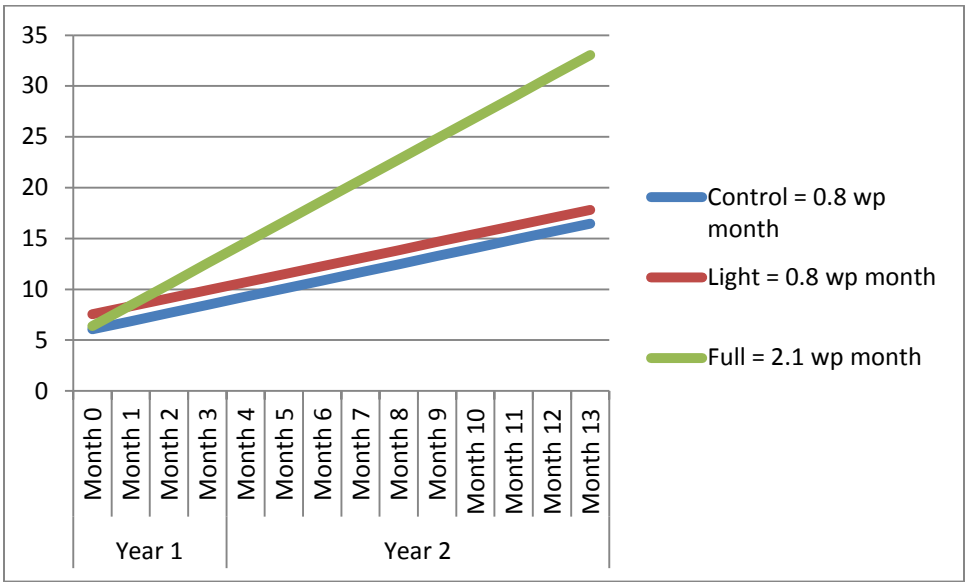
82. We also fit several models to determine whether EGRA Plus increased the reading of grade 2 or grade 3 children more. We found no differences in the EGRA Plus outcomes for either full or light treatment, at either midterm or final assessment. Only one model had a statistically significant difference. EGRA Plus light treatment increased the letter naming fluency by 4.7 more letters per minute for Grade 3 children than Grade 2. For the rest of the models, however, there were no statistically significant differences between the impact of the program by grade.

9.3 Learning Rate Increases

83. We felt it would be interesting to determine not only the absolute impact of the program, but also the learning trend over the duration of the program. Therefore, we fit causal models that investigated the month-by-month learning gains by treatment group (control, light treatment, and full treatment) over the life of the program.[10] Figure 5 below presents the monthly slope of learning gains for familiar words. The control schools increased familiar word fluency by an estimated 0.8 words per month, light treatment schools increased word fluency by 0.8 words per month, and full treatment increased outcomes by 2.1 words per month. This means, therefore, that the learning rate for full treatment schools was 2.6 times faster than that of children in control schools, confirming the points made above regarding program impact as compared to average gain between grades. While full treatment schools started at fluency rates below that of light and control schools, the final assessment scores were significantly higher.

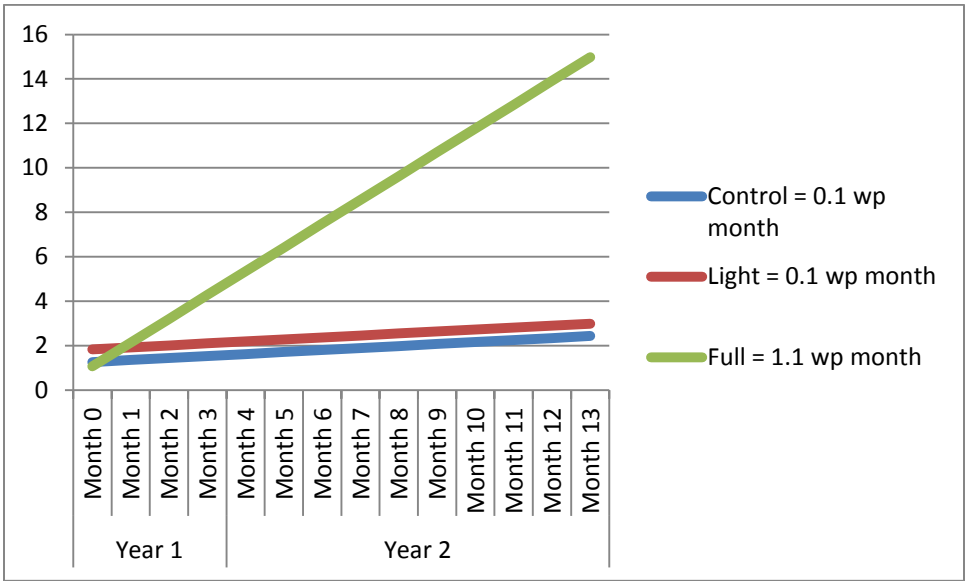
[10] These models were fit by giving the baseline data a value of 0, the midterm data a value of 3, and the final data a value of 13. This equates to the number of months that the program "taught" children. This analysis makes an assumption of linear monthly gains, however, which is likely not true. Moreover, the analysis ignores the summer reading loss that has been shown in a great deal of reading acquisition literature. It is useful, however, as a visual to estimate the effect of the treatment programs against the control schools. To simplify the figures, the grade effect is controlled for, as is gender. The main effects of midterm and final are also controlled for.

Figure 5: Learning Rates for Familiar Words Comparing Control, Light, and Full Treatment Schools Over the Two Years of EGRA Plus



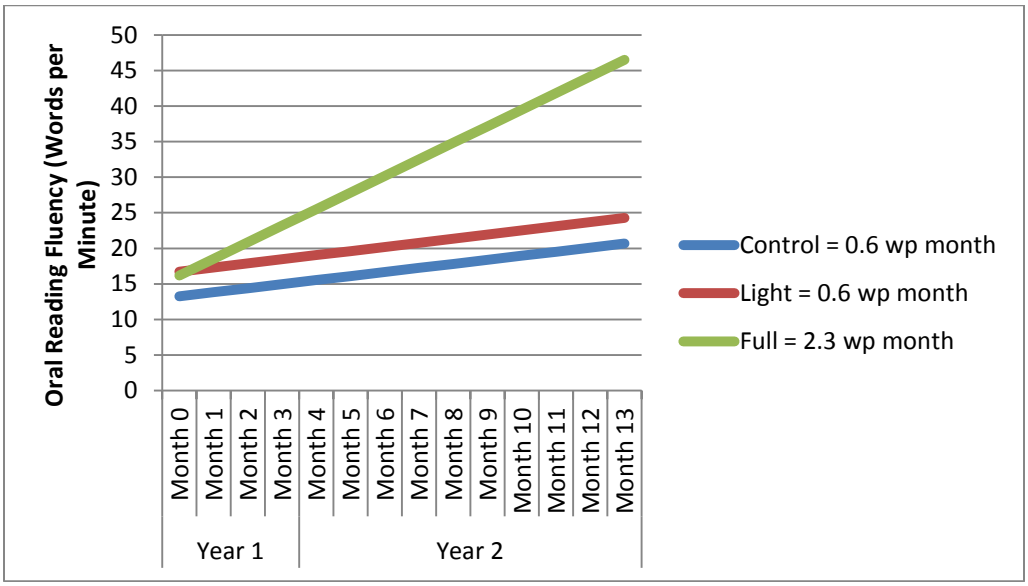
84. Figure 6 below presents the learning rates by month for unfamiliar word fluency, by treatment groups. It shows that the learning gains were very shallow for both light and control schools, with children in those schools gaining almost no fluency with decoding of new words. For full treatment schools, on the other hand, the learning rates were 1.1 words per month. While modest in absolute terms, this represents a rate 11.9 times faster for full treatment children than for those in control schools.

Figure 6: Learning Rates for Unfamiliar Words Comparing Control, Light, and Full Treatment Schools Over the Two Years of EGRA Plus



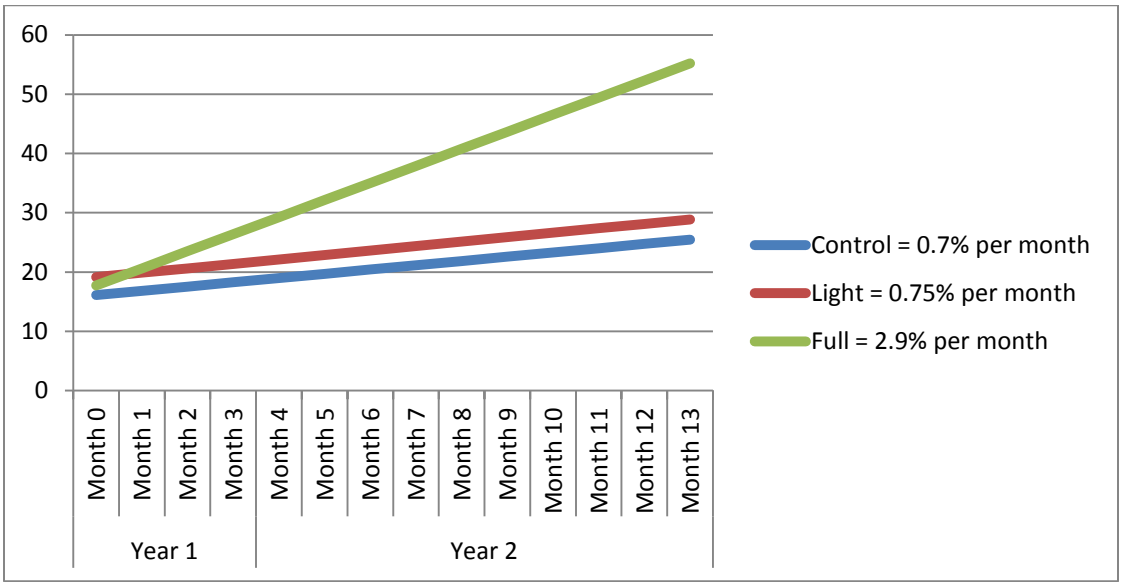
85. Figure 7 below presents the learning rates by month for oral reading fluency by treatment groups. While more steep than the slopes identified in the unfamiliar word analysis, the impact of the program on learning rates was still quite significant, since the word per minute learning rates for control (0.6 words per minute) and light schools (0.6 words per minute) were much slower than those for full treatment. This means that children learned to read 4.1 times faster in full treatment than control schools.

Figure 7: Learning Rates for Oral Reading Fluency Comparing Control, Light, and Full Treatment Schools Over the Two Years of EGRA Plus



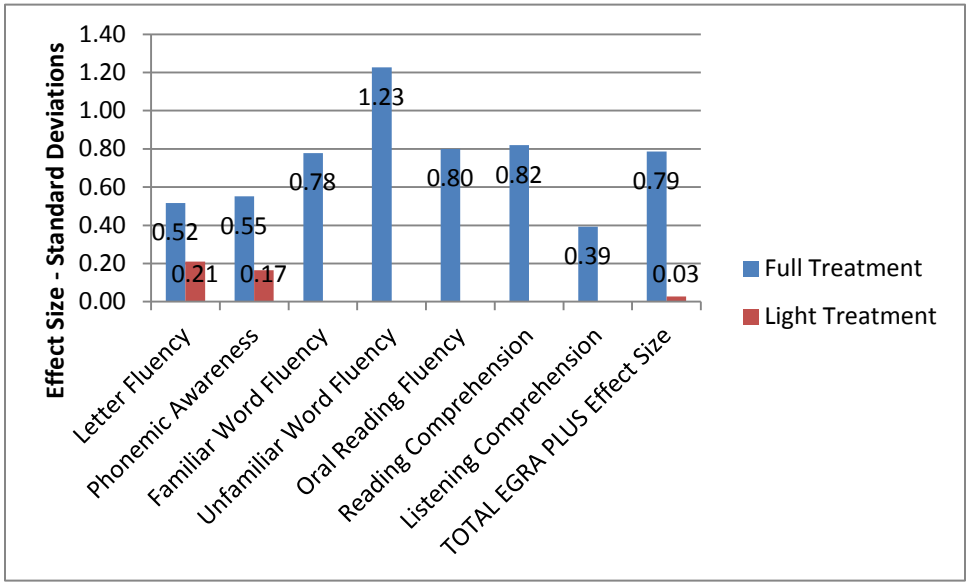
86. In order to compare the learning rates for reading comprehension across treatment groups, we analyzed the data to determine the learning rates by year, in Figure 8. It shows that children in control and light treatment schools increased their comprehension scores by 0.7% and 0.75% per month, respectively, while full treatment increased by 2.9% per month. This shows that children in full treatment schools were learning at a rate of four times more per month than their counterparts in control schools.

Figure 8: Learning Rates for Reading Comprehension Comparing Control, Light, and Full Treatment Schools Over the Two Years of EGRA Plus



87. In Figure 9 below, the effect sizes for each section are presented by treatment group. Recall that these would be much higher if a basic effect size calculation were performed, since those effect sizes do not remove the impacts from the control schools. These much more conservative estimates are remarkable because of their magnitude. Overall, using a conservative estimate of effect size, the overall full treatment effect size is 0.79 SD.[11] Light treatment had a negligible impact on achievement (0.03 SD). This appears to have been because the scores at the final assessment were significantly higher in the control schools, for a reason that requires further research.[12]

Figure 9: Effect Sizes by Full and Light Treatment and by EGRA Sections



[11] This effect size weighting procedure was devised by Dr. Luis Crouch and Dr. Marcia Davidson. Letter fluency was 5%, phonemic awareness was 10%, familiar words was 15%, oral reading fluency was 50%, reading comprehension was 25%, and listening comprehension was 10% of the total effect.

[12] Note that it is possible that treatment leakage occurred, and was responsible for the large increases in baseline schools. On the other hand, it is possible that other shifts occurred in the Liberian education sector during the period of EGRA Plus. More research is necessary to examine this more in depth.

10. The Further Research

88. The very large effect sizes experienced with EGRA Plus: Liberia suggest the need for further research to better understand the impact of the program on student achievement. Specifically, we suggest the following.

- **Examine more closely the change mechanisms at work in EGRA Plus.** The mechanisms that were responsible for the large impact sizes identified in this program warrant further investigation. That is to say, the EGRA Plus program was so successful that other programs and countries, and scale-up within Liberia itself, would benefit from investigating the reasons for the success of the project. Section 3.3 above presented a detailed discussion of causal influences, but post-hoc qualitative research is necessary to more adequately explain what happened to make the program quite so successful.
- **Understand the increases in reading outcomes in non-EGRA Plus schools.** EGRA Plus: Liberia showed that the control schools had significant gains both between the baseline and midterm assessments and between the midterm and final assessments. The relationship between baseline and midterm is easily explained as the learning effect of a grade, since baseline was in November 2008 and midterm was in June 2009. The significant increases between June 2009 and June 2010 for control schools are much more difficult to explain, since the learning effect is not the reason. Further research is necessary to determine whether this midterm-to-final-assessment effect was related to EGRA Plus (via some form of leakage) or whether it was due to changes in the literacy efforts in Liberia. (Note that even if unintended leakage to the control schools were revealed, this would be in itself an important finding.)
- **Understand the sex gap in the program effects.** A consistent pattern was identified in the results: While EGRA Plus increased reading outcomes for children across all the measures, effect sizes were larger for girls than for boys in most of the EGRA sections. This is partially because initial scores were lower for girls than for boys, yet it is not clear how the gender dynamic was mitigated by EGRA Plus, or whether it created achievement differences in the opposite direction.
- **Examine the relationship between math and improved reading.** One of the unexpected effects of EGRA Plus was the manner in which it improved mathematics outcomes for children. While EGRA Plus had no specific intervention in mathematics, the treatment program increased outcomes in mathematics by small to moderate amounts, with particularly sizeable gains in the number-sense portions of early mathematics achievement, namely number identification, quantity discrimination, and fractions. Modest gains were identified in addition, subtraction, and multiplication, as well. The mechanism by which this increase occurred is unclear, so further research is necessary to determine whether EGRA Plus increased mathematics scores by helping children read better (allowing for deeper understanding of the mathematics assessment), or whether general improvements to

pedagogical quality engendered by EGRA Plus transferred from reading to mathematics. The size of the effects means that the reason for the relationships needs further study and clarification.

- **Examine the cost-effectiveness of EGRA Plus.** The analyses presented in this paper allow for an understanding of whether the EGRA Plus program worked. However, it is less obvious how cost effective EGRA Plus was. Finding out will require deeper analysis of the inputs from the program. Our analysis shows that, given that EGRA Plus had the approximate effect of an additional two years of reading, the cost-effectiveness question is quite stark: What is the value of two years of schooling?

11. Recommendations

89. This final assessment report takes stock of the effectiveness of the EGRA Plus program. The discussion of the findings explains how EGRA Plus worked, and suggests several interventions and strategies that might be undertaken to sustain and replicate the findings.

- **Scale up EGRA Plus: Liberia.** The EGRA Plus program was remarkably effective. While control schools increased their reading outcomes over baseline by a significant amount at midterm (due to the grade learning effect) and at the final assessment (due to other improvements in the education sector or to program leakage), the program increased reading outcomes by nearly 1 standard deviation. This is a large effect size and is convincing evidence that the package of interventions in EGRA Plus should be replicated and expanded. The Liberia Teacher Training Program second phase (LTTP2) program could serve as an incubator for further interventions, and for an examination of whether the initial, and remarkable, increases from EGRA Plus can be replicated at scale. USAID agreed that beginning in January 2011, under LTTP2, the EGRA Plus program will be extended to all 180 schools, including control and light intervention schools. Beyond LTTP2, however, it appears that the rest of Liberia's children are likely to benefit greatly from this project. As a result, and given that the lesson plans and systems outcomes are already prepared, the government of Liberia should seriously consider whether the strategy could be scaled up to the rest of the country, resources allowing.
- **Move past focus on letters and words and focus on reading comprehension.** The gains on all of the EGRA outcomes were substantial and reading comprehension scores increased by nearly 1 standard deviation. That said, the reading comprehension scores, even at the full assessment, did not reach the expected level of proficiency. The full treatment children's ability to comprehend was highly correlated with their increased skills in oral reading fluency. However, the effect was not as large as it would have been if more emphasis had been placed on encouraging and developing children's metacognitive skills, including their ability to predict, categorize, and analyze events and situations in written text. This is evident given the gap in achievement between listening comprehension and reading comprehension. In other words, children could understand much more of what they heard than what they read. This shows that the children have the oral vocabulary to understand more of what they read. These skills must be explicitly taught and modeled.
- **Task the Liberian Ministry of Education with developing country-level benchmarks for reading.** Our research provides examples of benchmarks—that is, using the 90th percentile of reading scores as a benchmark. That measure was arbitrarily chosen by a non-Liberian evaluator, and was picked without an evaluation of the appropriate skills that each level of child will achieve based on the curriculum. Such a benchmark development process would help to streamline reading

intervention efforts, and allow for within-country, rather than cross-country, comparisons.

- **Target reading pedagogical techniques in teacher professional development.** The findings showed that Liberian teachers were sensitive to the intervention in this program. This suggests that with targeted efforts, and with the use of achievement data at the classroom and school level, teachers can improve how they teach children to read. We recommend that this finding be exploited in the Liberian Ministry of Education's efforts to train teachers at the pre-service and in-service levels. In other words, the targeted efforts used in a small project such as EGRA Plus should be replicated in in-service teacher professional development and adapted to the pre-service professional development.
- **Place considerably more emphasis on within-grade achievement.** While comparisons to international benchmarks are not ideal, Liberian children's progress within a grade was too modest to allow children to achieve reading fluency by grade 4 when most instruction is provided under the assumption that children can already read. If the grade 2 (beginning to end) gain in oral reading fluency is only 4 words on average, and grade 3 gains are nearly 2.5 words in control or standard Liberian schools (but 10 words per minute in full treatment schools), then children are not getting enough within a grade to be able to lessen the gaps between themselves and children elsewhere, even within sub-Saharan Africa.
- **Improve the achievement of girls in Liberian reading.** The baseline data showed that boys outperformed girls across the EGRA sections. This is dissimilar from the gender relationships identified in most other sub-Saharan African countries with EGRA studies. Under EGRA Plus, on the other hand, girls outperformed boys in many of the sections at the *final* assessment. What this shows is that girls can perform quite well under the right instructional conditions. This finding should influence how teachers teach girls. With the perspective that girls can achieve quite well if taught properly, then head teachers, communities, and higher education officials can and should demand high achievement for girls in the classrooms under their jurisdiction.
- **Move beyond community knowledge of reading achievement to teach the more complex aspects of reading.** The light treatment impacts on children showed that simply intensifying the community's focus on reading outcomes improved student outcomes. This was particularly the case in letter naming fluency. However, for the more technical aspects of reading that depend on decoding and comprehension strategies—such as reading comprehension, oral reading fluency, and unfamiliar word fluency—teachers need professional development to learn techniques and strategies for imparting these areas of expertise to children. In full treatment schools, relatively modest investments in teacher training paid large dividends. In other words, attention and focus on reading and increased accountability, by both teachers and communities, are powerful but insufficient; training and skills are also necessary. As

much of the worldwide literature shows, both accountability and support are key. One without the other is not as useful.

- **Underscore decoding skills as a critical step for improved reading outcomes.** The largest impacts of the EGRA Plus program were on children's ability to decode new words. These newfound skills in decoding new words were the jump start that children needed to improve their ability to read texts, and then to increase reading comprehension. Schools of teacher education and in-service programs should increase their focus on these decoding skills, since they seem to be a critical stepping-stone for improved outcomes in more complex reading tasks.
- **Use reading improvements to increase learning in other subjects.** The findings showed that a reading intervention can also have knock-on effects in other subjects, in this case mathematics. This suggests that while Liberia's Ministry of Education is rightly concerned about achievement levels across subjects, reading is an entry point to improving reading outcomes, as well as outcomes in other subjects. We did not study whether reading improvements also were responsible for increases in achievement in other subjects, but given the outcomes identified in mathematics, it is plausible that such a relationship exists. Therefore, we recommend that Liberia focus its human and financial resources on improving the quality of reading in Liberia's children, and then see whether and how these investments can affect what happens in other subjects, while at the same time using the techniques in the other subjects. It will certainly not be sufficient, but reading is a more appropriate initial place for pedagogical investment, since the improvements in this subject might have additional outcome improvements elsewhere and demonstrate that the combination of focus, subject pedagogy, and management gets results.
- **Expand the use of scripted programs for teacher professional development.** The experience of EGRA Plus makes quite clear that the use of scripted programs for teacher professional development can have significant impacts on reading outcomes. While some resistance was noted, in that some teachers did not want to do "extra" work, on the whole the teachers accepted the new methods. Moreover, increasing the scriptedness of the lesson plans increased the effectiveness between the midterm and final assessment. Both factors were quite revealing. It appears that these types of training methods have a high and significant likelihood of continuing to be effective in Liberia.

Appendix A: Calibration of Baseline, Midterm, and Final Assessments

90. This appendix offers more detail about the process by which we calibrated the versions of the EGRA instrument that were used at the three different time points.

91. In order to prevent teaching to the test, or memorization, the midterm and final assessments used different word lists and passages. While efforts were made to ensure that the levels of the stories and words were similar, using Spache analysis, this is often not sufficient to ensure calibration. Thus, in addition to the ex ante calibration, we made an empirical or statistical calibration. While this was also done for the midterm assessment, the relatively large differences in reading comprehension scores meant that this EGRA section was calibrated after the final assessment. This was done using a sample of 79 children who were not part of any of the previous three assessments. Children in both grades 2 and 3 participated, from several schools, in August 2010. Some children were given the baseline (2008) passage or set of words first, and then asked to read the midterm (2009) passage or set of words second, and then asked to read the final (2010) passage third. [13] The order was randomized so that we were able to remove the learning effect. The three assessments were well correlated, which was an important part of this calibration procedure. But the analysis also confirmed that the difficulty levels were slightly different, as Table A-1 shows.

Table A-1. Comparison of Calibration Results Across Three Versions of EGRA

Section	2008	2009	2010	Baseline to Midterm Adjustment	Baseline to Final Adjustment
Oral reading fluency	41.94	37.27	35.25	1.13	1.19
Reading comprehension	3.77	3.10	3.44	1.22	1.10

92. Therefore the results were adjusted, and the analyses presented in this report are calibrated results.

[13] Note that the same familiar word section was used in 2009 and 2010. As a result, and since the calibration exercise results for familiar words were not significantly different from those presented in the midterm report, no adjustments were made for the familiar word section.

Appendix B: Estimating the Impact of Full and Light Treatment on Outcomes, Disaggregated by Sex and Grade (extracted from differences-in-differences estimates)

93. This appendix investigates whether there were discrepancies by grade and sex on the impact of both full and light treatment. While grade 2 boys' achievement was lower than expected in letter naming fluency, grade 3 boys scored higher than expected on familiar word fluency, and grade 3 boys scored higher than grade 2 boys on oral reading fluency, few of the results deviated much from the aggregated findings (Table B-1).

Table B-1. Impact Disaggregated by Sex and Grade

		Grade 2		Grade 3	
	Treatment	Boys	Girls	Boys	Girls
Letter naming fluency (per minute)	Light	3.9	3.5	10.7**	8.1*
	Full	15.1***	15.7***	16.0***	13.8***
Phonemic awareness (words)	Light	0.6~	-0.2	0.8*	0.7*
	Full	1.9***	1.0**	1.2**	1.9***
Familiar word fluency (per minute)	Light	-0.3	-2.4	3.0	1.7
	Full	16.3***	11.4**	15.7***	14.8***
Unfamiliar word fluency (per minute)	Light	0.9	-1.1	0.7	0.2
	Full	11.3***	10.5***	13.1***	10.3***
Oral reading fluency (per minute)	Light	4.7	0.3	1.6	-2.2
	Full	25.5***	20.9***	20.6***	17.9***
Reading comprehension (%)	Light	2.8	-0.2	3.6	-2.9
	Full	30.1***	23.6***	27.4***	16.4***
Listening comprehension (%)	Light	-3.7	-0.4	2.6	4.9
	Full	8.6*	15.5***	13.6***	9.9**

****<.001, **<.01, *<.05, ~<.10

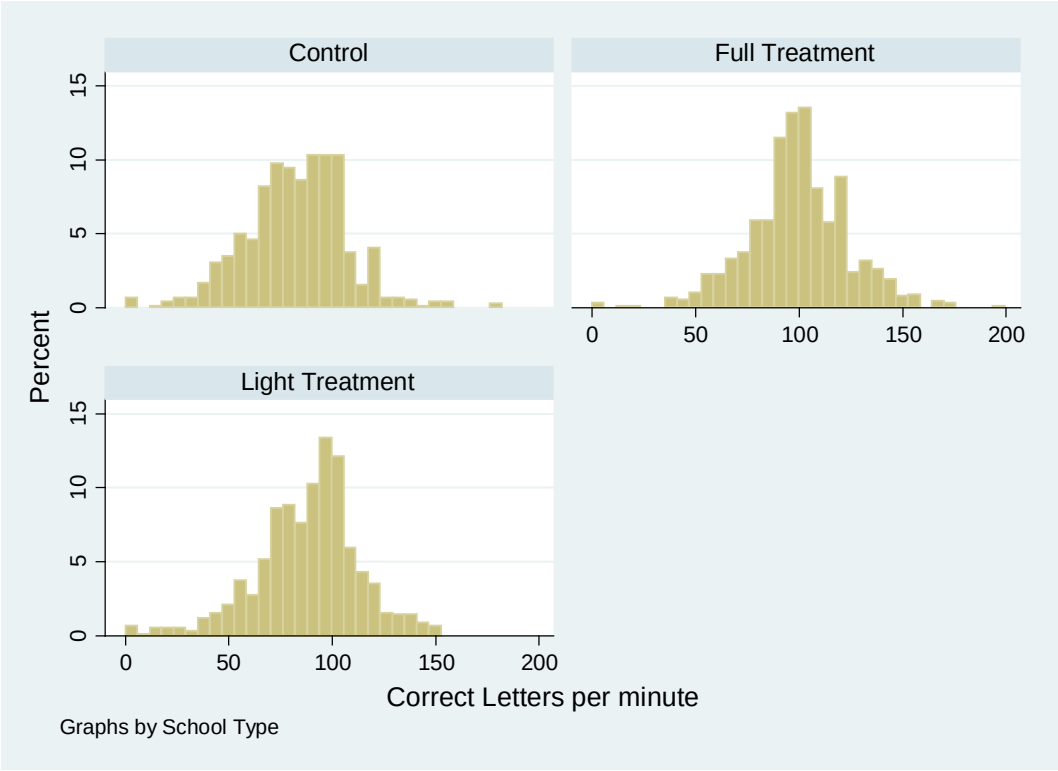
Annex C: Figure Analysis by EGRA Section

94. Here we present several graphics created to illustrate the relationship between treatment groups and achievements of the program as measured at the final assessment. Under the assumption (explained in other sections of the report) that the treatment and control groups were the same, these figures and the associated analyses show graphically the impact of EGRA Plus on student achievement at the final assessment, across treatment groups. By EGRA section, we look at which of several variables were predictive of reading outcomes, including grade and sex.

Letter Naming Fluency

95. Figure 3 shows the scores of control, full treatment, and light treatment children on the letter naming fluency section. Note that each bar presents the percentage of children from that treatment group who scored a particular number of letters per minute. Visual inspection shows that there were fewer children who scored zero or close to zero in the full treatment group than in either the control or light treatment groups. Similarly, more children scored 100 or more letters per minute in the full treatment group than in either of the other groups. In general, the full treatment group has a nearly normal distribution, while the control and light treatment groups have a slight leftward skew.

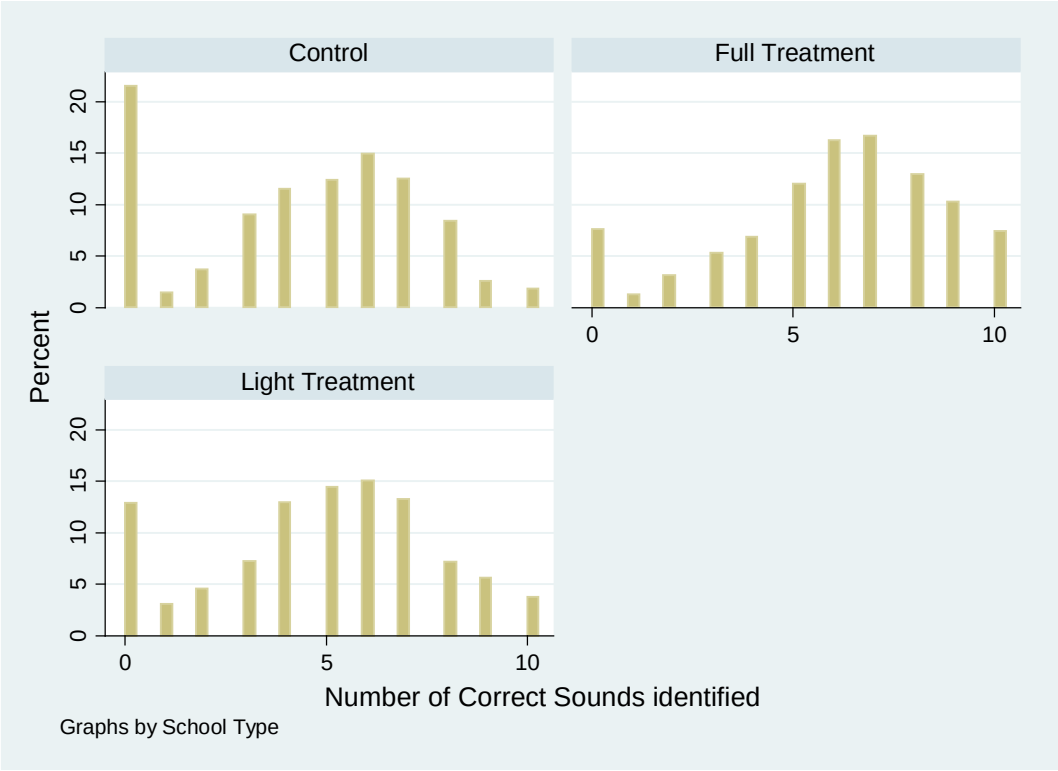
Figure C-1: Histograms Comparing Letter Naming Fluency Scores, by Treatment Group



Phonemic Awareness

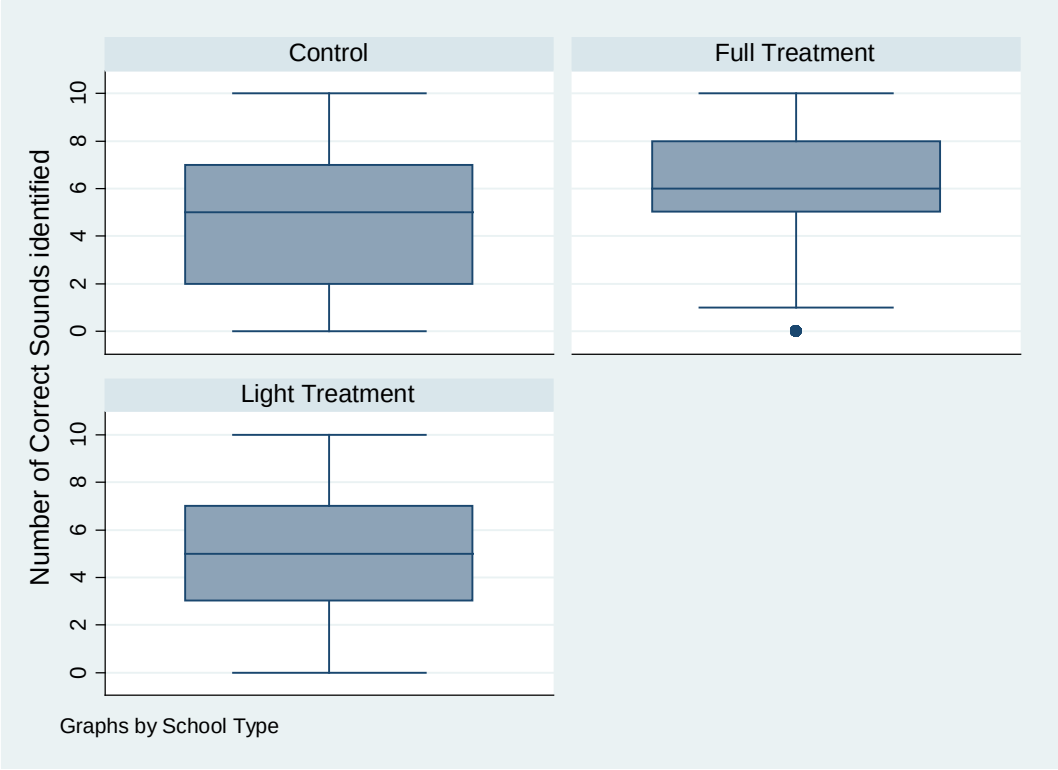
96. We also generated several figures to analyze the impact of the EGRA Plus program on phonemic awareness scores. In Figure 4 below, which presents box plots for each of the treatment groups on letter naming fluency, notable differences can be detected. First, a lower percentage of children scored zero on the phonemic awareness section in the full treatment schools than in either the light treatment or control schools. Other than those zero scores, the scores are nearly normally distributed for both control and light treatment. On the other hand, for full treatment, there is a rightward skew, with a larger percentage of children reading letters fluently.

Figure C-2: Histograms Comparing Phonemic Awareness Scores, by Treatment Group



97. Figure 5 below disaggregates by treatment group the achievement on the number of sounds identified. The mean scores (6.0 words correct) for the children in full treatment schools were much higher than those for either control or light treatment. The 75th- percentile scores also were much higher, with full treatment children reading eight words correctly on this section, compared to between six and seven words in control and light treatment schools. The 10th-percentile scores (the bottom line) were above zero for the full treatment children, allowing light treatment schools to have an outlier with respect to the number of phonemic awareness tasks correctly performed. On the other hand, the 10th-percentile score for both control and light treatment was zero words correct. This indicates a substantial gap in achievement on phonemic awareness between full treatment and light treatment and control schools.

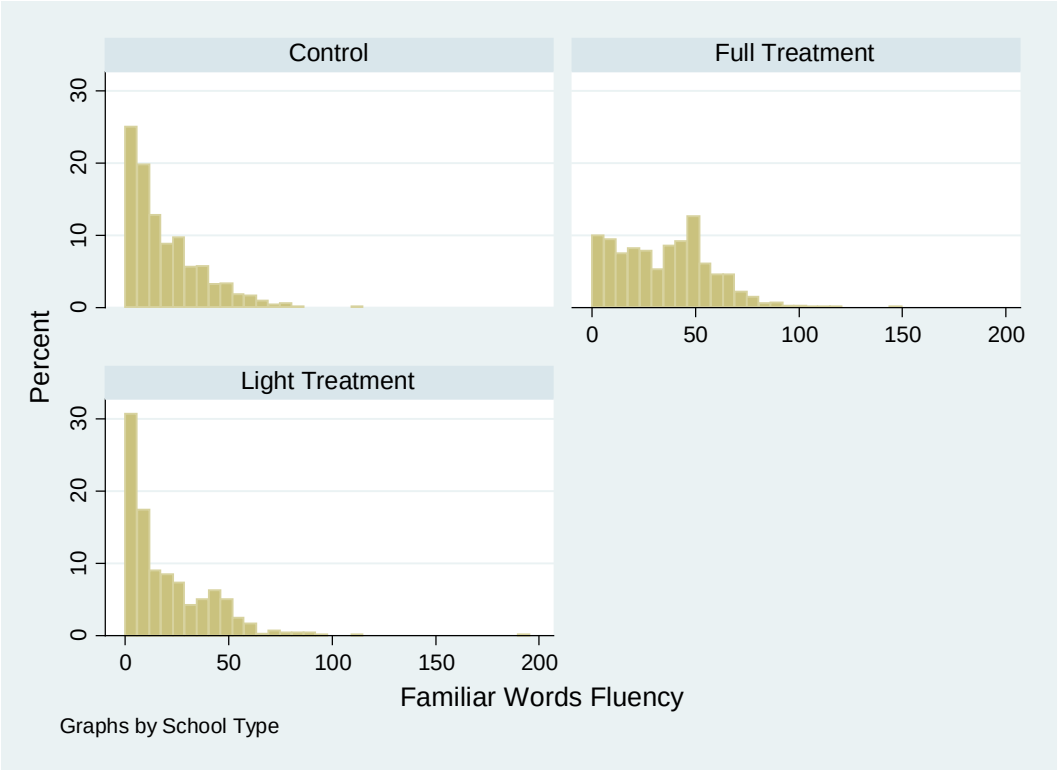
Figure C-3: Box Plots Comparing Phonemic Awareness Scores, by Treatment Group



Familiar Word Fluency

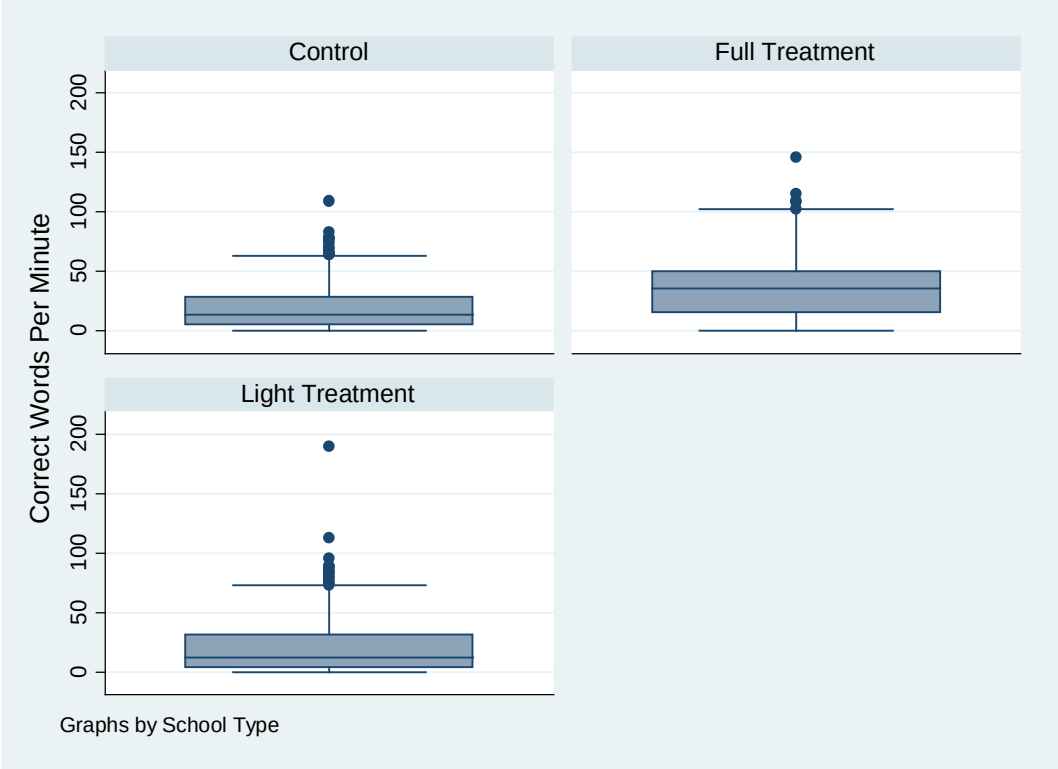
98. When we analyzed the results of the program's impact on the number of familiar words that children could identify in one minute, we found a relatively large impact of the full treatment when we compared children in each type of school (Figure 6). The score with the highest percentage for control and light treatment schools was zero words per minute, while for full treatment, the modal score was 50 words per minute. Moreover, the tail of the full treatment schools was more evenly distributed beyond the 50-words-per-minute mark. In other words, a significant percentage of children who could read 50 words per minute were in the full treatment schools.

Figure C-4: Histograms for Familiar Word Naming Fluency, by Treatment Group



99. In the box plots comparing treatment groups in Figure 7, it is easy to note substantial differences in word reading fluency. For example, the 75th- and 90th-percentile scores were higher for full treatment schools than for control or light treatment schools. The 90th percentile was nearly 100 words per minute for full treatment schools, but somewhere around 70 words per minute for control and light treatment. Similarly, the means were higher in full treatment schools, with control and light treatment mean scores less than 25 words per minute, and the full treatment schools nearer to 50 words per minute. The small differences between control and light treatment were also notable at the 10th and 25th percentiles, which shows that significant portions of the sampled learners were scoring at those levels. That was not the case for the full treatment schools, however. In short, on familiar word fluency, there was a consistent advantage for full treatment schools at the expense of both control and light treatment schools.

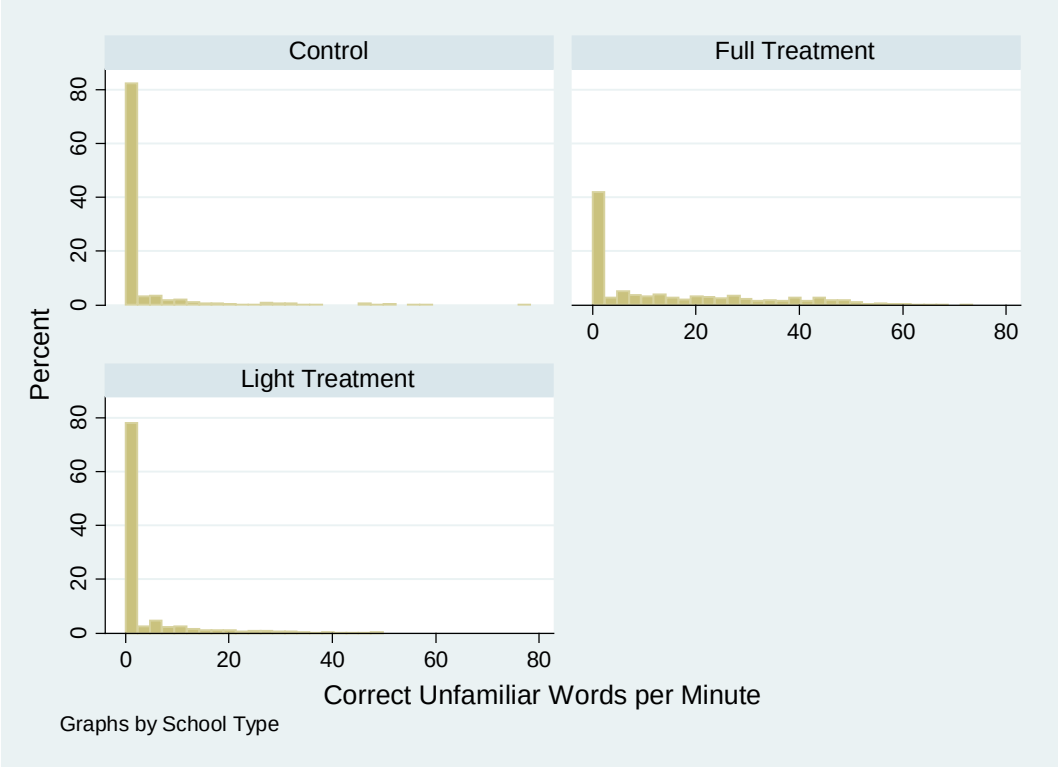
Figure C-5: Box Plots Comparing Familiar Word Fluency, by Treatment Group



Unfamiliar Word Fluency

100. The descriptive statistics above showed that the scores for unfamiliar words were quite low. This is borne out in Figure 8, which shows that nearly 80% of light treatment and more than 80% of control children read zero unfamiliar words per minute. Less than 40% of full treatment children, on the other hand, read zero unfamiliar words per minute. The histograms also show that the distribution of children reading more than zero words per minute on this section was much more substantial and more widely spread in the full treatment sample than in either full or light treatment.

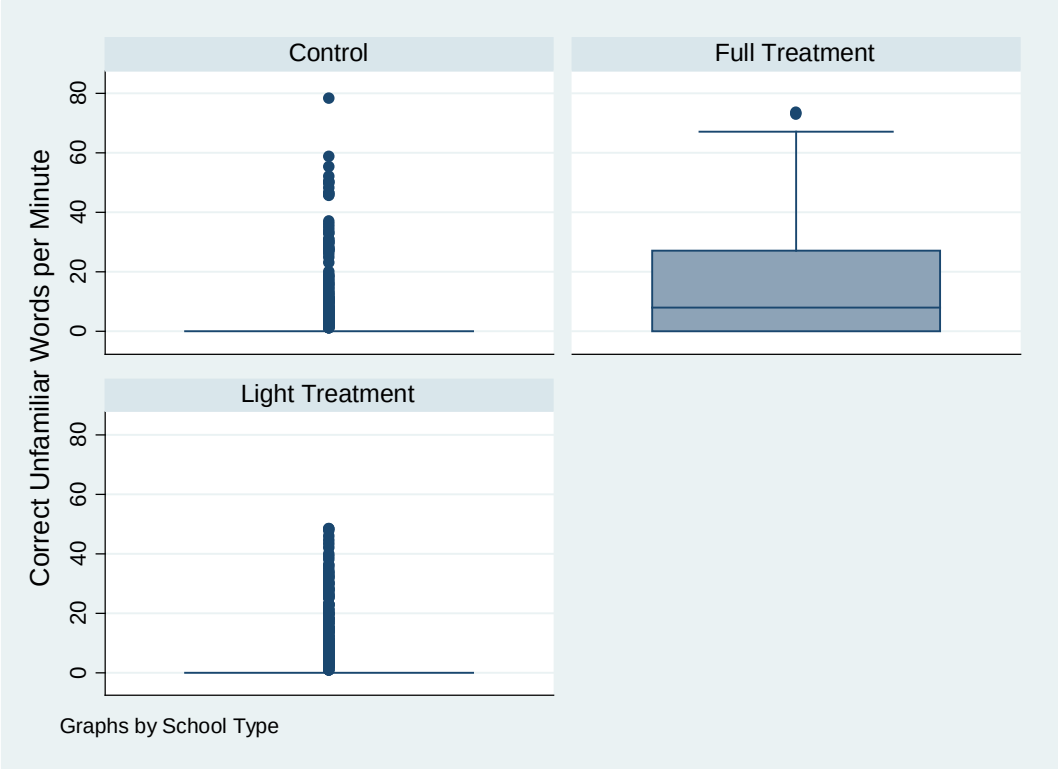
Figure C-6: Histograms Depicting Achievement on Unfamiliar Word Fluency, by Treatment Group



101. Figure 9 shows this point more clearly. It compares all three treatment groups and the two grades. It shows, particularly when we compare control and full treatment children, that the EGRA Plus program helped children move from zero scores to farther along the distribution. This is an important finding for equity: EGRA Plus not only helped high-achieving children expand their reading knowledge, but also helped the lower-achieving children increase their scores. The Figure 9 box plot illustrates an important point: Children in full treatment schools had enough variation in their scores that the mean, 75th percentile, and 90th percentile were all removed from zero. This shows that in full treatment schools, in particular for nonsense words, the program had an impact on the lowest achieving students.

102. Figure 9 also makes quite evident the wide gaps between the control/light treatment and full treatment schools. While the 25th-, 50th-, and 75th-percentile scores were all concentrated at zero words per minute for control and light treatment, the mean score for full treatment was more than 10 words per minute and the 75th percentile was significantly more than 20 words per minute. The majority of children in schools that had EGRA Plus full treatment support were much more capable of decoding. Moreover, the 90th percentile of the distribution was 60 words per minute, which is approximately the same as the greatest outlier in control schools, and higher than the entire light treatment sample. This EGRA section was one in which the full treatment had a significant impact on student outcomes, and particularly in the skill of decoding new words.

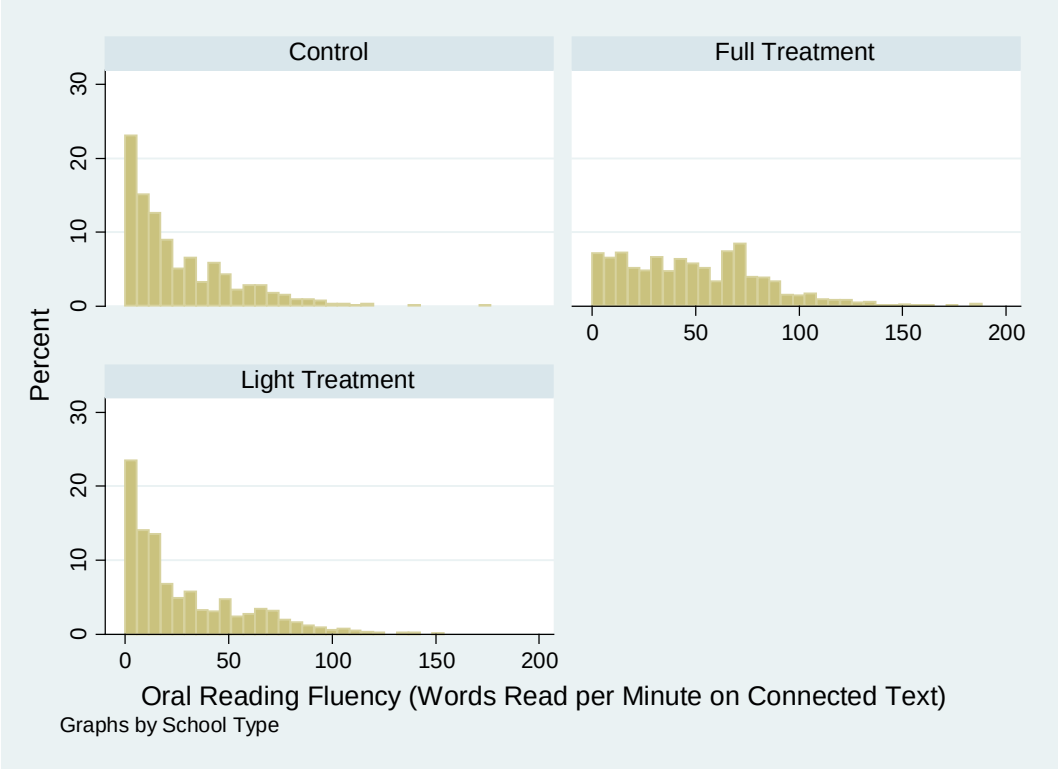
Figure C-7: Box Plot Showing Unfamiliar Word Fluency, by Treatment Group, for Grades 2 and 3 Combined



Oral Reading Fluency (Connected Text)

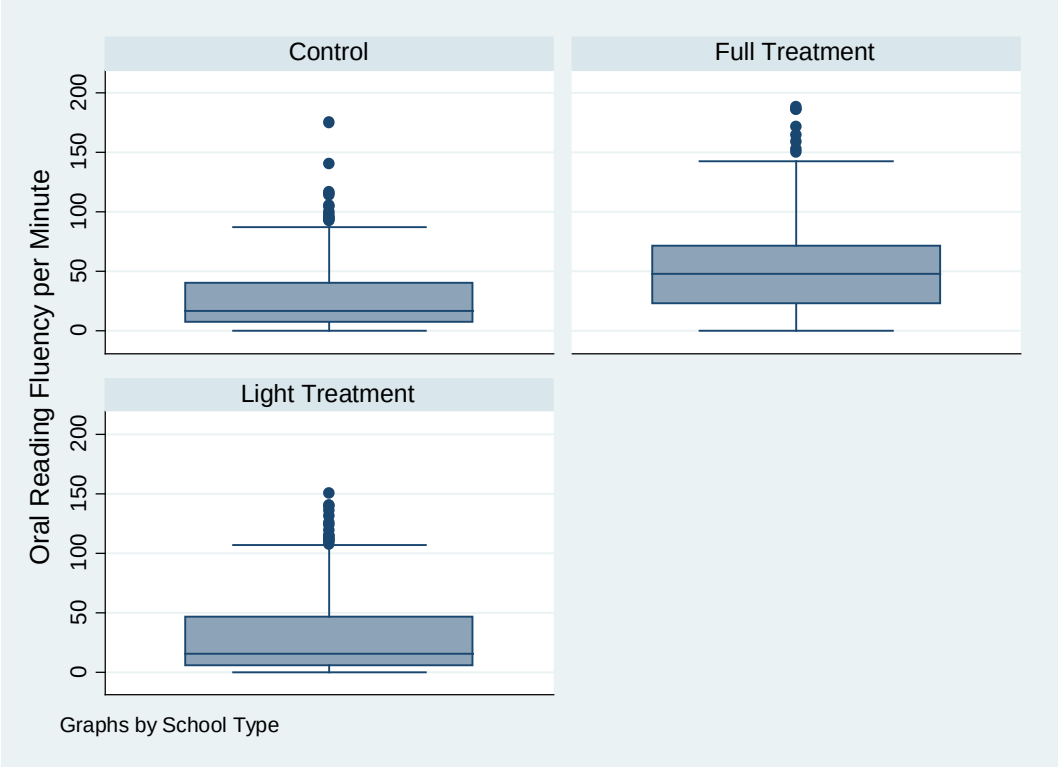
103. Figure 10 below shows the relationship between oral reading fluency and treatment status. In both control and light treatment schools, more than 20% of the sample read zero words per minute, and significant percentages read quite close to zero words per minute. On the other hand, while the oral reading fluency scores for the full treatment schools were skewed to the left, the percentages of children reading modest amounts on oral reading fluency were significantly less. The percentages of children who read 50 words per minute or more in the full treatment schools were substantial, and much more than the scattering of fluent readers in control and light treatment schools. Note also from Figure 10 that a higher percentage of light treatment school children could read at least a few words.

Figure C-8: Histograms Showing Oral Reading Fluency Scores, by Treatment Group



104. The box plots in Figure 11 were designed to show whether and how the EGRA Plus program had an impact on oral reading fluency scores for children in treatment schools. The mean score for full treatment was higher than the 75th-percentile oral reading fluency scores for both control and light treatment children. More powerfully, the 25th-percentile level for full treatment was higher than the 50th percentile for both of the other groups. The high scores also show the differences by treatment group, with the 90th-percentile score for the full treatment schools being 50 words per minute more than control, and nearly 50 words per minute more than light treatment. The substantial impact of the program therefore is apparent across the whole distribution.

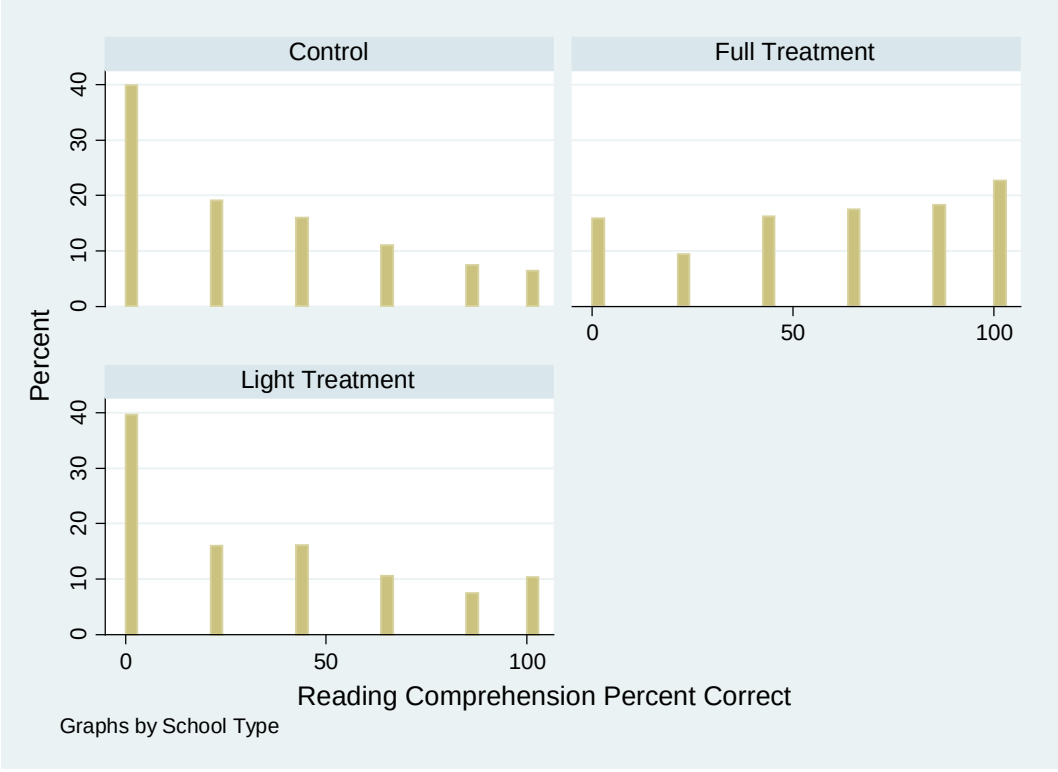
Figure C-9: Box Plots of Oral Reading Fluency Scores, by Treatment Group



Reading Comprehension

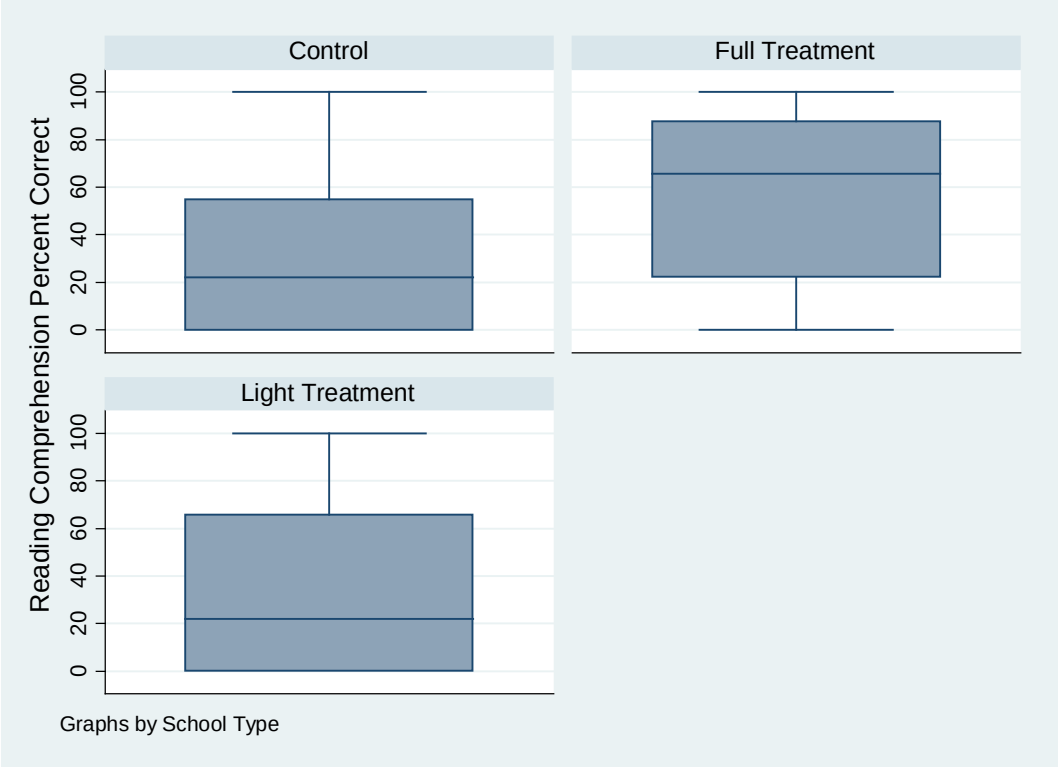
105. Figure 12 shows the relationships between achievement on reading comprehension and treatment status. Note that we would expect the children in treated schools to outperform their control colleagues since they outscored them on oral reading fluency and the two sections are linked. This might not be the case, however, if the program only increased children's ability to read and sound out words, rather than synthesize and understand what they read. For control and light treatment schools, 40% of children scored 0%correct on this section. The corresponding figure for full treatment was less than 20%. Looking at the other end of the distribution, nearly 20% of the full treatment children scored 80% correct; and more than 20% of them read the story at 100% comprehension. This far surpassed the achievement of both control and light treatment schools, where only about 10% of the entire distribution scored either 80% or 100%. The treatment, then, contributed heavily to students' understanding. This was in contrast to the midterm results, where the impact on reading comprehension was minimal. This was due partly to the lack of calibration between the reading comprehension sections (which has now been rectified) and to the more modest impacts on oral reading fluency found in the midterm. By the time of the final evaluation, children were benefiting a great deal from the full treatment, across sections, and particularly in reading comprehension.

Figure C-10: Histograms Showing Reading Comprehension Scores Overall, by Treatment Group



106. The box plot presented in Figure 13 reinforces the points made above. While it is clear that at the 75th percentile, light treatment schools outperformed control schools, neither group achieved at anything close to the level of the full treatment schools. For example, the 25th-percentile score for full treatment was close to the 75th-percentile score for both control and light treatment. The mean scores for full treatment were significantly higher than the 75th percentile for either full or light treatment. Thus, in the area of reading comprehension, the distribution between the full treatment and light treatment/control school outcomes was quite substantial.

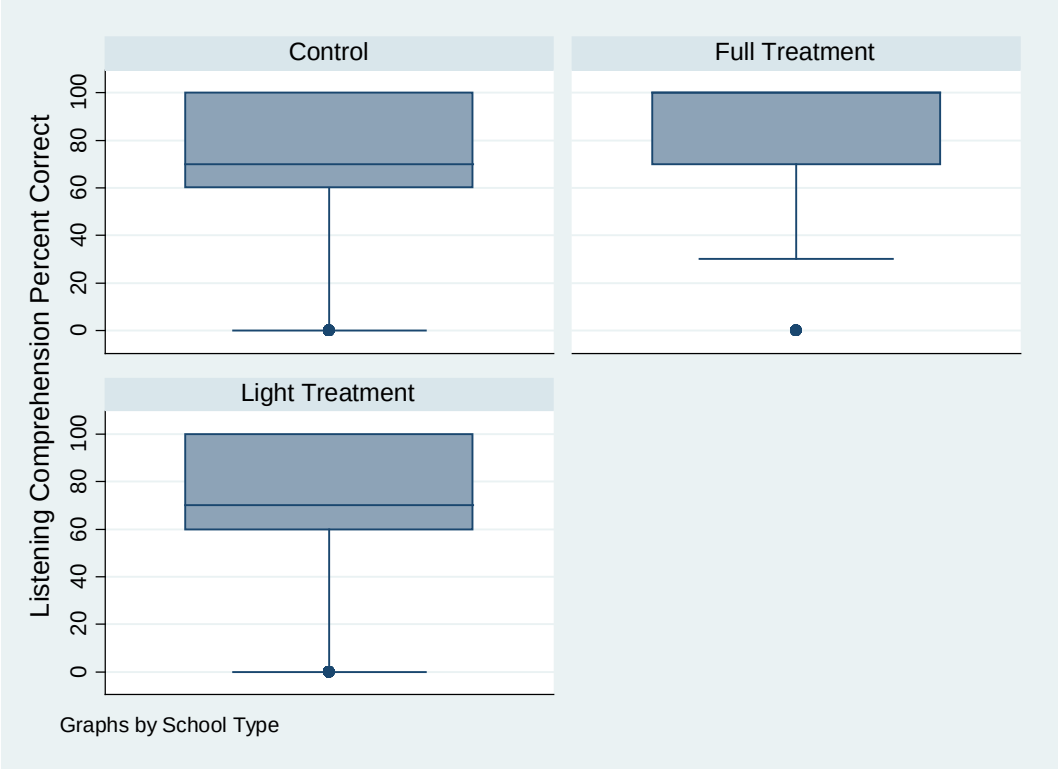
Figure C-11: Box Plot of Reading Comprehension Scores, by Treatment Status



Listening Comprehension

107. Section 10 closes with a brief investigation of the distribution of scores on listening comprehension, by treatment group, as depicted in Figure 14. Note that the figure indicates that the average score in the full treatment was 100%, while the average for both control and light treatment was 67%. As a result of the program, children were much more able to understand what they heard.

Figure C-11: Listening Comprehension Scores, by Treatment Status



Correlations Between Oral Reading Fluency and Reading Comprehension

108. In Figure 15 there are three scatterplots. They represent the relationships between children's oral reading fluency and those same children's reading comprehension scores. The scatterplots are divided by treatment group. It is interesting that there is a consistently linear relationship between oral reading fluency and reading comprehension, across all three samples. In other words, children's ability to read fluently was very useful in predicting their ability to comprehend what they read. The issue, therefore, is that far too few children could read with sufficient fluency to comprehend at a high level. The differences between the treatment groups depicted in Figure 15, then, are not in the slope of the predictive relationship, but in the density of the population. That is to say, children in full treatment schools were more likely to read at 50 words per minute, and therefore were much more likely to read with higher levels of comprehension.

Figure C-12: Scatterplots between Oral Reading Fluency and Reading Comprehension, by Treatment Group