USAID | LIBERIA
FROM THE AMERICAN PEOPLE

# EGRA Plus: Liberia

## Data Analytic Report:
## EGRA Plus: Liberia Midterm Assessment

**Early Grade Reading Assessment (EGRA) Plus: Liberia**
**EdData II Task Number 6**
**Contract Number EHC-E-06-04-00004-00**
**Strategic Objective 3**
**October 31, 2009**

This publication was produced for review by the United States Agency for International Development. It was prepared by RTI International and the Liberian Education Trust.

# EGRA Plus: Liberia

Data Analytic Report:

EGRA Plus: Liberia Midterm Assessment

Contract EHC-E-06-04-00004-00

October 31, 2009

Prepared for
USAID/Liberia

Prepared by:
Benjamin Piper
Medina Korda

RTI International
3040 Cornwallis Road
Post Office Box 12194
Research Triangle Park, NC 27709-2194

# Table of Contents

# List of Tables

# List of Figures

# 1. Executive Summary

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 measurement only and toward the types of interventions that can increase student achievement. Liberia, for example, has recently begun using EGRA-based interventions to improve the quality of learning outcomes in reading.

Liberia's path towards intervention started with a pilot assessment using EGRA in 2008, which was used to complete a system-level diagnosis with areas of improvement. The World Bank funded the pilot in 2008. The Ministry of Education (MOE) and USAID/Liberia decided to fund a two year intervention program, called EGRA Plus: Liberia, to improve student reading skills by implementing evidence-based reading instruction. EGRA Plus is both an intervention and an experiment, as it is designed as a randomized controlled trial. Three groups of 60 schools were randomly selected, for a total experiment of 180 schools. These groups were clustered within districts, such that several nearby schools were organized together. The intervention is targeted at grades 2 and 3. The design is as follows: The control group does not receive any interventions but will be assessed using EGRA. In the "full" treatment group, reading levels are assessed, teachers are trained how to continually assess student performance; teachers are provided frequent school-based pedagogic support, resource materials, and books; and, in addition, parents and communities are informed of student performance. In the "light" treatment group, the community is informed about reading achievement, and students are assessed.

This midterm assessment evaluates the impact of several months of instruction and program impact. Based on the initial pilot assessment (June 2008) and the curriculum review, RTI and local stakeholders determined that the remedial intervention should begin with the creation of an instructional model and key reading subskills that need to be taught. A clear model and a scope and sequence of instruction for each of the five key components of reading, and for each grade (2 and 3), was developed. The reading intervention implementation unfolded as planned with the exception of a delayed start due to a volunteer teacher strike that took place early January 2009.[1] Coupled with this was a need to commence the project's midterm assessment before the end of the academic year, in the last week of May 2009. In the end, teachers had slightly more than three months to teach reading. It is our experience with similar projects conducted elsewhere that in the first year of the project, one usually does not tend to see significant, if any, improvements.

In November 2008, RTI International and its subcontractor—Liberian Education Trust—collaborated with Liberian education officers to collect a nationally representative

[1] EGRA target teachers were trained before December 25, 2008, with the plan to commence the reading intervention on January 5, 2009. However, most of EGRA schools were closed until late January, and by the time the project revived the momentum, it was mid-February when the teachers started using the reading manuals.

comprehensive baseline early grade reading assessment in project and control schools in grades 2 and 3.[2] In May and June 2009, the midterm assessment was conducted in these same EGRA schools.[3] Students were assessed on a full battery of early grade reading subtasks, including letter naming fluency, phonemic awareness, familiar word fluency, unfamiliar word fluency, connected text oral reading fluency, reading comprehension, and listening comprehension. The test used for the midterm assessment was equated to that of the baseline assessment in order to ensure comparability of data. Analysis of the EGRA instrument itself showed that the assessment was reliable and the various subtasks assessing different parts of the underlying early grade reading skills tied together well as a reliable test. In fact, the Cronbach's alpha results for both baseline and midterm assessments showed reliability of about 0.85, which is quite good.[4]

At the baseline assessment, Liberian children were capable of identifying the names of letters for the most part, with the average grade 2 control child identifying 57.9 letters in a minute and the average grade 3 child identifying 69.0 letters. At the midterm, students in grade 2 in full treatment schools showed a 51.9% increase in letters read, and grade 3 students increased by 42.4%. Interestingly, children in light treatment schools increased their scores over baseline by 36.4% and 32.5% in grade 2 and grade 3, respectively. These were larger impacts than we expected, and with respect to program impact, the increases had between a .47 standard deviation (SD) and .75 SD effect size, remarkably large.[5]

Program impact on phonemic awareness was more moderate. Combining scores on grades 2 and 3 shows that the number of sounds identified increased by 29.8% and 16.9% in full and light treatment schools, with an effect size of .22 SD and .06 SD, respectively. This represented a substantive increase of 0.6 and 0.2 words read. For familiar words, children in full treatment schools increased by 68.5% and light treatment schools by 54.2%. Since control schools increased their skills as well, the effect size was .09 and .05 SD, still notable, but not very large. This represents an increase of 1.6 and 0.9 words per minute. For unfamiliar words, both control and light treatment schools decreased the number of words read per minute between baseline and midterm, with light treatment schools decreasing by 22.5%. On the other hand, full treatment schools increased by 78.7%, an increase of 1.5 words per minute on a baseline of 1.8 words. The effect size for full treatment schools was .25 SD.

Given the importance of oral reading fluency skills in future academic achievement and the ability to move from learning to read and reading to learn, much of this report focuses

[2] Baseline: 176 schools were assessed, including 57 control, 59 full treatment, and 60 light treatment schools, for a total of 2957 students.
[3] Midterm: 175 schools were assessed, including 56 control, 59 full treatment, and 60 light treatment schools.
[4] Cronbach's alpha is a measure of how well a set of variables (in this case, Early Grade Reading Assessment subtasks) measure an underlying construct (in this case, early grade reading skill). In short, it is a measure of test reliability.
[5] Note that the effect sizes reported here are Cohen's *d*. Small effect sizes are from 0 to .40, moderate from .40 and .75, and large higher than .75.

on oral reading fluency levels and the impact of various predictor variables on this construct. Compared against baseline, full treatment children increased the number of words read correctly by 51.2%, while light treatment schools increased by 28.9%. Substantively, this means that full treatment schools increased their number of words read from 19.4 to 29.5 words per minute, while light treatment increased from 21.0 to 27.1. Compared against the gains for control schools, these effect sizes are positive, at .42 SD and .19 SD, for full and light treatment schools. This means that at midterm, children in full treatment schools were reading 7.2 words more per minute than those in control (29.5 compared to 21.0). The gap was also large for light treatment schools, with a difference of 6.1 words (27.1 compared to 21.0). This analysis shows clearly that there were moderate absolute gains, as well as statistically significant program effects, with effect sizes either small or moderate, and notable for a program in the education sector.

Given the close relationship between reading comprehension and connected text fluency in EGRA, it is unsurprising that there is a strong correlation between the two scores. Comparing the midterm and baseline assessment scores, we find that full treatment schools increased their scores by 1.6% over baseline, while light treatment scores decreased by a 18.2%. This was less than the control schools decrease of 30.8%. This means that, at midterm, children in full treatment schools scored 6.6 percentage points higher than those in control, with light treatment school students scoring 4.4 points higher. The program's effect size was .34 SD and .13 SD for full and light treatment schools. This equates to an increase of 8.2 percentage points more than control for full treatment and 3.2 percentage points more than control for light treatment. For listening comprehension, the increases for full and light treatment schools were 128.3% and 130.4%, respectively. In fact, the scores increased from 34.3% to 78.3% correct for full schools and 33.6 to 77.5% for light treatment schools, large increases both. It should be noted that control schools increased their scores by 103.9%, so only by taking into account the baseline scores can a true program effect be estimated. Substantively, full treatment schools increased by 9.9% and light treatment schools by 9.8% more than baseline schools, an effect size of .33 SD and .32 SD, respectively.

The large sample size allows more precision in the estimation of differences between grades and gender. In all subtasks, grade 3 students scored statistically significantly higher than grade 2 students, with more than 10 additional words read correctly per minute on the oral reading fluency subtask. On the other hand, there were no differences between boys' and girls' achievement, except in oral reading fluency and reading comprehension, where girls outperformed boys. This suggests that more work is necessary to ensure that the program increases the skills of boys in the more complex portions of reading.

Finally, the research design lends itself to more sophisticated analyses using differences-in-differences.[6] These analyses show that the full treatment group increased student achievement for every subtask, often with relatively large impacts on student achievement. The light treatment group increased student achievement in letter fluency, unfamiliar word fluency, reading comprehension, and listening comprehension.

In summary, given the impediments to program implementation, the short time frame, and the relatively modest cost of the program, EGRA Plus: Liberia outperformed expectations with respect to impact on student achievement, particularly in the full treatment schools. Those effects can be shown in ***Table 1*** below. Note that the effects were moderate in size, always larger in full than light treatment schools, with effect sizes for full treatment schools from .15 to 77 SD. Impacts were largest in letter fluency, and smallest in familiar words.

**Table 1. Program effects and effect sizes, by treatment group and outcome**

| Outcome measure | Treatment group | Program effect | *p*-value | Effect size |
|---|---|---|---|---|
| Letter fluency | Light | 13.62 | <.001 | .54 SD |
| | Full | 19.56 | <.001 | .77 SD |
| Phonemic awareness | Light | .18 | .27 | *No effect* |
| | Full | .62 | <.001 | .27 SD |
| Familiar word fluency | Light | 1.00 | .30 | *No effect* |
| | Full | 1.79 | .07 | .13 SD |
| Unfamiliar word fluency | Light | -.66 | .09 | *No effect* |
| | Full | 1.56 | <.001 | .26 SD |
| Oral reading fluency | Light | 3.43 | .02 | .17 SD |
| | Full | 7.39 | <.001 | .38 SD |
| Reading comprehension | Light | 3.21 | .04 | .13 SD |
| | Full | 8.53 | <.001 | .35 SD |
| Listening comprehension | Light | 9.61 | <.001 | .47 SD |
| | Full | 9.92 | <.001 | .48 SD |

[6] Differences-in-differences is an identification strategy that attempts to make causal inference about a treatment effect by removing the secular trend using a pre and post, treatment and control design. Skoufias, E. & Shapiro, J. (2006). *The pitfalls of evaluating a school grants program using non-experimental data.* Working paper. Washington, DC: World Bank.

When compared against the Performance Management Plan (PMP) of February 2009, the results from the EGRA Plus are mixed. The PMP noted that the impact over baseline would be a 20% increase for connected text fluency and reading comprehension in full treatment schools, while light treatment schools would see a 5% increase for those same subtasks. ***Table 2*** below shows mixed progress towards that goal. It shows that for both boys and girls, for both grade 2 and grade 3, the light treatment schools met their target. By the same token, for both genders and both grades, the full treatment schools increased by more than 20%, often much more than 20%. The results were mixed, however, for reading comprehension, where only grade 2 girls in full treatment schools made the PMP's target. It appears very likely, however, that the lower absolute changes on the midterm reading comprehension assessment stemmed from the fact that the reading comprehension passage questions were not equated, while the oral reading fluency passage was, since in every case, there was a significant program effect over the changes in the control schools.[7]

**Table 2. Disaggregated analysis of percentage increases over baseline, by treatment status, grade, and gender**

| | | Grade 2 | | Grade 3 | |
|---|---|---|---|---|---|
| | **Treatment** | **Boys** | **Girls** | **Boys** | **Girls** |
| Oral reading fluency | Control | -3.9% | 68.5% | 2.0% | 22.6% |
| | Light | **7.8%** | **60.2%** | **6.4%** | **54.5%** |
| | Full | **28.2%** | **150.7%** | **22.3%** | **61.2%** |
| Reading comprehension | Control | -40.7% | -11.5% | -37.5% | -25.0% |
| | Light | -29.0% | -11.2% | -29.3% | -3.7% |
| | Full | -7.8% | **44.4%** | -8.2% | 2.5% |

# 2. Introduction

The EGRA Plus: Liberia program (2008–2010) is an experimental intervention. The intervention is part of a joint collaboration of the Liberian Ministry of Education, the World Bank, and USAID/Liberia. A baseline, midterm, and final assessment will be conducted and assessed against agreed-upon targets for improved student performance. The baseline assessment was conducted in November 2008, the midterm assessment was conducted in June 2009, and the final assessment is scheduled for June 2010.

[7] This type of finding is why many scholars prefer accounting for program impact in terms of percentage change over baseline and control, so that any secular changes are accounted for in program impact.

The EGRA Plus: Liberia program uses empirical data from early grade reading assessments to track progress toward quality improvements in early grade reading instruction, with particular focus on phonics-based instruction. The research and intervention design allows for the comparison of three different treatment groups. The first is a control group which will receive no program interventions. The second group, the "light" intervention, is a set of schools where parents and community members are provided student achievement data in the area of literacy. The final group, the "full" intervention, provides an intensive teacher-training based program targeting reading instructional strategies, in addition to the information on student achievement provided to parents and communities in "light" treatment schools.

In this report, we present the project's performance at the midterm by comparison with baseline assessment results.[8] In what follows, we briefly provide a description of 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 2957 students. [9] The assessment was to be conducted in all of the project's schools: 60 control, 60 light, and 60 full intervention schools.[10] In each school, either 10 or 20 students were assessed, depending on the size of the school and number of teachers. The assessment itself had several components, which have been tested in a variety of other low-income countries, as well as the June 2008 pilot assessment in Liberia.

The June 2009 midterm assessment was conducted in the same EGRA schools. A total of 175 schools for a total of 2805 students was included in this survey. As was the case with the baseline assessment, either 10 or 20 students were assessed, with a goal of having at a minimum 10 students from grade 2 and 10 students per grade 3, depending on the size of the school and number of teachers. For both assessments, students were randomly selected using a systematic sampling procedure implemented by assessors, rather than teachers, in order to prevent teachers to select only the best students.

As noted in Section 1, analysis of the EGRA itself showed that the assessment is reliable and the various subtasks assess different parts of the underlying early grade reading skills as well as tying together well as a reliable test. In fact, the Cronbach's alpha results show reliability of 0.85, which is quite good.

Beginning portions of the analytic report lay out the various subtasks of the assessment, and point out how they are related to important characteristics of early reading skills and proficiency. The analysis presented here focuses on a particular set of research questions

[8] In addition, ***Appendix A*** compares grade 2 baseline results with grade 3 midterm results, as a way of taking into account the instructional delays that occurred between baseline and midterm.

[9] The missing four schools were assessed in January and February 2009, but were not included in the baseline data analysis.

[10] The sampling procedure used in this study and in the intervention is a means of identifying the true impact of the program. Without having a counterfactual, a comparison group, it is impossible to know whether any impacts we see are the result of program effects, typical growth over the course of the school year, or changes that apply to all students equally. Having a control group allows us to differentiate among those possibilities. In this case, there is one control group and two experimental groups—one having a full intervention and one a light intervention.

designed to inform the early stages of the program intervention as well as to provide a baseline of early grade reading skills across Liberia. This analytic report is organized in the following way:

- First, descriptive statistics are presented for both predictor and outcome variables. Then, we compare these descriptive statistics across important characteristics, particularly student gender, treatment group, and grade level.
- Second, we assess the reliability of the assessment itself using a variety of statistical methods and follow this by presenting correlations of relevant variables.
- Third, we use simple comparisons between treatment and control groups to estimate the impact of the program.
- Fourth, we present graphic depictions of student achievement across various metrics and present some models that predict student achievement in early reading.

# 3. Early Grade Intervention in Reading

The EGRA Plus: Liberia intervention, designed based on the findings of the World Bank pilot assessment of reading in 2008,[11] is based on a three-stage intervention strategy. First, a baseline EGRA was implemented in a nationally representative set of Liberian primary schools. This assessment serves as the baseline for the impact evaluations, but also informs 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.

Second, RTI International, in collaboration with Liberia Education Trust and the Ministry of Education, is implementing a teacher professional development program that encompasses intensive, week-long capacity-building workshops using early grade teaching skill techniques, ongoing professional development, external support, and existing processes and procedures for ongoing feedback. The intervention is 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, along with the fostering of interactive meetings between school managers and community members. This set of school and community action activities serves as the main intervention in light intervention schools, while full intervention schools also receive on-site professional development and supervision support for grade 2 and 3 teachers.

[11] Crouch, L., & Korda, M. (2008). *EGRA Liberia: Baseline assessment of reading levels and associated factors.* Report prepared as part of a process of collaboration between USAID and the World Bank. Research Triangle Park, North Carolina: RTI International. https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=158

The third major intervention activity is an additional two rounds of early grade reading assessments in Liberia, allowing for a truly longitudinal research design. This design allows researchers and the Ministry of Education to identify whether and how the interventions have had a significant impact on student achievement, as well as by what causal mechanisms the program was successful.

Now that the project has reached the end of Year 1, it is important to share some of the background information on how the intervention has unfolded. The approach to supporting teachers was adjusted and intensified due to some of the challenges that teachers, Coaches (master trainers), and education officers faced.

The implementation of the reading intervention in 60 full treatment schools commenced with teacher training in December 2008. At this time, resources were provided to teachers in hopes that if they were trained and provided with materials prior to the holidays, teachers would spend time preparing for teaching reading. The school academic year did not resume on January 5, 2009, as per the academic calendar, but rather on January 19, due to a volunteer-teacher strike caused by the dismissal of all unqualified volunteer teachers. Vis-à-vis the EGRA Plus project, this delay undoubtedly had a negative effect on the momentum created in December 2008.

While some schools, mainly in Monrovia, started teaching on time (January 5), most of the schools did not open their doors to children until late January 2009. Even when classes resumed, teachers focused on wrapping up exams and reports for Period 3, and in most cases, the EGRA reading intervention did not start until mid-February 2009. This disruption also had an impact on teacher and Coach morale, since nearly 30% of EGRA teachers were volunteer teachers.

This situation presented a huge challenge to the project, for two reasons. First, the EGRA team needed to train replacement teachers, and to continue encouraging volunteer teachers to consider the EGRA project as a way to improve their skills. Second, a number of volunteer teachers left schools, causing a bigger burden on the teachers staying behind, given that now they had to teach more children than before. As a result, there were instances where grades 2 and 3 were combined into one class. In some schools, the principals started to teach and Coaches began helping with teaching.

The same factors and assumptions described for full treatment schools above also apply to light treatment schools. The original plan was for Coaches to visit the light intervention schools as soon as schools opened in January 2009 in order to share the EGRA assessment results and provide initial training. This was delayed until February, which is when the workshop training was conducted in all light treatment schools. Other factors that affected the intervention were an ingrained pattern of insufficient time spent teaching reading in classrooms, a low skill base on which to scaffold reading instructional strategies, and a lack of general pedagogic skills such as lesson planning.

The reading program is very specific and organized; it demands planning skills from teachers and, most importantly, dedication. If followed, this program will lead to

significantly improved student performance in reading in less than one year. However, teaching reading, rather than language arts, is new to many teachers in Liberia and they find it challenging. Teachers also struggle with lesson planning and delivery. Working toward clearly specified goals while measuring their progress along the way is also demanding of teachers simply because it requires time, skills, and dedication. Our analysis of the curriculum in Liberia (Davidson & Crouch, June 2008)[12] indicates that while curriculum goals are specified, insufficient information is provided as to how to achieve those goals. We also believe that teachers need to be held accountable for delivery and that accountability mechanisms, such as strong and empowered parent-teacher associations (PTAs), need to be supported and strengthened systematically. Throughout the EGRA Plus: Liberia project, this accountability has begun to be put into place. Teachers are continually assessed and are supported by Coaches, and they know that the project is tracking improvements in progress.

Some teachers complained that EGRA work was extra effort imposed in addition to the regular school curriculum. Coaches, in response, reminded teachers that teaching reading is a subject that is part of the curriculum. While teaching language arts is very important, teaching children how to read proficiently as early as possible is the most important precondition for the child's further cognitive development. Without reading, children will lag behind and it will become harder and harder for them to catch up as they get older. They will also perform poorly on other subjects given their insufficient reading skills.

Another interesting research and policy issue might be the organization and effectiveness of PTAs. In most of the target districts, some PTAs were recognized as a formality, in that the PTAs were structured but are not fully functional. The baseline data show that when asked, almost all principals reported that they held regular PTA meetings. When probed further, they indicated that for the most part, the majority of parents came to the PTA meetings. We suggest that better understanding of issues like this will provide invaluable planning information for the MOE and point to the ways by which PTA support and influence can be leveraged to the extent possible. Note that in some districts, the PTAs are not functional at all, whereas in others, Coaches have succeeded in reviving the PTAs. The EGRA team will continue to work with schools and Coaches to organize cluster-level PTA meetings and garner broader support for the EGRA efforts. We hope that this effort will extend to other non-EGRA discussions between schools and PTAs.

An important obstacle to the implementation of EGRA Plus is classroom "time on task." Some teachers' attendance is not regular. They come late, or leave early for various reasons such as second employment or going to the market. Schools in some rural areas are only open between 10:00 am and noon. On market days, some schools are closed as teachers *and* students go to the market. What is interesting is that attendance in public schools is highest during examination or testing periods, or when food is distributed. This is more pronounced in rural areas. Students often choose to work for companies in their

[12] Unpublished manuscript, available from the authors: Luis Crouch, lcrouch@rti.org; Marcia Davidson, Marcia.davidson@utah.edu

area rather than go to school, resulting in low student attendance and/or dropout. This is also the case with rural families; they keep their children to help on the farm. As a result, reading instruction seems to take place three or four times a week, whereas the MOE requested all teachers in the project to teach reading five times a week.

The combination of the barriers and obstacles above is certain to have had some negative effect on the delivery and implementation of the program. In sum, the actual teaching of reading by teachers took place primarily between mid-February and the last week of May 2009, when the midterm assessment commenced (hence the inclusion of Appendix A in this report). This equates to approximately 3.5 months of teaching, quite a limited amount of time for the treatments to take effect. EGRA Plus has learned and adapted to these challenges and has reanalyzed the academic calendar for 2009/2010. The calendar will be adjusted as a result of the lessons learned, and teachers will receive a more specific manual with daily lesson plans.

# 4. Sustainability and Scale-Up

Year 2 of EGRA Plus provides an opportunity to scale up the project and work to ensure sustainability. A component of the EGRA Plus: Liberia project is 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 present and trained, including two EGRA assessment workshops, three EGRA reading workshops, and one workshop on data analysis and reporting.

One of these reading workshops marked the beginning of more in-depth involvement of District Education Officers (DEOs) 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 will be fully involved in the training activities and the support provided to EGRA target schools. They were all trained instructional methods for reading during the project's refresher course that took place in August 2009. Between September and December 2009, each DEO will visit at least four schools together with Coaches. This will give them an opportunity to practice some of their skills in teaching reading as well as to provide pedagogic support to teachers. At the end of Semester 1, they will attend a retreat/refresher training in December 2009, once again, together with Coaches. The same number of visits to EGRA schools is planned for the second semester (January–May 2010). Finally, DEOs will be invited to attend the final reading policy workshop planned for the end of the project.

At the national level, the capacity building of MOE staff will also be deepened to allow more opportunities for turning newly acquired knowledge into practice. Dozens of MOE staff have learned how to assess student reading, and most of them were also deployed for data collection. In Year 2 of the project, they will partner with project staff to learn how to calibrate (equate) instruments, be co-facilitators of assessor training, supervise

data collection, do data entry and analysis, supervise the implementation of reading intervention, and assist with the training and support provided to teachers.

The goal of these capacity-building efforts is to provide a foundation for expansion of the 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 on how to ensure that all children in Liberia can experience the same increases in their reading skill early.

# 5. EGRA Subtask Descriptions

This section briefly introduces the various subtasks, so that the analysis below is meaningful. The EGRA tool consists of a variety of subtasks, 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. *Print orientation*: awareness of the direction of text, and the knowledge that a reader should read down the page.
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 one minute, which saves time and also prevents children from having to spend time on something they are having a hard time with.
3. *Phonemic awareness*: awareness of how sounds work with words. This is generally considered a pre-reading skill, and 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, starts with a different sound (e.g., *ball*, in “mouse, ball, moon”).
4. *Familiar word recognition*: ability to read high-frequency words. This assesses whether children can process words quickly. It is timed to one minute.
5. *Unfamiliar or nonsense word recognition*: 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 subtask assesses the child’s ability to “decode” words fluently. It is timed to one minute.
6. *Connected text oral reading fluency*: ability to read a passage, about 60 words long, that tells a story. It is timed to one minute.
7. *Comprehension in connected text*: ability to answer up to five questions based on the proportion of the passage read.
8. *Listening comprehension*: being able to follow and understand a simple oral story. This 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 pre-reading skill.

## 5.1 Adjustments to the EGRA Instruments for Midterm Assessment

In order to prevent "teaching to the test," or memorization, the midterm assessment used different word lists and passages. Although every effort was made to calibrate the difficulty ex ante using various analyses in May 2009, such as Spache analysis, this type of ex ante calibration typically is not good enough, in our experience. Thus, in addition to the ex ante calibration, we also conducted an empirical or statistical calibration. In this section we discuss both the Spache analysis and the calibration.

The advantage of EGRA as a tool for measuring reading fluency is that it is an assessment of skills and not the content. For every EGRA that is done in a particular setting, the content of the EGRA tests is entirely changed. In other words, the story used for assessing student performance in reading connected text is never the same, which eliminates the possibility of "test leaking" and "teaching to the test." However, for EGRA Plus: Liberia, it is important that the data collected be comparable from baseline to midterm to final assessments if we are to make any inferences about improvements in student performance over time. To this end, the EGRA team calibrated (or equated) the midterm assessment student instrument to be of equal difficulty to the one used in November 2008 baseline, as follows:

- First, the new passage was developed by RTI's Reading Specialist, Dr. Marcia Davidson. She used the Spache readability online tool to determine the grade level of this passage. It was important that the new passage be as close as possible in terms of its difficulty level to the one used in the baseline.
- Once the passages were equated using Spache analysis, they had to be tested in a "live" setting. The EGRA team went to four schools in Monrovia and tested 80 students.[13] The sample of 80 children was independent of the sample of children in any of the project schools. Each student was asked to read both passages (baseline and midterm), and the time taken to read each passage was recorded. The order in which students were asked to read the passage was alternated in order to create a randomization effect (e.g., Student 1 read the old passage first and then the new passage, Student 2 read the new passage first and then the old, and so on until all 80 students were tested). Children in both grades 2 and 3, in several schools, were part of the sample.
- Once the 80 observations were collected, the data were entered and analyzed by the Task Coordinator (and, as noted, two assessments were excluded from the data entry). An analysis of the averages showed that in general, the correlation between the two (2008 and 2009) was excellent. But the analysis also confirmed that the levels of difficulty appeared slightly different. These differences were adjusted for during the analysis stage.

While letters and unfamiliar words were only reshuffled in the 78-observation instrument, we needed to include new familiar words. Dr. Davidson recommended calibrating these

[13] Two assessments were excluded from the final data entry and analysis as the data were incomplete, which resulted in an actual sample of 78 rather than 80 as had been intended.

as well, which the EGRA team did using the same approach described above. A fuller description of these analyses can be found in ***Appendix B.***

Other sections of the Student, Teacher, and Principal instruments were reviewed and adjusted jointly with the assessors during their training. Each question in the instruments was discussed and approved by all of the participants. Note that the Student, Teacher, and Principal instruments were vetted by the Liberian stakeholders in June 2008 at the time the pilot assessment was conducted, and in November 2008 at the time of the baseline assessment.[14] The same was done in June 2009 with the workshop participants.

## 5.2 EGRA Assessor Training

The training was organized for May 11–15, 2009, and it was facilitated by the Task Coordinator and EGRA Technical Coordinator. The training was also attended by the MOE EGRA coordination committee. 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 was 47, from which the 18 best assessors were selected. The total number of MOE staff trained at this training was 15. At the core of the training approach was the use of an interrater reliability tool for both training and selection of assessors. The idea behind this tool is that a person—usually the trainer—is chosen to represent the "gold standard." This person pretends to be a student and then intentionally makes a number of mistakes in a given instrument. The closer the assessor is to the gold standard, the better his/her performance is. In other words, if the trainer made four intentional reading errors, then all assessors should have caught the same mistakes.

This approach allowed the trainers to pinpoint the struggles that assessors were experiencing. For instance, they would mix up the sounds for the letters "m" and "n"; they often did not hear the word "the"; or they would not be accurate in marking the student sheet. The mistakes were discussed in plenary, which allowed all participants to compare their mistakes together and explain why some mistakes were being made.

The interrater reliability tool also assisted trainers in the final selection of assessors. After the candidates completed one interrater exercise, the instruments they had filled out were collected by trainers and scored. Scoring consisted of adding up mistakes that the candidate assessors had made across different subtasks. The added scores then were used for ranking the assessors—the lower the number of mistakes, the higher the chance that candidate would be selected. Two such tests were used to rank the candidate assessors. The EGRA team looked at both tests and chose the better of the two performances for ranking the assessors. One of these tests was unannounced, as a way of reducing pretest anxiety. The other test was announced, to avoid the possibility that some assessors would

[14] The three instruments (student, teacher, principal) for both baseline and midterm are available from the EdData II website, https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=159 and https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=284.

think it was an exercise and would not perform as well as if they had been warned. At the end, the candidate assessors were ranked and chosen based foremost on their performance, but also on whether they had participated in the baseline assessment. The selection process was done together with the MOE's EGRA coordination committee.

## 5.3 EGRA Data Collection

Data collection commenced on May 25, 2009, and lasted four weeks. Nine teams were formed, each consisting of two members. They were both tasked with conducting assessments and had the same responsibilities. One of them was chosen to be the team leader and to make sure that all subtasks were completed. Overall, schools were cooperative and open to assessment based on their familiarity with the project. Teams needed to put in extra effort in a few cases, however, where schools were closed due to the rainy season. Together with the EGRA team, the assessors managed to reach their schools. Out of the total 180 schools, 179 EGRA schools were assessed.

It needs to be mentioned, however, that the data collection faced a major obstacle in one district. Due to a number of misunderstandings by the community, which were based only on the color of the vehicle, they concluded that one of our teams was a group of kidnappers, and as a result the community attacked our vehicle. Fortunately, the assessors received only minor injuries. The EGRA team and the MOE (the EGRA project is very grateful for MOE's assistance in this matter) conducted an inquiry into what happened and a report was provided to USAID. In discussions with the community and authorities, it was agreed that the incident was a huge misunderstanding.

The instruments were submitted by the assessors in mid-June. The EGRA team checked every instrument and checked the assessors' scoring. At the same time, the EGRA Task Coordinator checked instruments for missing data and found that there were very few instances of missing data.

## 5.4 EGRA Data Entry

An EGRA data entry application was developed in June 2008 by Mr. Farwenee Dormu of the MOE, with guidance and support from RTI. According to Mr. Dormu, the EGRA database was the first database that the MOE had developed since the end of the conflict in Liberia. Mr. Dormu was grateful to be given an opportunity to engage in this important work and to use it to build the capacity of the EMIS staff. Lessons learned from the November 2008 baseline were used to develop a brief manual for data entry. Entry of the baseline EGRA data was completed at the end of January 2009. Both the Principal Investigator and the Task Coordinator determined that compared to June 2008, the accuracy of data entry was greatly improved. For the midterm assessment, RTI developed a data entry application using Visual Basic that reduced the time for data entry to a third of what would have been needed previously.

# 6. Research Design

***Table 3*** below shows the achieved sample for both the baseline and midterm assessment. Notably, only one school that was included in the baseline sample was not retained in the midterm assessment: one of the control schools. This table also shows the sample of children in the baseline and midterm assessments 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, while the numbers of children in full treatment schools were quite close. Note that the samplings for each assessment, baseline and midterm, were done randomly and independent of each other. In other words, no attempt was made to resample children assessed in the baseline at the midterm. However, it is possible that children in the baseline assessment would also be found in the midterm assessment, although since children's names were not used, it is impossible to tell with any certainty. Table 3 also shows that the impact analysis contained in this report is based on 2970 baseline and 2805 midterm participants, for a total of 5775 children, a substantial sample size for this type of analysis.

**Table 3. Achieved EGRA sample for baseline and midterm, by treatment group, for schools and students**

| | | Treatment | | | |
|---|---|---|---|---|---|
| | | Control | Full | Light | Total |
| Schools | Baseline | 57 | 59 | 60 | 176 |
| | Midterm | 56 | 59 | 60 | 175 |
| Students | Baseline | 951 | 980 | 1030 | 2970 |
| | Midterm | 874 | 973 | 958 | 2805 |

More details about the sample used in this analysis can be found in ***Table 4*** below. Disaggregating by baseline/midterm assessment, the gender, grade, and treatment status of all of the children can be found. Interestingly, while there were more boys than girls in the baseline sample (1626 and 1331, respectively), there were more girls than boys in the midterm assessment (1310 and 1452, respectively). If the findings from the baseline[15] hold, then this suggests that analyses should be done with control variables for gender, such that the differential sampling by gender does not skew the results. This is particularly true when we consider the treatment status of children's schools. Where light and control schools were more heavily male than female in the baseline, these same schools were now

[15] RTI International. (2009, April). *EGRA Plus: Liberia data analytic report: EGRA Plus: Liberia baseline assessment.* Report prepared for USAID/Liberia under EdData II Task 6, Contract No. EHC-E-06-04-00004-00. Research Triangle Park, North Carolina: RTI. https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=286

more female than male in the midterm assessment. It is important to note that these variations are logical given the sampling method, and are not of a concern as long as gender and treatment status control variables are included in latter analyses. The columns to the right in Table 4 below are indicative of grade level. Note that in both midterm and baseline assessments, there were more grade 2 than grade 3 children, both boys and girls. This is plausibly as 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 4. Achieved sample, by baseline, midterm, grade, and treatment group**

| | | Treatment | | | | Level | | |
|---|---|---|---|---|---|---|---|---|
| | Gender | Control | Full | Light | Total | Grade 2 | Grade 3 | Total |
| Baseline | Boys | 543 | 512 | 571 | 1626 | 816 | 801 | 1617 |
| | Girls | 406 | 462 | 463 | 1331 | 720 | 603 | 1323 |
| | Total | 949 | 974 | 1034 | 2957 | 1536 | 1404 | 2940 |
| Midterm | Boys | 397 | 474 | 439 | 1310 | 713 | 591 | 1310 |
| | Girls | 464 | 477 | 511 | 1452 | 740 | 694 | 1452 |
| | Total | 874 | 973 | 958 | 2805 | 1466 | 1310 | 2805 |

Basic descriptive statistics for both the baseline study (columns to the left) and midterm assessment (columns to the right) can be found in ***Table 5*** below. Simple comparisons of means across assessments gives some hints of how much children have learned during the seven intervening months. On the midterm, children scored 19.04 letters per minute higher, 7.12 familiar words per minute more, and .23 unfamiliar words per minute more. Interestingly, children scored lower on the midterm on oral reading fluency, by 1.74 words per minute, and lower on reading comprehension, by 3.96 percentage points. It is important to note that there is, of course, a connection between oral reading fluency scores and reading comprehension, since children are asked comprehension questions about the oral reading passage. That said, it seems that either children did less well on the midterm assessment story and associated comprehension questions, or that the story was more difficult on the midterm assessment. This discussion is given a thorough treatment and analysis in subsequent sections. In any case, care must be given to performing a simple comparative analysis across baseline and midterm assessments given the disparate numbers by gender of participants as well as the differing class sizes across assessments. Even so, when we compare maximum scores, children performed a bit better at the midterm than they did at baseline. Of course, this comparison subsumes any differences between treatment and control groups, which are analyzed more thoroughly elsewhere in the report.

A final note of interest regarding the descriptive statistics from the baseline and midterm assessments is the relatively large standard deviations, particularly for the familiar word, unfamiliar word, and oral reading fluency scores. It appears that there was a great deal of variation in all of these measures, and a table in Section 9 below discusses how many of the children in these subtasks were either discontinued or scored very low.

**Table 5. Descriptive statistics for baseline and midterm assessment**

| Item | Baseline November 2008 | | | | | Midterm June 2009 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *N* | Mean | Standard deviation | Min | Max | *N* | Mean | Standard deviation | Min | Max |
| Letter naming fluency | 2971 | 61.16 | 25.30 | 0 | 180 | 2712 | 80.20 | 26.51 | 0 | 230.77 |
| Phonemic awareness | 2971 | 3.49 | 2.29 | 0 | 10 | 2805 | 4.18 | 2.63 | 0 | 10 |
| Familiar word fluency | 2946 | 9.26 | 13.90 | 0 | 76.67 | 2694 | 14.81 | 16.24 | 0 | 96.79 |
| Unfamiliar word fluency | 2950 | 2.24 | 6.02 | 0 | 53.6 | 2696 | 2.47 | 5.91 | 0 | 74.54 |
| Connected text fluency | 2952 | 19.58 | 20.03 | 0 | 96.58 | 2648 | 26.00 | 25.22 | 0 | 174.84 |
| Reading comprehension | 2971 | 25.01 | 24.23 | 0 | 80 | 2648 | 21.05 | 23.83 | 0 | 100 |
| Listening comprehension | 2986 | 33.49 | 20.49 | 0 | 60 | 2713 | 74.45 | 30.34 | 0 | 100 |

While Table 5 above presents descriptive for both the baseline and midterm assessments, ***Table 6*** below allows an analysis of the combination of the baseline and midterm assessments. This table is useful in that this combined data set is used for analysis later in the report.

**Table 6. Descriptive statistics for baseline and midterm, combined data set**

| Item | Combined baseline and midterm | | | | |
|---|---|---|---|---|---|
| | *N* | Mean | Standard deviation | Min | Max |
| Letter naming fluency | 5683 | 70.25 | 27.57 | 0 | 230.77 |
| Phonemic awareness | 5776 | 3.83 | 2.49 | 0 | 10 |
| Familiar word fluency | 5640 | 11.91 | 15.31 | 0 | 96.79 |
| Unfamiliar word fluency | 5646 | 2.35 | 5.97 | 0 | 74.54 |
| Connected text fluency | 5600 | 22.62 | 22.85 | 0 | 174.84 |
| Reading comprehension | 5619 | 23.14 | 24.12 | 0 | 100 |
| Listening comprehension | 5699 | 52.99 | 32.81 | 0 | 100 |

# 7. EGRA Reliability Analysis

In order to examine whether and how the subtasks in the Liberian Early Grade Reading Assessment at the midterm were reliable, and critically, whether it can be argued that they test an underlying skill, the reliability tests below were performed. Initially, simple Pearson's bivariate correlations were examined and are presented in ***Table 7*** below. Note that the findings are remarkably similar to those of the baseline assessment (see p. 17 of the baseline assessment report), largely because this version of the assessment was adapted from the baseline assessment. Note that the lowest correlations are between the listening comprehension and phonemic awareness subtasks and the rest of the subtasks. There are a couple of potential reasons for this. First, it appears that these subtasks assess different skills from the rest of the construct. Second, neither of these subtasks is timed, which means that achievement is less a function of speed, which differentiates them from the rest of the assessments.

**Table 7. Pearson's correlations for EGRA subtasks**

| | Letter naming fluency | Phonemic awareness | Familiar word fluency | Unfamiliar word fluency | Connected text fluency | Reading compre-hension | Listening compre-hension |
|---|---|---|---|---|---|---|---|
| Letter naming fluency | 1.00 | | | | | | |
| Phonemic awareness | 0.36*** | 1.00 | | | | | |
| Familiar word fluency | 0.58*** | 0.32*** | 1.00 | | | | |
| Unfamiliar word fluency | 0.32*** | 0.25*** | 0.56*** | 1.00 | | | |
| Connected text fluency | 0.60*** | 0.35*** | 0.88*** | 0.56*** | 1.00 | | |
| Reading comprehension | 0.46*** | 0.38*** | 0.62*** | 0.42*** | 0.72*** | 1.00 | |
| Listening comprehension | 0.39*** | 0.30*** | 0.29*** | 0.18*** | 0.35*** | 0.38*** | 1.00 |

After the correlational matrix analysis, a Cronbach's alpha reliability test was performed in order to assess whether the entire subtask was representative of an underlying construct, hopefully early grade reading skills. Not surprisingly, the lowest item–test correlations were found for both the listening comprehension and phonemic awareness subtasks, which mirrors what was found in the correlational analysis in ***Table 8*** below. The Cronbach's alpha of the entire test is still 0.85, which is within the accepted range for

a low-stakes assessment such as EGRA and is in line with what was found in the baseline report.

**Table 8. Cronbach's alpha statistics for midterm assessment**

| Item | Item–test correlation | Item–rest correlation | Average inter-item correlation | Alpha |
|---|---|---|---|---|
| Letter naming fluency | 0.74 | 0.62 | 0.44 | 0.82 |
| Phonemic awareness | 0.62 | 0.43 | 0.49 | 0.85 |
| Familiar word fluency | 0.84 | 0.76 | 0.40 | 0.80 |
| Unfamiliar word fluency | 0.65 | 0.51 | 0.47 | 0.84 |
| Connected text fluency | 0.88 | 0.82 | 0.39 | 0.79 |
| Reading comprehension | 0.79 | 0.69 | 0.42 | 0.81 |
| Listening comprehension | 0.57 | 0.41 | 0.49 | 0.85 |
| Overall test | | | 0.44 | 0.85 |

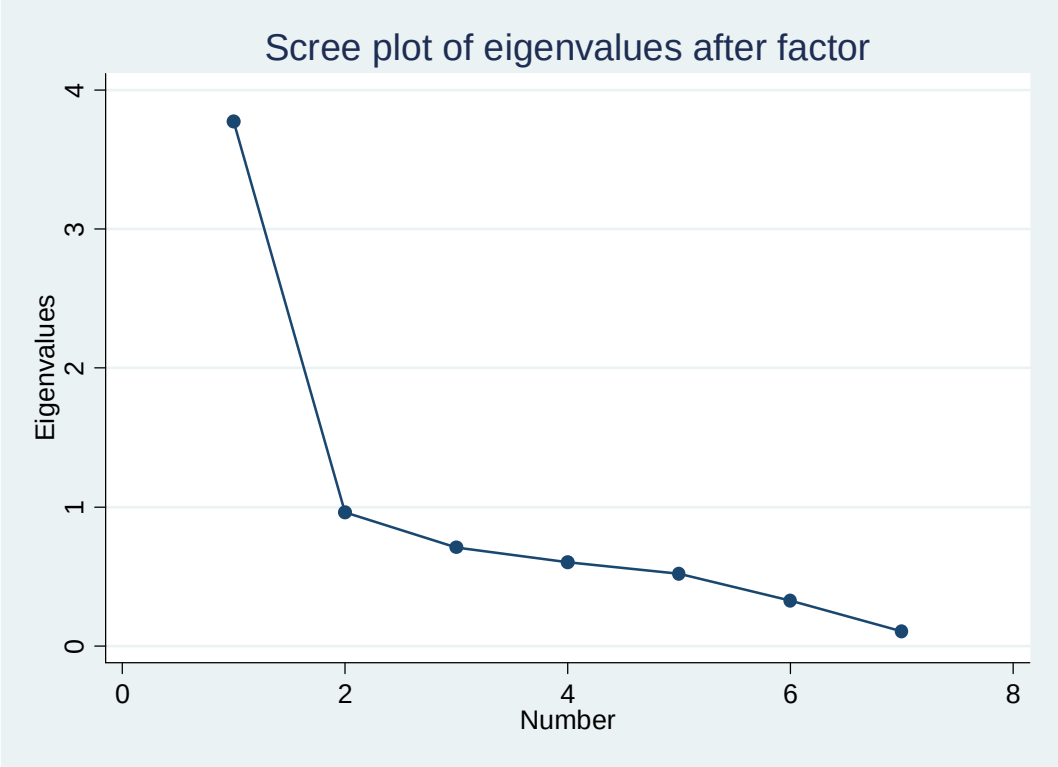
Following the Cronbach's alpha analysis above, a principal components analysis was performed to investigate, once again using another method, whether there was an underlying construct that the subtasks were evaluating. The principal component loaded highly on all of the subtasks, although (once again) the loadings were lower for phonemic awareness and listening comprehension. The details are found in the first column of ***Table 9*** below. The second column shows the unique contribution of each subtasks, and of particular interest is that both phonemic awareness and listening comprehension were adding unique and important information to the entire assessment.

**Table 9. Principal component analysis for early reading component**

| Principal component 1 loading | | Uniqueness of each component | |
|---|---|---|---|
| Letter naming fluency | 0.74 | Letter naming fluency | 0.46 |
| Phonemic awareness | 0.54 | Phonemic awareness | 0.71 |
| Familiar word fluency | 0.88 | Familiar word fluency | 0.23 |
| Unfamiliar word fluency | 0.65 | Unfamiliar word fluency | 0.58 |
| Connected text fluency | 0.91 | Connected text fluency | 0.17 |
| Reading comprehension | 0.80 | Reading comprehension | 0.36 |
| Listening comprehension | 0.52 | Listening comprehension | 0.73 |

Principal components analyses are often followed by a creation of a visual screeplot to determine how much of the variation of the total assessment (in this case, EGRA) is explained by the new principal component that was created with the characteristics of Table 9 above. ***Figure 1*** shows that the first component explains 3.77 eigenvalues of variation. In short, this means that nearly half of the entire variation of all the subtasks is subsumed within this new component, which can be argued to represent early grade reading skill. The second principal component in Figure 1 below represents less than one eigenvalue, which means that the first principal component does a good job of identifying the underlying construct. This bodes well for our ability to argue that the set of subtasks estimates the underlying skill well enough, and mirrors the findings in the baseline report.

**Figure 1. Screeplot of eigenvalues for principal components analysis**




# 8. Passage and Word Calibration

In this section, we present the calibration process used to equate the baseline and midterm assessment oral reading fluency story and the familiar word subtask (see details in ***Appendix B***). For the main body of the text, it is sufficient to share the adjustments for the analyses. The midterm results are to be adjusted as follows:

- Connected test fluency in the midterm passage should be multiplied by 1.26 to make it comparable to fluency in the baseline passage.

- Familiar word fluency in the midterm list should be multiplied by 0.93 to make it comparable to fluency in the baseline list.

# 9. Analysis of Discontinued Assessments

While the descriptive statistics above and the fuller analysis below provide several opportunities to compare the achievement of children in different treatment groups, an analysis of the discontinued assessments provides another take on the impact of the program. In several subtasks in the EGRA, a subtask is discontinued when the child reaches the stop rule, designed so that a child completely overmatched by a subtask does not have to endure the entire subtask, getting item after item incorrect. For letters, familiar word, unfamiliar, and oral reading fluency, the stop rule is that the child gets every item on the first line incorrectly. For phonemic awareness, the stop rule is when a child answers the first five items incorrectly. In all cases, discontinued subtasks show the subset of children who can be characterized as nonreaders. Comparing the numbers of discontinued students across the control, full, and light treatment groups allows us to determine whether the program is able to help those children who have had very limited success in reading skills.

***Table 10*** below presents this analysis, and shows that, for the most part, boys were more likely to discontinue than girls, which is surprising given that girls performed less well than boys. When we compare the treatment groups, the percentage of children who discontinued in control schools was higher than in full or light treatment, for each of the five discontinuable subtasks. For example, less than 4% of full treatment children discontinued phonemic awareness, while 11.8% of control children did. Similarly, for familiar words, 18.6% of control children discontinued while only 10.2% of full treatment children discontinued. For unfamiliar words, the numbers of discontinued students was very high, with little difference between control and light treatment students. On the other hand, full treatment students were 10% less likely to discontinue. This suggests that the program is helping a significant percentage of the lowest students access the decoding skills necessary for unfamiliar words. For oral reading fluency, 25.4% of control, 15.3% of full, and 18.7% of light treatment students discontinued. Across the subtasks, there remains a reasonably sized gap between full treatment and control discontinued, and a smaller gap between light treatment and control. It appears that the program helps some of the very lowest students.

**Table 10. Discontinued subtasks, by treatment status and gender (midterm)**

| | Control | Full Treatment | Light Treatment | Boy | Girl | Total |
|---|---|---|---|---|---|---|
| Letter naming fluency | 11 (1.2%) | 4 (.4%) | 6 (.6%) | 12 (.9%) | 9 (.6%) | 21 (.8%) |
| Phonemic awareness | 101 (11.8%) | 36 (3.9%) | 47 (5.0%) | 93 (6.6%) | 86 (6.7%) | 184 (6.7%) |
| Familiar word fluency | 157 (18.6%) | 94 (10.2%) | 132 (14.2%) | 231 (18.3%) | 149 (11.9%) | 383 (14.2%) |
| Unfamiliar word fluency | 686 (81.1%) | 652 (70.8%) | 740 (79.5%) | 1057 (83.8%) | 993 (70.9%) | 2078 (77.0%) |
| Connected text fluency | 212 (25.4%) | 139 (15.3%) | 169 (18.7%) | 306 (24.6%) | 209 (15.3%) | 520 (19.6%) |

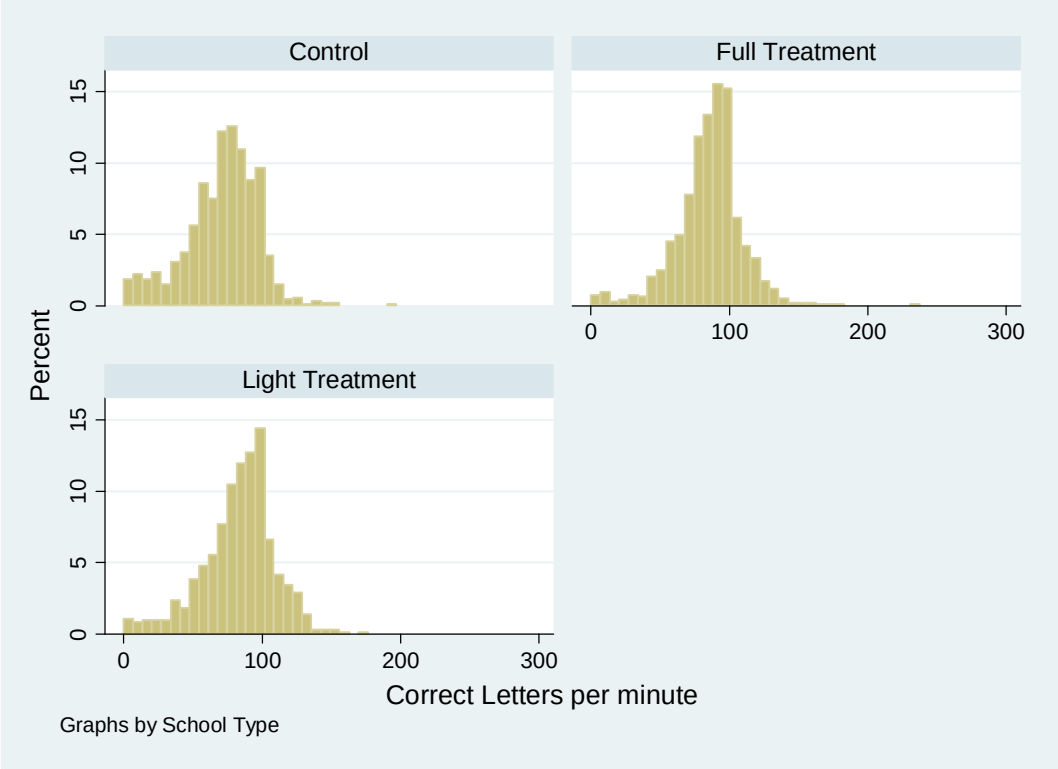
# 10. Subtask Figure Analysis

In this section, several figures are presented, created to illustrate the impact on achievement of the program, as measured at the midterm. This section is organized by subtask and looks at which of several variables are predictive of reading outcomes, including grade and gender.

## 10.1 Letter Naming Fluency

***Figure 2*** below shows the scores of control, full treatment, and light treatment children on the letter identification fluency subtask. Note that each bar presents the percentage of children from that treatment group that scored a particular number of letters per minute. Visual inspection shows fewer children who scored 0 or close to 0 in the full treatment group than in either the control or light treatment groups. Similarly, more children scored nearly 100 in the full and light treatment groups than the control group. In general, the full treatment group had a nearly normal distribution, while the control and light treatment groups had a slight leftward skew.

**Figure 2. Histograms comparing letter naming fluency scores, by treatment group**



In ***Figure 3,*** a gender comparison can be made for letter naming fluency achievement. Note that more boys scored at the very bottom of the distribution. However, visual inspection suggests that boys also had more scores to the right of 100 letters per minute. This counterintuitive set of findings suggests that boys were more likely found at both the bottom and the top of the distribution. Analysis of the grade differences show clearly what would be expected: that there were many fewer low scores in grade 3, and that the peak was much higher, near 100 letters per minute.

**Figure 3. Histograms comparing letter naming fluency scores, by gender (left) and grade (right)**

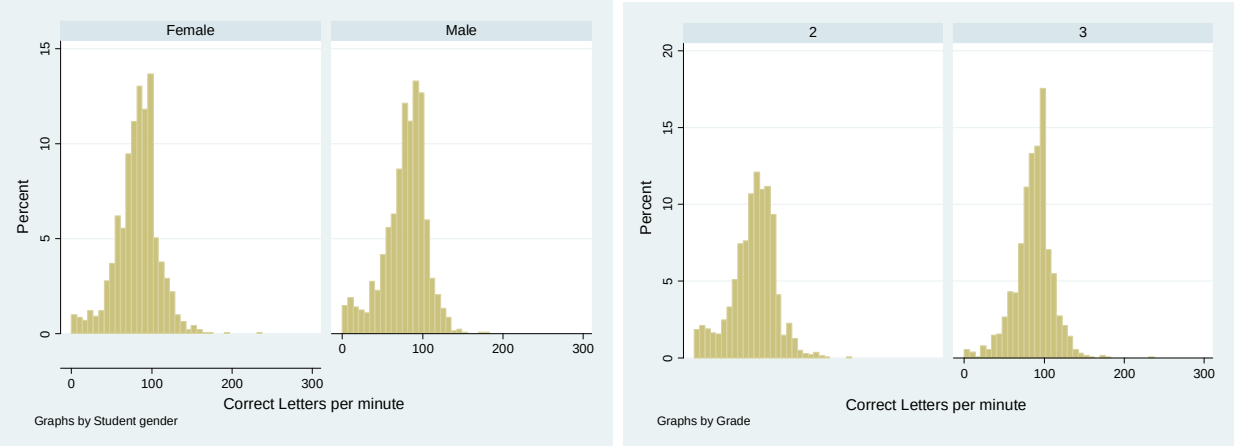
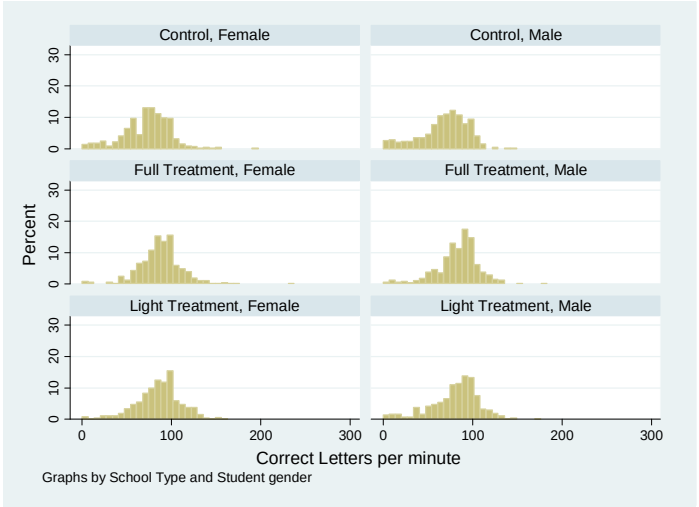


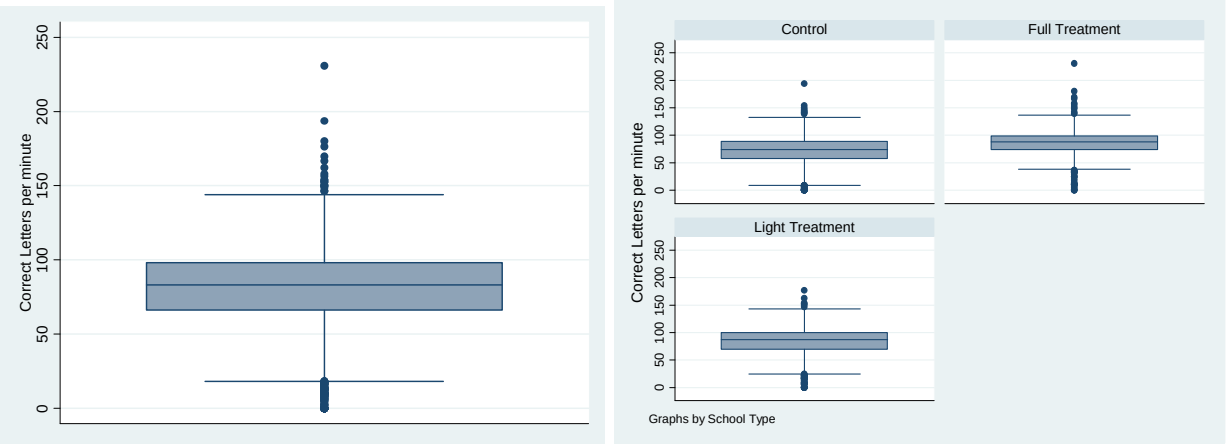

***Figure 4*** below compares letters per minute scores for boys and girls, by treatment groups. It shows, quite clearly, that the scores of boys and girls in control schools were centered to the left of both the light treatment and full treatment children (both boys and girls). This suggests that the program is having an impact for both genders, although a bit more modest for students in the light treatment group.

**Figure 4. Histograms comparing letter naming fluency scores, by treatment and gender**



The box plots below ***(Figure 5)*** present the overall achievement of children on the letters per minute subtask, on the left, and the letters per minute scores disaggregated by treatment status, on the right. There is a relatively small spread of the 25th-percentile, mean, and 75th-percentile scores, all of which fell between 60 and 100 letters per minute. On the other hand, the 10th- and 90th-percentile scores are quite widely spread, at around 20 and 150 letters per minute, respectively. Comparing the box plots of the various treatment groups reveals that the mean score for full treatment children is comparable to the 75th percentile score for control children. The 10th percentile for the full treatment group is much higher than that of the control children, which is quite near to 0. This reflects the discussion above regarding the ability of the EGRA Plus program to impact the scores of the lowest achieving children.

**Figure 5. Box plots comparing letter naming fluency overall (left) and by treatment (right)**



## 10.2 Phonemic Awareness

This subsection of the report investigates the figures produced to analyze the impact of the EGRA plus program on phonemic awareness scores. Note that these figures are of a discrete outcome measure, which explains the bars for each of the scores. In ***Figure 6*** below, in the graphic on the left, the 16% of children who discontinued this subtask are represented by the tall bar at the zero mark. The rest of the scores are nearly normally distributed, with the majority of children falling between 3 and 7 sounds correctly identified. The graphic on the right compares the achievement of boys and girls. Slightly more boys scored 4 or more sounds correct than did girls.

**Figure 6. Histograms comparing phonemic awareness scores overall (left) and by gender (right)**

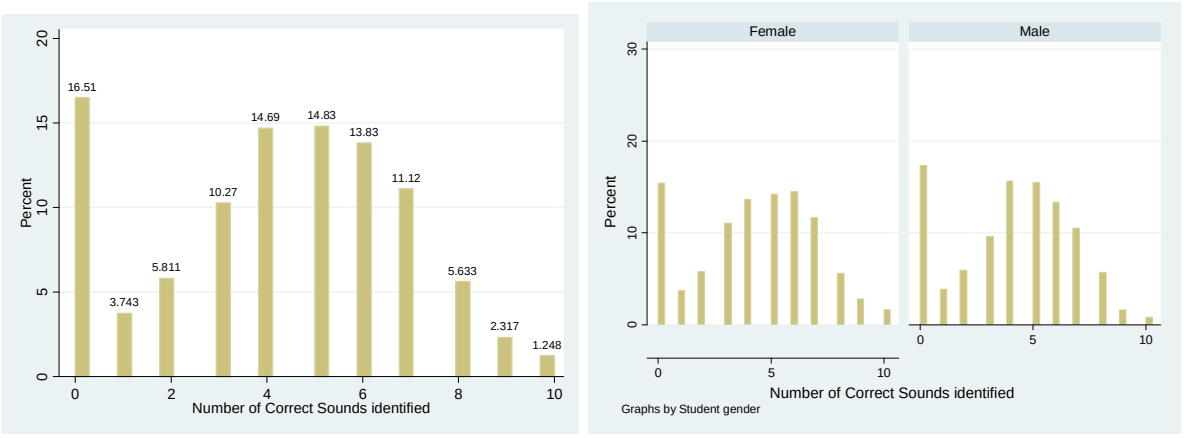
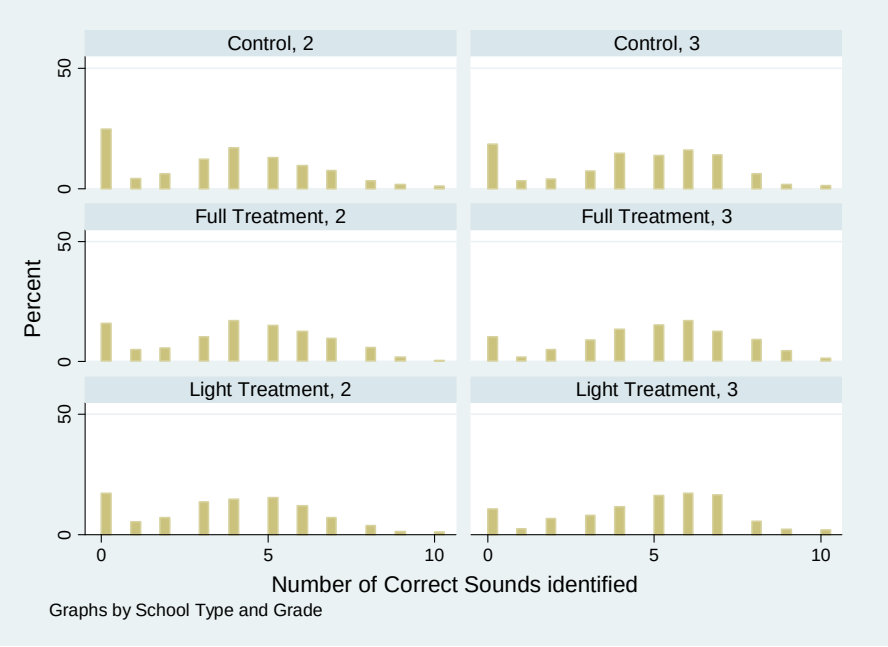


***Figure 7*** below disaggregates achievement on the number of sounds identified by grade and treatment group. Comparing the grade 2 children with the grade 3 children shows clearly that children were less likely to score 0 and more likely to score 5, 6 or 7 in grade 3. Comparing treatment groups, this figure also shows that full and light treatment children in grade 3, in particular, were more likely to score 5, 6 or 7, than were children in control schools. The relationships are not as clear in grade 2.

**Figure 7. Histograms comparing phonemic awareness scores, by grade and treatment**



### 10.3 Familiar Word Fluency

When we analyze the results of the program on the impact on the number of familiar words that children could identify in one minute, ***Figure 8*** shows the relatively large impact of grade when we compare the graphs on the left (grade 2) with those on the right (grade 3). Between grade 2 and grade 3, for all three treatment groups, there were fewer lower scores, particularly those centered around 0. The treatment effect is easier to see in grade 3, since it is evident that more of the children in full treatment scored farther to the right on the familiar words per minute subtask.

**Figure 8. Histograms for familiar word naming fluency, by treatment and grade**

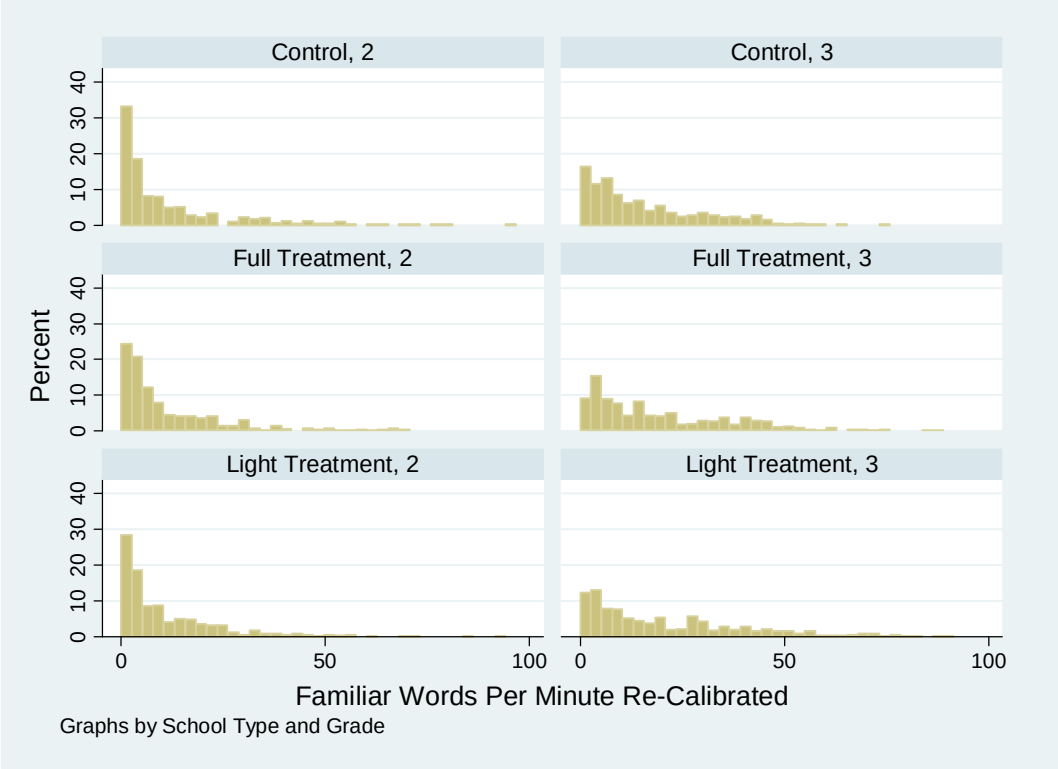
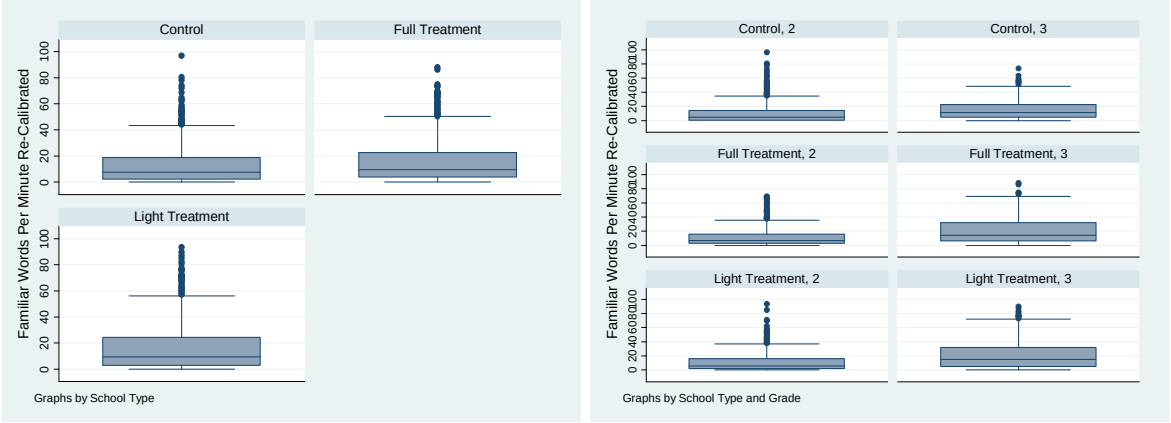


***Figure 9*** shows the differences between the treatment groups, and in the box plots on the right, the differences between the treatment groups, disaggregated by grade. The 75th and 90th percentile scores are clearly higher for full and light treatment schools than for control schools. Similarly, the means appear to be higher, though all three treatment group scores are clustered near zero, making it difficult to differentiate the scores visually. The grade effect is evident when we compare the graphs on the right: In all three treatment groups, grade 3 students outperformed grade 2. The full and light treatment scores for grade 3, in particular, were much higher (at the 25th percentile, mean, and 75th percentile) than the scores for the control groups. This is an indication of the impact that EGRA Plus has had on familiar word fluency.

**Figure 9. Box plots comparing familiar word fluency by treatment (left) and treatment and grade (right)**



## 10.4 Unfamiliar Word Fluency

The descriptive statistics section above showed that the scores for unfamiliar words were quite low. This is borne out in ***Figure 10,*** which shows, on the left, the difference between grade 2 and grade 3 scores. Note that a full 80% of children in grade 2 scored 0 (were discontinued) on the subtask, while around 70% of grade 3 children were discontinued. The graph on the right shows how achievement on unfamiliar words differed for control and full treatment schools. The most notable aspect of this figure is that children in full treatment schools were less likely (by nearly 20% compared to control, and around 10% compared to light treatment) to be discontinued on this subtask. This had commensurate impacts on the spread of scores for full treatment, which had at least some children scoring up to 20 unfamiliar words per minute.

**Figure 10. Histograms depicting achievement on unfamiliar word fluency, by grade (left) and treatment status (right)**

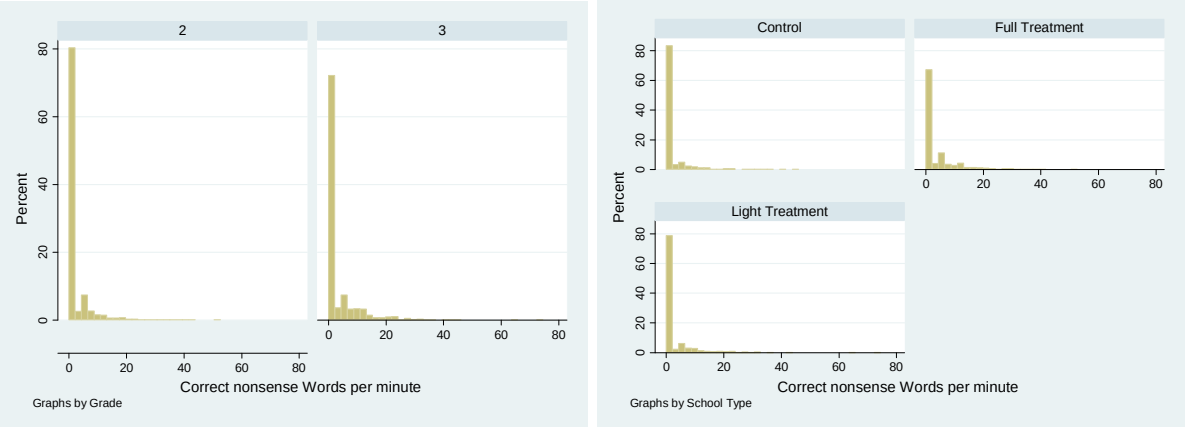


***Figure 11*** shows this point even more clearly. This bar charts on the left compare all three treatment groups and the two grades. They show, particularly when we compare control and full treatment children, that the EGRA plus program has helped children move from 0 scores to farther along the distribution. This is an important finding for equity: EGRA Plus not only is helping the brighter and more clever children expand their reading knowledge, but also is helping the lower achieving children increase their scores. The box plots to the right show another 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 is having an impact on the lowest achieving students.

**Figure 11. Histograms and box plots showing unfamiliar (nonsense) word recognition fluency, by treatment and grade**

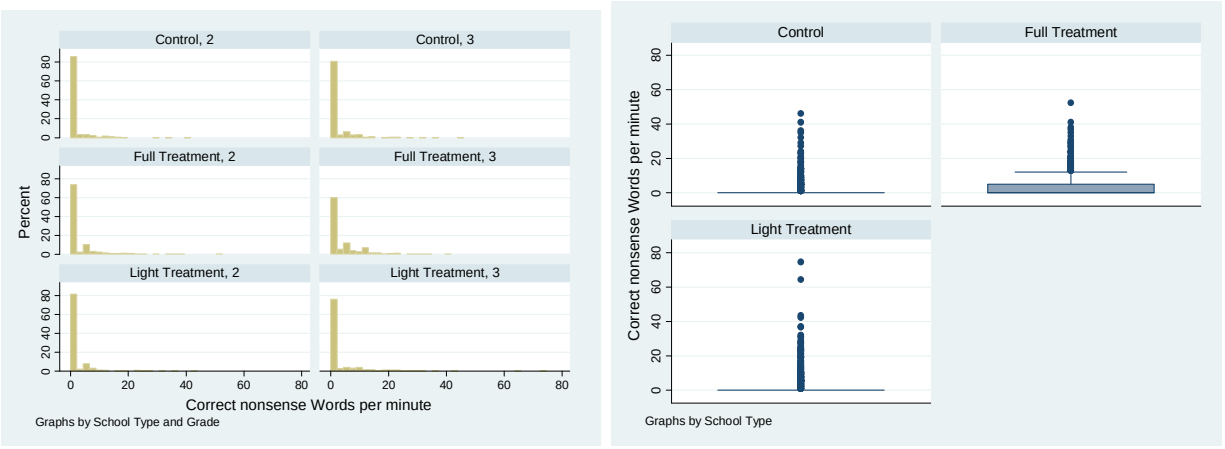
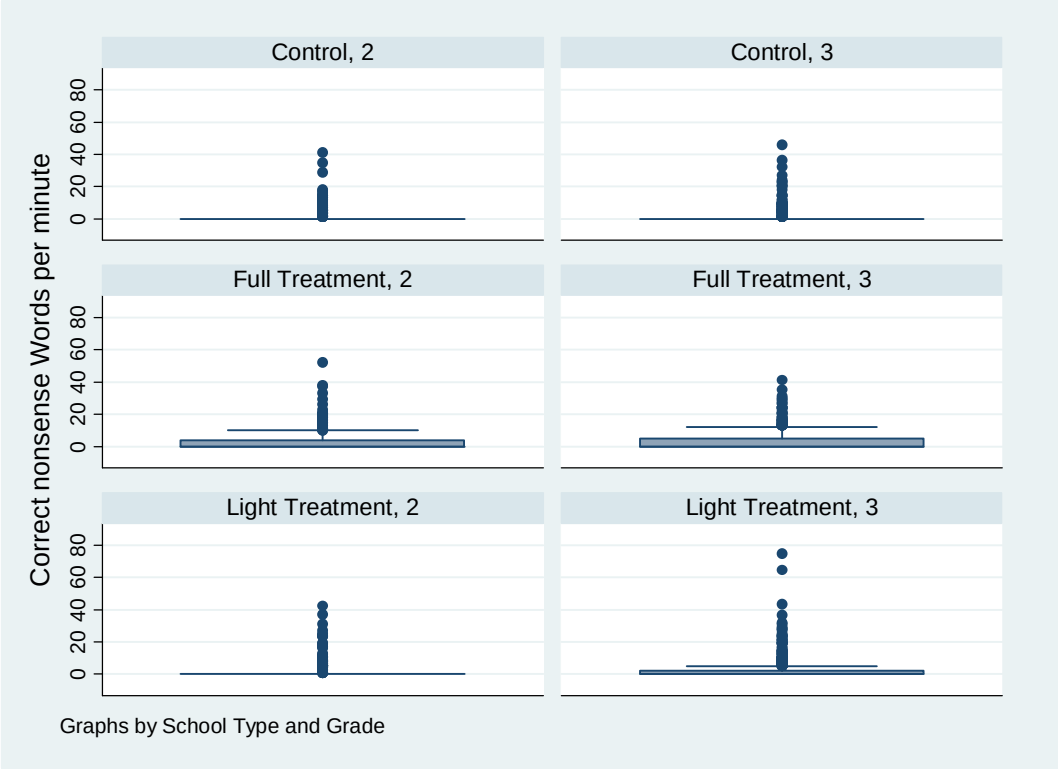


***Figure 12*** makes the same point more forcefully. While for control schools, in grade 2 and grade 3 there were no differences in 50th, 75th and 90th percentile scores, there were differences in full treatment schools, both for grade 2 and 3; and in light treatment schools as well, for grade 3. This means that grade 2 full treatment children outperformed grade 3 control children, which is, again, evidence of the impact of EGRA Plus.

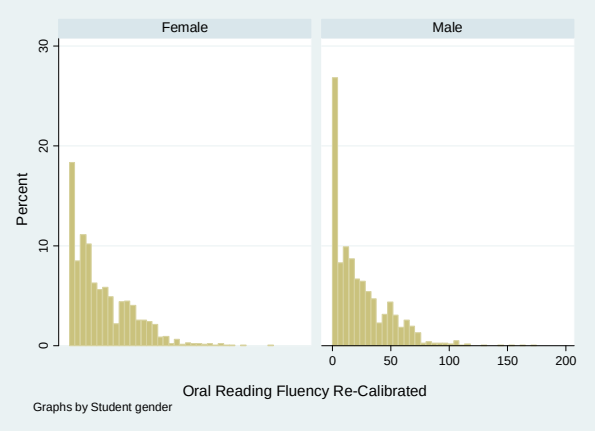
**Figure 12. Box plots showing achievement on unfamiliar word fluency, by grade and treatment**

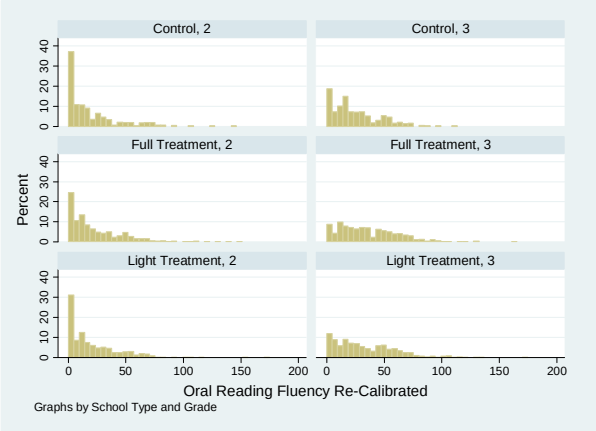


## 10.4 Oral Reading Fluency

In ***Figure 13,*** the graphs on the left show the differing achievement of boys and girls on the oral reading fluency measure. Once again, we find that while more boys scored 0 than girls (28% to 20%), boys' achievement was higher overall. This finding has been echoed in previous discussions and needs further research to unpack. The graphs on the right depict oral reading fluency scores by treatment group and grade. Note that while the control and light treatment histograms look very similar for grade 2 and grade 3, the full treatment group has fewer 0 scores for both grade 2 and 3, as well as a wider distribution of scores beyond the lowest scores. It appears that the EGRA plus program is helping children read more fluently, although the differences remain modest.

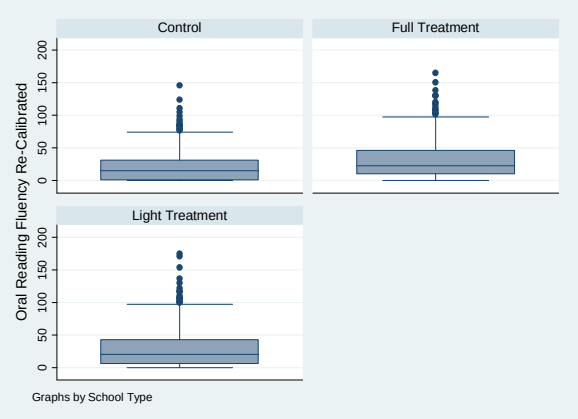
**Figure 13. Histograms showing oral reading fluency scores, by gender (left) and by grade and treatment status (right)**

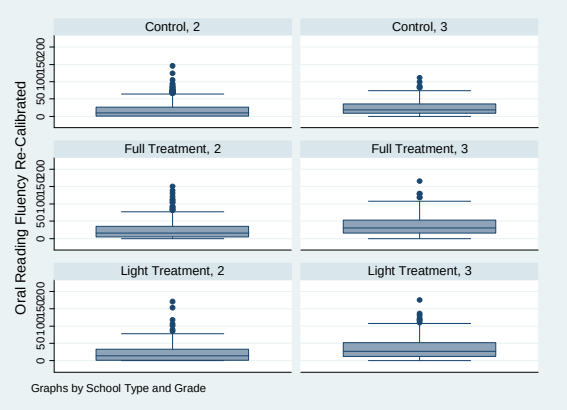


The box plots in ***Figure 14*** show how and whether the EGRA plus program is having an impact on oral reading fluency scores for children in treatment schools. It appears that that is the case. For example, in the graphs on the left, children in both full treatment and light treatment scores had a 10th-percentile marker, which means that there was a difference between the 10th- and 25th-percentile scorers for both groups. That is not the case for control schools, which means, once again, that both treatments are having an impact for the lowest level of readers. Similarly, it is notable that the mean score for the full treatment schools was very close to the 75th percentile for the control schools. These differences are made more obvious in the graphs on the right, where children in full and light treatment schools achieved higher scores at every percentile than did children in control schools.

**Figure 14. Box plots of oral reading fluency scores by treatment (left) and treatment/grade (right)**

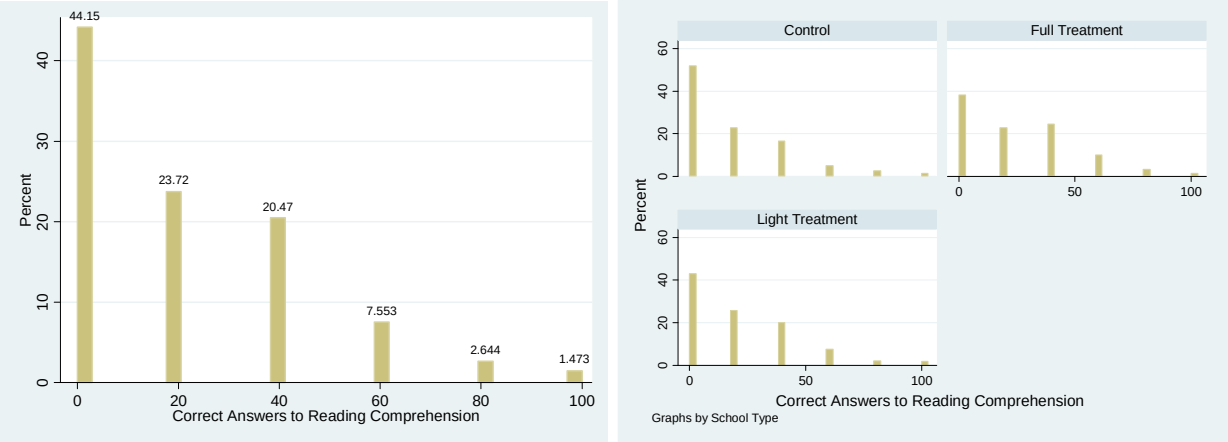


### 10.5 Reading Comprehension

***Figure 15*** below 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 peers since they outscored them on oral reading fluency, and the two subtasks are linked. This might not be the case, however, if the program only increases children's ability to read and sound out words, rather than synthesize and understand what they read. The graph on the left shows that most children (more than two thirds) were only able to answer one or two questions correctly. The graphs on the right show that, for full treatment schools, around 40% of children answered two or more questions correctly. For control schools, that number was less than 30%. This 10% of children in full treatment schools who were able to answer at least two questions correctly can be argued to be one of the major impacts of the EGRA plus program.

**Figure 15. Histograms showing reading comprehension scores overall (left) and by treatment status (right)**




# 11. EGRA Plus Program Impact

In order to determine whether the EGRA Plus program had an impact on student achievement in reading, it is important to compare the scores of children from the three groups of schools. For example, in ***Table 11*** below, scores are disaggregated by grade 2 and grade 3, as well as by control schools, full treatment schools, and light treatment schools. While the analysis below compares the achievement by these children with achievement in the baseline assessment, Table 11 shows whether there are differences in scores by control and treatment schools.

**Table 11. Midterm statistics and program impact, by grade**

| Item | School type | Grade 2 | | | | Grade 3 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *N* | Mean | Standard deviation | Percent difference from control | *N* | Mean | Standard deviation | Percent difference from control |
| Letter naming fluency | Control | 437 | 65.73 | 28.58 | | 404 | 76.7 | 23.3 | |
| | Full | 483 | 79.4 | 25.04 | 20.8% | 435 | 92.32 | 21.36 | 20.4% |
| | Light | 494 | 76.13 | 27.63 | 15.8% | 438 | 91.09 | 22.32 | 18.8% |
| Phonemic awareness | Control | 456 | 3.57 | 2.64 | | 409 | 4.3 | 2.69 | |
| | Full | 501 | 4.06 | 2.56 | 13.7% | 459 | 4.89 | 2.55 | 13.7% |
| | Light | 509 | 3.77 | 2.52 | 5.6% | 442 | 4.74 | 2.52 | 10.2% |
| Familiar word fluency | Control | 432 | 10.96 | 15.01 | | 402 | 15.19 | 14.02 | |
| | Full | 481 | 11.42 | 13.83 | 4.2% | 434 | 19.63 | 17.45 | 29.2% |
| | Light | 487 | 11.1 | 13.78 | 1.3% | 437 | 21.08 | 19.6 | 38.8% |
| Unfamiliar word fluency | Control | 434 | 1.33 | 4.16 | | 404 | 1.92 | 5.14 | |
| | Full | 479 | 2.58 | 5.84 | 94.0% | 434 | 4.01 | 6.51 | 108.9% |
| | Light | 491 | 1.74 | 4.84 | 30.8% | 433 | 3.26 | 7.92 | 69.8% |
| Connected text fluency | Control | 432 | 17.99 | 22.8 | | 396 | 24.21 | 20.65 | |
| | Full | 473 | 23.15 | 24.41 | 28.7% | 427 | 36.13 | 26.21 | 49.2% |
| | Light | 475 | 20.78 | 23.24 | 15.5% | 425 | 34.25 | 27.76 | 41.5% |
| Reading comprehension | Control | 432 | 15.37 | 23.5 | | 396 | 19.95 | 22.07 | |
| | Full | 473 | 19.37 | 21.91 | 26.0% | 427 | 29.37 | 25.83 | 47.2% |
| | Light | 475 | 16.08 | 21.91 | 4.6% | 425 | 26.64 | 24.34 | 33.5% |
| Listening comprehension | Control | 433 | 64.06 | 33.88 | | 405 | 70.32 | 33.28 | |
| | Full | 483 | 74.7 | 29.12 | 16.6% | 441 | 81.9 | 26.51 | 16.5% |
| | Light | 492 | 73.39 | 29.05 | 14.6% | 438 | 82.03 | 25.9 | 16.7% |

There were quite large differences on most subtasks. For example, grade 2 children in full treatment schools outperformed their control school counterparts by 13.67 letters per minute. The difference between light treatment and control children in grade 2 was only slightly smaller, at 10.30 letters per minute. This differences is reflected in grade 3, with full treatment schools identifying 15.52 more letters per minute, and light treatment schools identifying 14.39 more letters. Table 11 makes that comparison across the range of subtasks, and remarkably, on this midterm assessment, full treatment children outperformed control children on every subtask, for both grade 2 and grade 3. This provides evidence that the program has had an impact on student achievement. Similarly, light treatment children outscored control children on every single subtask, for both

grades. This suggests that the light treatment school program also has been effective with respect to student reading skills. Finally, when the full and light treatment schools are compared, in nearly every subtask, save grade 3 listening comprehension and grade 3 familiar words, full treatment school children scored higher than light treatment school children. Note that this analysis is a simple comparison of means, and does not take into account the standard errors that would allow us to determine whether these differences are statistically significant. That said, given the fuller technical discussion below, Table 11 shows that the program has had at least a moderate impact on student achievement across the range of subtasks.

## 11.1 Program Impact Comparing Grade 2 and Grade 3

It is important to be more specific and make some note of the magnitude of the differences between treatment and control schools. While below the impacts are presented in terms of effect sizes, here it is sufficient to note that full treatment schools and light treatment schools differed by at least a moderate amount.

- For letters, for both grade 2 and 3, the difference between treatment and control schools ranged between 15.8% to 20.8%; in all cases the treatment schools outperformed controls.
- The impacts were a bit smaller for phonemic awareness, ranging from 5.6% to 13.7%.
- For familiar words, the difference was larger at grade 3, with the differences for full (29.2%) and light (38.8%) quite large, although for grade 2 the gaps were between 1.3% and 4.2%.
- For unfamiliar words, notably since this was the subtask with which all children had the most difficulty, children in full treatment schools basically doubled the score of children in control schools (94.0% for grade 2, 108.9% for grade 3). This suggests that the program took many children who had no skills in unfamiliar word decoding and taught them some skills. Substantively, however, note that the differences were at the maximum, just over 2 words per minute, or still quite small.
- For oral reading fluency, differences ranged from 15.5% to 49.3%, with full treatment children doing better, and the impacts were higher at grade 3, where both full and light treatment schools differences were greater than 40%. This is a relatively large increase in student achievement, particularly taking into account the difficulties of program implementation.
- Commensurate with the differences in oral reading fluency, children in grade 3 treatment schools dramatically outperformed those in control schools on reading comprehension (47.2% and 33.5% for full and light treatment, respectively). The gaps were moderate in grade 2, much closer to the differences in oral reading fluency, as expected. For treatment groups at both grade 2 and grade 3, the gap

between both full and light treatment compared against control schools was around 16%.

In short, this provides strong and highly consistent evidence that children in treatment schools outperformed those in control schools on EGRA subtasks at the midterm, with a slightly larger achievement gap for those children in full treatment schools.

## 11.2 Program Impact Comparing Entire Baseline and Entire Midterm

***Table 12*** compares the entire baseline sample (from November 2008) with the entire midterm assessment sample (from June 2009) disaggregated by test item and treatment status (control, full or light). The columns to the left show the mean and standard deviation for each of these groups at the baseline. The next set of columns depicts the midterm scores for these same groups. The columns to the right show the program impact, described several ways. The first column, "Raw gains over baseline," shows the difference in scores between baseline and midterm as an absolute difference. The next column, "Raw increase over control," shows the difference in the gains between baseline and midterm less than the gains for the control group. This is the true program impact column. The next column, "Percent increase over baseline," changes the "Raw gains over baseline" column to a percent increase against the baseline score. This is reflective of the need from the Project Monitoring Plan to discuss the increase over baseline. The final column, "Effect size," takes the increase over control column and converts it to standard deviations for each of the subtasks, using Cohen's *d.* This is the column that includes the gains due to the treatment compared against the baseline, but converted to a comparable figure (standard deviations).

### Table 12. Comparing grade 2 and grade 3 baseline and midterm, with program impact

| Item | School type | Baseline, grades 2 and 3 | | | Midterm, grades 2 and 3 | | | Program impact | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | Mean | Standard deviation | *N* | Mean | Standard deviation | Raw gains over baseline | Raw increase over control | Percent increase over baseline | Effect size (SD) |
| Letter naming fluency | Control | 948 | 63.16 | 25.49 | 849 | 70.92 | 26.72 | 7.76 | | 12.29% | |
| | Full | 979 | 58.48 | 24.8 | 926 | 85.65 | 24.21 | 27.17 | 19.41 | 46.46% | 0.73 |
| | Light | 1036 | 61.81 | 25.42 | 937 | 83.23 | 26.31 | 21.42 | 13.66 | 34.65% | 0.52 |
| Phonemic awareness | Control | 951 | 3.41 | 2.35 | 874 | 3.85 | 2.7 | 0.44 | n/a | 12.90% | |
| | Full | 980 | 3.42 | 2.17 | 973 | 4.44 | 2.6 | 1.02 | 0.58 | 29.82% | 0.22 |
| | Light | 1039 | 3.61 | 2.37 | 958 | 4.22 | 2.57 | 0.61 | 0.17 | 16.90% | 0.06 |

| Item | School type | Baseline, grades 2 and 3 | | | Midterm, grades 2 and 3 | | | Program impact | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | Mean | Standard deviation | *N* | Mean | Standard deviation | Raw gains over baseline | Raw increase over control | Percent increase over baseline | Effect size (SD) |
| Familiar word fluency | Control | 940 | 8.31 | 12.6 | 842 | 13.01 | 14.65 | 4.7 | n/a | 56.56% | |
| | Full | 971 | 9.16 | 13.76 | 923 | 15.43 | 16.21 | 6.27 | 1.57 | 68.45% | 0.09 |
| | Light | 103 | 10.25 | 15.08 | 929 | 15.81 | 17.48 | 5.56 | 0.86 | 54.24% | 0.05 |
| Unfamiliar word fluency | Control | 947 | 1.68 | 4.72 | 846 | 1.62 | 4.65 | -0.06 | n/a | -3.57% | |
| | Full | 971 | 1.83 | 5.3 | 921 | 3.27 | 6.2 | 1.44 | 1.5 | 78.69% | 0.25 |
| | Light | 1027 | 3.16 | 7.46 | 929 | 2.45 | 6.5 | -0.71 | -0.65 | -22.47% | -0.11 |
| Connected text fluency | Control | 945 | 18.18 | 18.58 | 836 | 21.03 | 22 | 2.85 | n/a | 15.68% | |
| | Full | 968 | 19.44 | 19.79 | 907 | 29.48 | 26.18 | 10.04 | 7.19 | 51.65% | 0.42 |
| | Light | 1031 | 21.04 | 21.44 | 905 | 27.11 | 26.31 | 6.07 | 3.22 | 28.85% | 0.19 |
| Reading compre-hension | Control | 948 | 25.42 | 24.35 | 836 | 17.58 | 22.96 | -7.84 | n/a | -30.84% | |
| | Full | 979 | 23.84 | 23.74 | 907 | 24.21 | 24.35 | 0.37 | 8.21 | 1.55% | 0.34 |
| | Light | 1036 | 25.77 | 24.61 | 905 | 21.08 | 23.67 | -4.69 | 3.15 | -18.20% | 0.13 |
| Listening compre-hension | Control | 951 | 32.83 | 20.55 | 846 | 66.93 | 33.8 | 34.1 | n/a | 103.87% | |
| | Full | 980 | 34.27 | 19.69 | 932 | 78.25 | 28.06 | 43.98 | 9.88 | 128.33% | 0.33 |
| | Light | 1039 | 33.63 | 21.01 | 935 | 77.48 | 27.92 | 43.85 | 9.75 | 130.39% | 0.32 |

Table 12 above shows that the program has had an unexpectedly large impact on nearly all of the subtasks, save oral reading fluency and reading comprehension. Looking at the "Percent increase over baseline" column, for letter naming fluency, the change for full and light treatment was 46.5% and 34.7%, respectively, representing an increase of 27.2 and 21.4 letters per minute. For phonemic awareness, the change was 29.8% and 16.9%, respectively. For familiar words, the difference was 68.5% and 54.2%, quite a large gap between baseline and midterm, representing 6.3 and 5.6 words per minute. For unfamiliar words, full treatment children outperformed their baseline counterparts by 78.7%, while for the light treatment children, they did 22.5% worse. Substantively, however, the change was less than 1.5 words in either direction. Critically, for oral reading fluency, full treatment schools increased their words per minute by 51.7% (10.0 words), and light treatment schools increased by 28.9% (6.1 words). Control schools increased as well (15.7%), but not by nearly as much (2.9 words). Given the connection with oral reading fluency, it is not surprising that the order of impact was similar between treatment and

control schools for reading comprehension. Children in full treatment schools increased their reading comprehension scores by 1.6%, while those in light treatment decreased by 21.1%.[16] Again, note that these scores should be read in context, since control children decreased their reading comprehension scores by 30.8%, and the reading comprehension scores were not calibrated for difficulty. For listening comprehension, full treatment and light treatment schools increased their scores by 128.3% and 130.4%, respectively.

In summary, comparing the full baseline data set against the full midterm, children's scores increased by most measures. Moreover, the magnitude of the increases makes substantive sense given the complexity of reading achievement: While teaching letters is relatively easy, teaching phonemic awareness is more difficult. Identifying familiar words is relatively simple, but teaching decoding skills evident in unfamiliar words is harder. Likewise, oral reading fluency is more difficult because it depends on the combination of all of the above skills, and reading comprehension is dependent on reading fluency. Therefore, the findings match the theory, in that during the first year of EGRA Plus, the program was able to have relatively large impacts on the portions of reading that are easiest to impact, and more modest impacts where the complexity of teaching reading is larger.

Note that while citing the percentage increase over baseline is important, it does not take into account the scores from the baseline study collected before implementation. Accounting for the baseline scores enables the researcher to estimate the "secular trend," which in the example of the Liberian program, means a great number of outside forces working on student achievement above and beyond the program. For example, if a subtask was easier on the midterm than it was on the baseline, then reporting only on the change over the baseline does not account for the secular trend. The effect size column, as calculated here, notes the gain over the baseline that is also greater than the control, and converts that difference to an effect size. As far as the estimation of a true program effect is concerned, this is a better method. For letter naming fluency, the effect sizes for both full (.73 SD) and light treatment (.52 SD) were large, by social science research standards. For phonemic awareness, the effect size for full schools (.22 SD) was moderate, while the effect in light schools (.06 SD) was small. For familiar words, the effect size was small for both full (.10 SD) and light (.06) schools. In the unfamiliar word fluency subtask, the effect was moderate for full schools (.25 SD), and actually negative for light schools (-.11 SD). In oral reading fluency, the effect was moderate for full treatment schools (.42 SD) and small for light treatment schools (.19 SD). Moderate effects were found for reading comprehension (.34 SD) and listening comprehension (.33 SD). It is worth repeating that as far as true program impact is concerned, these figures are more appropriate.

[16] Note that the scores under reading and listening comprehension were converted to percentage correct scores.

## 11.3 Program Impact Comparing Baseline Grade 2 and Midterm Grade 2

While the discussion above compares the entire baseline against the entire midterm assessment sample, ***Table 13*** below compares only grade 2 students in the baseline and midterm assessments. The columns on the right, "Percent increase over baseline" and "Effect size," are the important ones with respect to program impact. Regarding the increase over baseline, full treatment schools had an increase over baseline of over 15% for every subtask at grade 2, with many of the subtasks increasing over baseline above 100% (familiar words, unfamiliar words and listening comprehension). For light treatment schools, there were increases for every subtask except reading comprehension, which is likely at least partially due to the different stories and questions used in the baseline and midterm assessments, and the lack of calibration for the reading comprehension questions. A better estimate of the program impact, expressed in effect sizes in the column to the absolute right, shows that while the gains were not large, children in light treatment schools scored higher than those in control schools by .02 and .07 SD in oral reading fluency and reading comprehension, respectively. The effect size was moderate for full treatment schools at .38 and .36 SD. Notably, for every subtask save familiar words, where the effect size was small, the effect size for children in full treatment schools was of moderate magnitude, with large impacts found in letter naming fluency. Recall that in social science research, detecting even small effect sizes is noteworthy. It seems that the program had an impact on student achievement in early reading subtasks in every respect, except for grade 2 children in light treatment schools on familiar words and unfamiliar words.

**Table 13. Program impact at baseline and midterm for grade 2**

| Item | School type | Baseline, grade 2 | | | Midterm, grade 2 | | | Program impact | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | Mean | Standard deviation | *N* | Mean | Standard deviation | Raw gains over baseline | Raw increase over control | Percent increase over baseline | Effect size (SD) |
| Letter naming fluency | Control | 499 | 57.9 | 23.6 | 437 | 65.73 | 28.58 | 7.83 | | 13.52% | |
| | Full | 494 | 52.27 | 25.31 | 483 | 79.4 | 25.04 | 27.13 | 19.3 | 51.90% | 0.73 |
| | Light | 545 | 55.82 | 25.32 | 494 | 76.13 | 27.63 | 20.31 | 12.48 | 36.38% | 0.47 |
| Phonemic awareness | Control | 499 | 3.26 | 2.33 | 456 | 3.57 | 2.64 | 0.31 | n/a | 9.51% | |
| | Full | 499 | 3.05 | 1.98 | 501 | 4.06 | 2.56 | 1.01 | 0.7 | 33.11% | 0.27 |
| | Light | 547 | 3.22 | 2.25 | 509 | 3.77 | 2.52 | 0.55 | 0.24 | 17.08% | 0.09 |
| Familiar word fluency | Control | 495 | 6.26 | 10.37 | 432 | 10.96 | 15.01 | 4.7 | n/a | 75.08% | |
| | Full | 489 | 4.95 | 9.48 | 481 | 11.42 | 13.83 | 6.47 | 1.77 | 130.71% | 0.10 |
| | Light | 540 | 6.88 | 12.47 | 487 | 11.1 | 13.78 | 4.22 | -0.48 | 61.34% | -0.03 |

| Item | School type | Baseline, grade 2 | | | Midterm, grade 2 | | | Program impact | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | Mean | Standard deviation | *N* | Mean | Standard deviation | Raw gains over baseline | Raw increase over control | Percent increase over baseline | Effect size (SD) |
| Unfamiliar word fluency | Control | 497 | 1.35 | 4.37 | 434 | 1.33 | 4.16 | -0.02 | n/a | -1.48% | |
| | Full | 492 | 1.23 | 4.38 | 479 | 2.58 | 5.84 | 1.35 | 1.37 | 109.76% | 0.23 |
| | Light | 541 | 2.16 | 6.4 | 491 | 1.74 | 4.84 | -0.42 | -0.4 | -19.44% | -0.07 |
| Connected text fluency | Control | 495 | 14.45 | 15.63 | 432 | 17.99 | 22.8 | 3.54 | n/a | 24.50% | |
| | Full | 488 | 12.97 | 15.81 | 473 | 23.15 | 24.41 | 10.18 | 6.64 | 78.49% | 0.38 |
| | Light | 540 | 16.03 | 18.76 | 475 | 20.78 | 23.24 | 4.75 | 1.21 | 29.63% | 0.07 |
| Reading comprehension | Control | 496 | 21.41 | 23 | 432 | 15.37 | 23.5 | -6.04 | n/a | -28.21% | |
| | Full | 494 | 16.8 | 20.55 | 473 | 19.37 | 21.91 | 2.57 | 8.61 | 15.30% | 0.36 |
| | Light | 544 | 20.48 | 22.65 | 475 | 16.08 | 21.91 | -4.4 | 1.64 | -21.48% | 0.07 |
| Listening comprehension | Control | 499 | 30.58 | 20.4 | 433 | 64.06 | 33.88 | 33.48 | n/a | 109.48% | |
| | Full | 494 | 30.45 | 19.88 | 483 | 74.7 | 29.12 | 44.25 | 10.77 | 145.32% | 0.35 |
| | Light | 547 | 30.02 | 21.29 | 492 | 73.39 | 29.05 | 43.37 | 9.89 | 144.47% | 0.33 |

## 11.4 Program Impact Comparing Grade 3 Baseline and Grade 3 Midterm

Similar to Table 13, which examines grade 2 scores, ***Table 14*** below explores the impact of the full and light treatment programs on their percentage increase in grade 3 over baseline, and the effect size. The pattern follows what was found in grade 2, for the first several subtasks, except for a very interesting dip in the scores compared to baseline for reading comprehension. For all three groups (control, full, and light treatment), grade 3 children scored much lower on the midterm than they did on the baseline assessment for this subtask. The difference was largest for control children, who scored 33.3% lower on reading comprehension.[17] The next drop was for light treatment children, who scored 16.3% lower. Finally, for full treatment children, the difference was 5.1%.

When the changes over baseline are converted to effect sizes of impacts over the control groups, we find that the children in full treatment schools had effect sizes of at least .12 SD (familiar words) and up to .75 SD (letter naming). For light treatment children, most of the effects were positive, although there was a slight negative effect for phonemic awareness (-.02 SD) and a moderately negative effect (-.16 SD) for unfamiliar words.

[17] Note again that scores for reading and listening comprehension were transformed from raw scores to percentage correct.

This latter effect is possibly because while it is easier to increase students' knowledge of names and frequently used letters based on simple increases of exposure to this type of activity, unfamiliar words requires systematic training on phonics and decoding, which light treatment schools were not provided. Further disaggregation of these scores by treatment and gender can be found in ***Appendix C.*** Note that one area of interest will be the unfamiliar word, reading, and comprehension scores for boys at midterm, which seem to drive the unexpectedly low findings for these subtasks.

**Table 14. Program impact at baseline and midterm for grade 3**

| Item | School type | Baseline, grade 3 | | | Midterm, grade 3 | | | Program impact | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | Mean | Standard deviation | *N* | Mean | Standard deviation | Raw gains over baseline | Raw increase over control | Percent increase over baseline | Effect size (SD) |
| Letter naming fluency | Control | 451 | 68.99 | 26.25 | 404 | 76.7 | 23.3 | 7.71 | | 11.18% | |
| | Full | 478 | 64.83 | 22.73 | 435 | 92.32 | 21.36 | 27.49 | 19.78 | 42.40% | 0.75 |
| | Light | 482 | 68.74 | 23.77 | 438 | 91.09 | 22.32 | 22.35 | 14.64 | 32.51% | 0.55 |
| Phonemic awareness | Control | 450 | 3.6 | 2.35 | 409 | 4.3 | 2.69 | 0.7 | n/a | 19.44% | |
| | Full | 478 | 3.79 | 2.29 | 459 | 4.89 | 2.55 | 1.1 | 0.4 | 29.02% | 0.15 |
| | Light | 482 | 4.09 | 2.41 | 442 | 4.74 | 2.52 | 0.65 | -0.05 | 15.89% | -0.02 |
| Familiar word fluency | Control | 444 | 10.6 | 14.36 | 402 | 15.19 | 14.02 | 4.59 | n/a | 43.30% | |
| | Full | 475 | 13.34 | 15.86 | 434 | 19.63 | 17.45 | 6.29 | 1.7 | 47.15% | 0.09 |
| | Light | 481 | 14.1 | 16.85 | 437 | 21.08 | 19.6 | 6.98 | 2.39 | 49.50% | 0.13 |
| Unfamiliar word fluency | Control | 449 | 2.06 | 5.07 | 404 | 1.92 | 5.14 | -0.14 | n/a | -6.80% | |
| | Full | 472 | 2.43 | 5.98 | 434 | 4.01 | 6.51 | 1.58 | 1.72 | 65.02% | 0.29 |
| | Light | 477 | 4.35 | 8.4 | 433 | 3.26 | 7.92 | -1.09 | -0.95 | -25.06% | -0.16 |
| Connected text fluency | Control | 449 | 22.33 | 20.61 | 396 | 24.21 | 20.65 | 1.88 | n/a | 8.42% | |
| | Full | 473 | 26 | 21.27 | 427 | 36.13 | 26.21 | 10.13 | 8.25 | 38.96% | 0.48 |
| | Light | 482 | 26.7 | 22.89 | 425 | 34.25 | 27.76 | 7.55 | 5.67 | 28.28% | 0.33 |
| Reading comprehension | Control | 451 | 29.89 | 25.03 | 396 | 19.95 | 22.07 | -9.94 | n/a | -33.26% | |
| | Full | 478 | 30.96 | 24.58 | 427 | 29.37 | 25.83 | -1.59 | 8.35 | -5.14% | 0.35 |
| | Light | 483 | 31.84 | 25.45 | 425 | 26.64 | 24.34 | -5.2 | 4.74 | -16.33% | 0.20 |

| Item | School type | Baseline, grade 3 | | | Midterm, grade 3 | | | Program impact | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | Mean | Standard deviation | *N* | Mean | Standard deviation | Raw gains over baseline | Raw increase over control | Percent increase over baseline | Effect size (SD) |
| Listening comprehension | Control | 451 | 35.39 | 20.4 | 405 | 70.32 | 33.28 | 34.93 | n/a | 98.70% | |
| | Full | 479 | 38.29 | 18.69 | 441 | 81.9 | 26.51 | 43.61 | 8.68 | 113.89% | 0.29 |
| | Light | 483 | 37.85 | 20.05 | 438 | 82.03 | 25.9 | 44.18 | 9.25 | 116.72% | 0.30 |

## 11.5 Program Impact Comparing Baseline and Midterm, Disaggregated by Gender

A final table disaggregated more fully by gender makes an important point. ***Table 15*** compares achievement on the various subtasks disaggregated by both school type and gender. It shows a general pattern of increasing scores, particularly for treatment schools. However, for unfamiliar words and reading comprehension, boys' scores actually declined between baseline and midterm. In each case, the scores for boys in full treatment schools decreased by less than those in control schools, which explains why the effect sizes remained relatively large for boys in full treatment schools. However, there remained some underachievement by boys in the midterm assessment in some of the portions of early grade reading that are most correlated with future achievement. While the program has been successful in teaching letters and familiar words, and remains successful in outperforming control schools, more work is necessary for boys in the more complex areas of reading.

## Table 15. Program impact at baseline and midterm for grade 2 and grade 3, by gender

| Item | School type | Gender | Baseline, grades 2 and 3 | | | Midterm, grades 2 and 3 | | | Program impact | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *N* | Mean | Standard deviation | *N* | Mean | Standard deviation | Raw gains over baseline | Raw increase over control | Percent increase over baseline | Effect size (SD) |
| Letter naming fluency | Control | Male | 541 | 65.03 | 25.26 | 385 | 68.93 | 26.95 | 3.9 | | 6.00% | |
| | | Female | 405 | 60.62 | 25.64 | 445 | 72.73 | 26.59 | 12.11 | | 19.98% | |
| | Full | Male | 511 | 61.17 | 24.05 | 453 | 83.57 | 23.81 | 22.4 | 18.5 | **36.62%** | **0.73** |
| | | Female | 462 | 55.28 | 25.32 | 456 | 87.7 | 24.39 | 32.42 | 20.31 | **58.65%** | **0.80** |
| | Light | Male | 570 | 64.2 | 25.58 | 426 | 80.61 | 27.65 | 16.41 | 12.51 | **25.56%** | **0.49** |
| | | Female | 461 | 58.75 | 24.9 | 504 | 85.6 | 25 | 26.85 | 14.74 | **45.70%** | **0.58** |
| Phonemic awareness | Control | Male | 540 | 3.59 | 2.3 | 397 | 3.68 | 2.62 | 0.09 | | 2.51% | |
| | | Female | 404 | 3.23 | 2.37 | 464 | 3.99 | 2.76 | 0.76 | | 23.53% | |
| | Full | Male | 511 | 3.44 | 2.17 | 474 | 4.4 | 2.55 | 0.96 | 0.87 | **27.91%** | **0.38** |
| | | Female | 461 | 3.41 | 2.17 | 477 | 4.5 | 2.63 | 1.09 | 0.33 | **31.96%** | **0.14** |
| | Light | Male | 569 | 3.75 | 2.43 | 439 | 4.08 | 2.56 | 0.33 | 0.24 | **8.80%** | **0.10** |
| | | Female | 463 | 3.54 | 2.27 | 511 | 4.36 | 2.55 | 0.82 | 0.06 | **23.16%** | **0.03** |
| Familiar word fluency | Control | Male | 538 | 10.02 | 13.77 | 386 | 11.7 | 13.49 | 1.68 | | 16.77% | |
| | | Female | 400 | 6.03 | 10.43 | 444 | 14.05 | 15.55 | 8.02 | | 133.00% | |
| | Full | Male | 505 | 10.82 | 14.69 | 452 | 13.16 | 15.36 | 2.34 | 0.66 | **21.63%** | **0.05** |
| | | Female | 460 | 7.39 | 12.5 | 454 | 17.68 | 16.75 | 10.29 | 2.27 | **139.24%** | **0.16** |

| Item | School type | Gender | Baseline, grades 2 and 3 | | | Midterm, grades 2 and 3 | | | Program impact | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *N* | Mean | Standard deviation | *N* | Mean | Standard deviation | Raw gains over baseline | Raw increase over control | Percent increase over baseline | Effect size (SD) |
| | Light | Male | 566 | 11.53 | 15.61 | 422 | 13.4 | 16.51 | 1.87 | 0.19 | **16.22%** | **0.01** |
| | | Female | 459 | 8.6 | | 500 | 17.82 | 18.06 | 9.22 | 1.2 | **107.21%** | **0.09** |
| Unfamiliar word fluency | Control | Male | 540 | 2.18 | 5.51 | 389 | 1.02 | 4.08 | -1.16 | | -53.21% | |
| | | Female | 405 | 1.03 | 3.31 | 444 | 2.17 | 5.09 | 1.14 | | 110.68% | |
| | Full | Male | 508 | 2.28 | 5.89 | 452 | 2.21 | 4.41 | -0.07 | 1.09 | **-3.07%** | **0.18** |
| | | Female | 457 | 1.35 | 4.55 | 452 | 4.39 | 7.51 | 3.04 | 1.9 | **225.19%** | **0.32** |
| | Light | Male | 566 | 3.81 | 8.07 | 418 | 1.59 | 6.08 | -2.22 | -1.06 | **-58.27%** | **-0.18** |
| | | Female | 456 | 2.34 | 6.57 | 504 | 3.16 | 6.74 | 0.82 | -0.32 | **35.04%** | **-0.05** |
| Connected text fluency | Control | Male | 540 | 20.38 | 18.96 | 385 | 20.26 | 21.78 | -0.12 | | -0.59% | |
| | | Female | 403 | 15.27 | 17.69 | 438 | 21.69 | 22.41 | 6.42 | | 42.04% | |
| | Full | Male | 505 | 21.98 | 20.35 | 443 | 26.3 | 25.51 | 4.32 | 4.44 | **19.65%** | **0.22** |
| | | Female | 457 | 16.53 | 18.79 | 446 | 32.58 | 26.46 | 16.05 | 9.63 | **97.10%** | **0.48** |
| | Light | Male | 565 | 22.86 | 21.38 | 414 | 23.97 | 25.19 | 1.11 | 1.23 | **4.86%** | **0.06** |
| | | Female | 461 | 18.6 | 21.27 | 484 | 29.75 | 27 | 11.15 | 4.73 | **59.95%** | **0.24** |
| Reading comprehension | Control | Male | 541 | 27.69 | 24.92 | 385 | 16.88 | 21.52 | -10.81 | | -39.04% | |
| | | Female | 405 | 22.42 | 23.27 | 438 | 14.88 | 15.37 | -7.54 | | -33.63% | |
| | Full | Male | 511 | 26.11 | 23.7 | 443 | 23.02 | 24.53 | -3.09 | 7.72 | **-11.83%** | **0.32** |

| Item | School type | Gender | Baseline, grades 2 and 3 | | | Midterm, grades 2 and 3 | | | Program impact | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | *N* | Mean | Standard deviation | *N* | Mean | Standard deviation | Raw gains over baseline | Raw increase over control | Percent increase over baseline | Effect size (SD) |
| | | Female | 462 | 21.17 | 23.48 | 446 | 25.61 | 24.02 | 4.44 | 11.98 | **20.97%** | **0.49** |
| | Light | Male | 569 | 26.89 | 23.95 | 414 | 18.74 | 22.1 | -8.15 | 2.66 | **-30.31%** | **0.11** |
| | | Female | 462 | 24.2 | 25.28 | 484 | 23.1 | 24.61 | -1.1 | 6.44 | **-4.55%** | **0.27** |
| Listening comprehension | Control | Male | 543 | 33.96 | 20.23 | 385 | 16.88 | 21.52 | -17.08 | | -50.29% | |
| | | Female | 406 | 31.23 | 20.89 | 445 | 66.49 | 33.83 | 35.26 | | 112.90% | |
| | Full | Male | 512 | 35.74 | 19.7 | 455 | 77.49 | 27.27 | 41.75 | 58.83 | **116.82%** | **2.87** |
| | | Female | 462 | 32.73 | 19.65 | 460 | 79.5 | 28.07 | 46.77 | 11.51 | **142.90%** | **0.56** |
| | Light | Male | 571 | 34.89 | 20.67 | 421 | 76.53 | 29.04 | 41.64 | 58.72 | **119.35%** | **2.87** |
| | | Female | 463 | 31.92 | 21.37 | 507 | 78.17 | 26.98 | 46.25 | 10.99 | **144.89%** | **0.54** |

# 12. Liberia Comparisons and Benchmarks

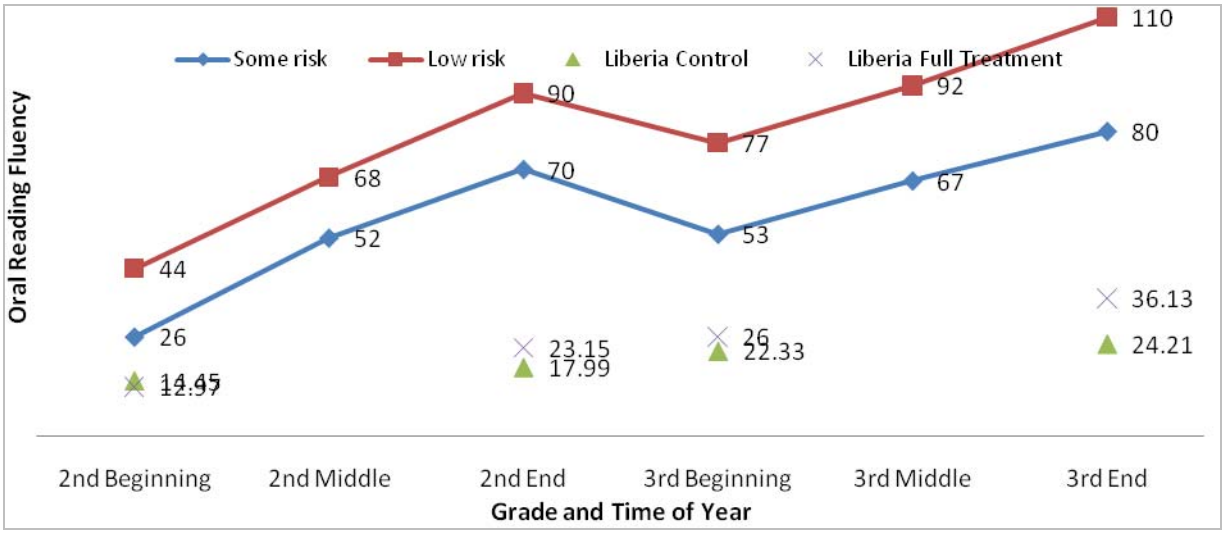
## 12.1 Comparisons with International Benchmarks

The findings in this report are cautiously optimistic about the impact that EGRA Plus can have on basic reading skills. The impact appears to be moderate in size, although it seems consistently smaller in the area of reading comprehension, which is unsurprising, given the complexity of this skill and the short time that the program had to impact student achievement.

This section provides some comparisons between oral reading fluency in Liberia (both control and full treatment schools) and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS) benchmarks for oral reading fluency. ***Figure 16*** below shows several things. The blue line shows the "some risk" benchmark for oral reading fluency, while the red line shows the "low risk" benchmark. Given that children in both treatment and control schools scored below the "some risk" benchmark, this provides some evidence that while the program has had an impact, there remains quite a long way to go.

One important thing to note from Figure 16 is that at every level, children in full treatment schools outperformed control schools, and the gap between those groups increases from the beginning of grade 2 to the end of grade 3. Visual inspection of this figure shows that the slope of the international benchmark curves are more pronounced than for the Liberia curves. In other words, children in the benchmark schools increase their oral reading fluency within grades more than Liberian children do. This is important to note for the EGRA Plus program, since it can be argued that more needs to be done within both grade 2 and grade 3 to increase the slope of learning between the beginning and the end of the year. Some of this modest slope is, of course, due to the impediments to implementation and successful pedagogy in the 2008–2009 academic year. Finally, it appears that the Liberian children in grade 3 (both control and full treatment) are not gaining proportionally as much in grade 3 as they do internationally. In other words, it appears that some focus on grade 3 achievement is necessary, and that while the program appears to have made a dent in the problem of early reading achievement, much work remains to be done.

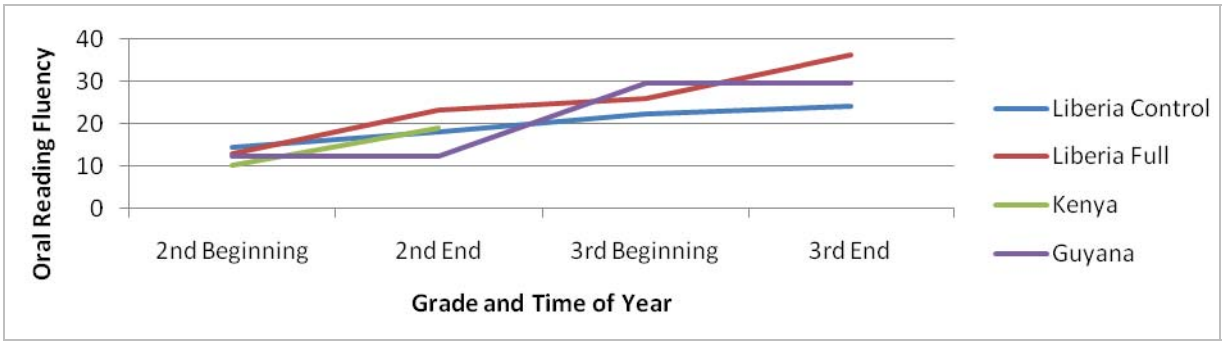
**Figure 16. Oral reading fluency scores compared to international benchmarks**




## 12.2 Comparisons with Kenya and Guyana

While the discussion above is interesting, and there is some value in comparing Liberian children to what is found in the U.S. DIBELS benchmarks, it might be even more valuable to compare against the oral reading fluency scores in other countries, namely Kenya and Guyana. Note that even this is fraught with problems given the language differences and the local adaptation of EGRA in each country. Even in countries where English is assessed, the assessments can be quite different since each story is locally created. That said, it is worth taking a look at the comparisons between students. A cursory glance at ***Figure 17*** shows that children in Liberian schools scored as low as students in both Kenya and Guyana, and lower than Guyanese students in grade 3. The slope of their learning between the beginning and end of grade 2 and 3 was less than what occurs in both Kenya and Guyana, and the increase between grade 2 and 3 was smaller than what exists elsewhere. Therefore, compared to Kenya and Guyana, children in control schools are learning to read much too slowly. The story is slightly different for children in full treatment schools, where the increases shown in Figure 17 were more dramatic.

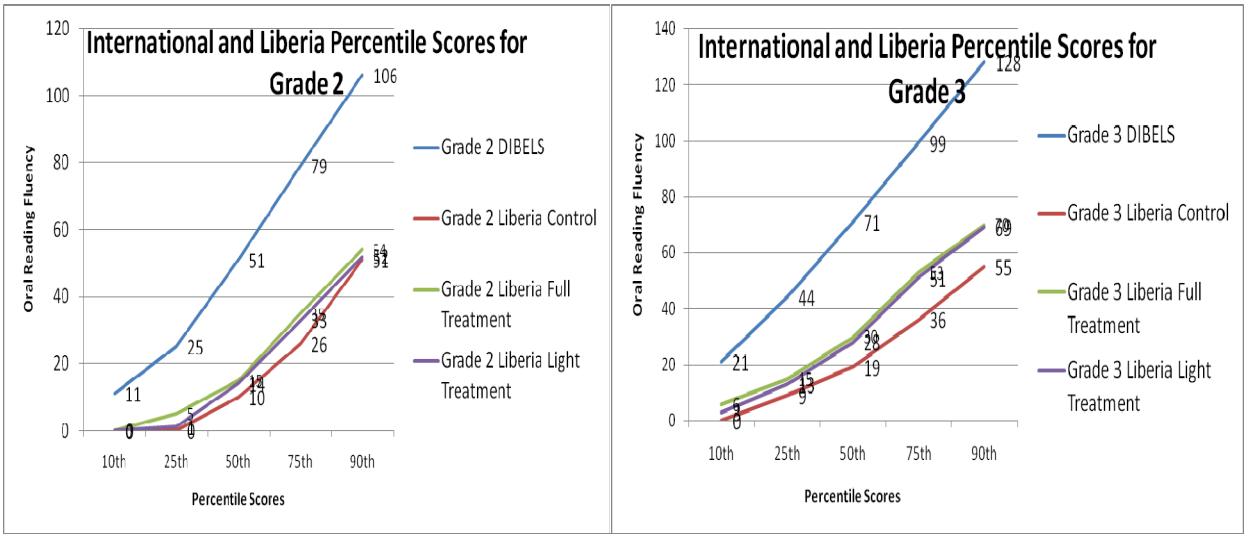
**Figure 17. Oral reading fluency scores in Liberia compared to other developing countries**




## 12.3 Percentile Score Comparisons with DIBELS

***Figure 18*** below shows Liberia's grade 2 and 3 student achievement in oral reading fluency across treatment groups against the grade 2 and 3 international benchmarks. Note that Figure 18 is organized differently from Figure 17. In Figure 18, the percentile scores relate to the distribution of scores within each separate data set. In other words, for the DIBELS scores, all of the children assessed are ranked by percentile, and this figure shows how they fall out, from the 10th, 25th, 50th, 75th, and 90th percentiles at grade 2. What this figure shows is that while the gaps between the 10th-percentile child in all of Liberia's schools and that of the DIBELS children is small (11 words per minute), the gap increases rapidly across the distribution. When we compare control and treatment schools within Liberia, note that both the light and full treatment schools were most able to limit the gap at the 25th-, 50th-, and 75th-percentile scores, as compared to the control children. In other words, while the gap widens across the distribution, treatment schools were best able to limit that gap in the middle portion of the distribution.

**Figure 18. Liberia percentile scores compared to international benchmarks, by grade**

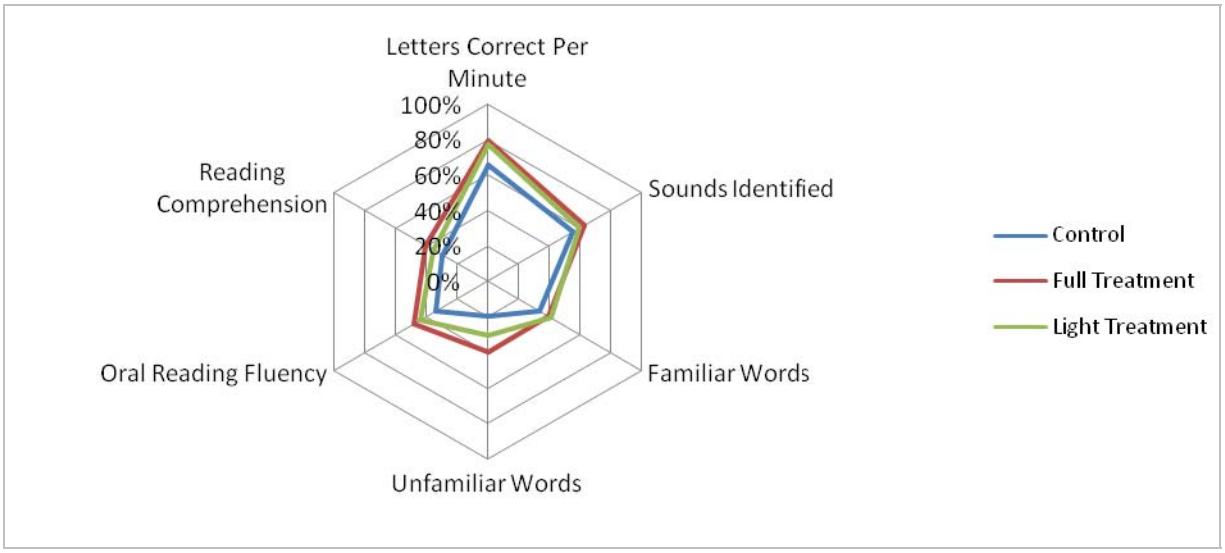


The grade 3 graph in Figure 18 above shows a similar story: a rapidly widening gap between the Liberian distribution and the DIBELS one. However, the gaps for children in treatment schools (both light and full) are less than for control schools, particularly at the 50th and 75th percentiles. This provides hopeful evidence that children in treatment schools will be able to lessen the gaps between themselves and readers elsewhere.

## 12.4 Liberian Benchmark Example

In the hopes of contributing to a sort of a Liberian benchmark, ***Figure 19*** below was created. This figure extracts the 90th percentile of Liberia's distribution of children on several subtasks: letters per minute, phonemic awareness, familiar words, unfamiliar words, oral reading fluency and reading comprehension. This 90th percentile score is the "benchmark," then, and serves as the outward point on the radial plot. This was done to create a sort of Liberia-specific ideal for student achievement. Substantively, this means that 108 letters correct per minute, 7 sounds correctly identified, 38.7 familiar words read, 8.3 unfamiliar words read, 61.5 words read on connected text, and 60% reading comprehension were kept as the targets. The blue (control), red (full treatment), and green (light treatment) lines show how closely each group of children was to this reading target.

**Figure 19. 90th percentile of Liberian benchmarks, compared to treatment groups**




It is easy to see that the "red" children (full treatment) were closest to the targets in general, and the "blue" children (control) were farthest away. Some other points are worth making here: While there were only nominal differences between the light and full treatment schools at the letters correct, sounds identified, and familiar word subtasks, the gap widens at unfamiliar words, oral reading fluency, and reading comprehension. This is likely because—as pointed out earlier—while it is relatively easy to teach letter sounds

and a small number of familiar words, decoding (unfamiliar words), reading (oral reading fluency), and comprehension require more effort and technique. It is also worth noting how far away from the "target" all three groups are, particularly at the more complex subtasks of unfamiliar words, oral reading fluency, and reading comprehension. There remain large differences among the treatment groups for these subtasks, but substantively, all three groups have quite far to go.

# 13. EGRA Impact Analysis

Impact studies take a variety of forms and use different strategies to assess the impact of a program on student outcomes. In the sections that preceded this one, we used simple tabulation analyses to determine whether the program had an impact on student achievement. This is acceptable, but regression models have a variety of benefits over the more simple comparison techniques. For example, the *t*-tests inherent in the models allow for an estimate of whether or not an individual predictor (gender or grade, for example) has a statistically significant impact on a particular outcome. In addition, the research design of this particular study lends itself to an analytic method called differences-in-differences analysis. This analysis falls in the category of causal analytic methods, which attempt to use statistical techniques to estimate the actual causal impact of a program of interest.

The differences-in-differences technique uses the pre/post and treatment/control nature 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. What is done to perform this analysis is to create a combined data set with both the baseline, and in this case, the midterm assessment. Children are either identified as baseline or midterm and as treatment or control. In this case the analysis is slightly more complicated because there are two treatment groups. However, using a system of dummy variables in the regression analysis, one can estimate the effects 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. Finally, post-hoc general linear hypothesis (GLH) tests can compare whether the impacts of the two treatment groups are equivalent—or, to put it another way, whether the full treatment program works better than the light treatment group. The models below have several parameters, which are defined here.

- Post – This represents whether a child is in the baseline or midterm (post) data set.
- Light Treatment – This represents whether the child is in the light treatment group.
- Light Treatment*Post – This identifies the children who were in both the post and light treatment group.

- Full Treatment – This represents the children in the full treatment group.
- Full Treat*Post – this represents the children who were in the post and full treatment group.
- Gender (girl) – This variable shows the effect of being a girl, compared against boys.
- Grade (3) – This variable shows the effect of grade 3, compared against grade 2.
- Control Group – This is the constant variable, which in this design is the average score of a boy in grade 2 in the control group.

## 13.1 General Findings

This set of models shows that the full treatment program has had a statistically significant impact on student achievement on all of the subtasks, while light treatment has had an impact on letters per minute, oral reading fluency, and listening comprehension. The models also show no statistically significant gender differences save for oral reading fluency, and, of course, grade 3 children outperforming grade 2.[18]

## 13.2 Subtask-Specific Findings

### *13.2.1 Letters Per Minute*

This model shows that both the full and light treatment programs have had an effect on achievement in letter naming fluency. Girls did no worse than boys (*p*-value .42), and grade 3 children did better than grade 2 (12.7 letters). Children in the control group scored 56.8 letters, with children in the midterm (rather than baseline assessment) scoring 7.9 letters higher, showing that children were learning letters in both grade 2 and 3. The main effect of being in a treatment group (regardless of baseline or midterm) was 1.2 letters less (light) and 5.0 letters less (full). Critically, the causal effect of being a child in a light treatment group was an additional 13.6 letters per minute. The effect of being a child in a full treatment group was 19.6 letters per minute. In other words, both treatment groups increased student achievement in letters. The GLH test is statistically significant, meaning that the full treatment group experienced a bigger impact than the light treatment group.The model does a reasonably good job of predicting achievement on letters per minute, since the $R^2$ is .20.

[18] Models were fit that investigated whether the treatment programs had a differential impact by gender and/or grade. In both cases, the treatment programs did not have a differential impact on achievement depending on grade or gender. That said, given the smaller sample sizes of the disaggregated data sets, in some cases the program effect was no longer statistically significant at the .05 level.

**Table 16. Differences-in-differences regression analysis for letter naming fluency**

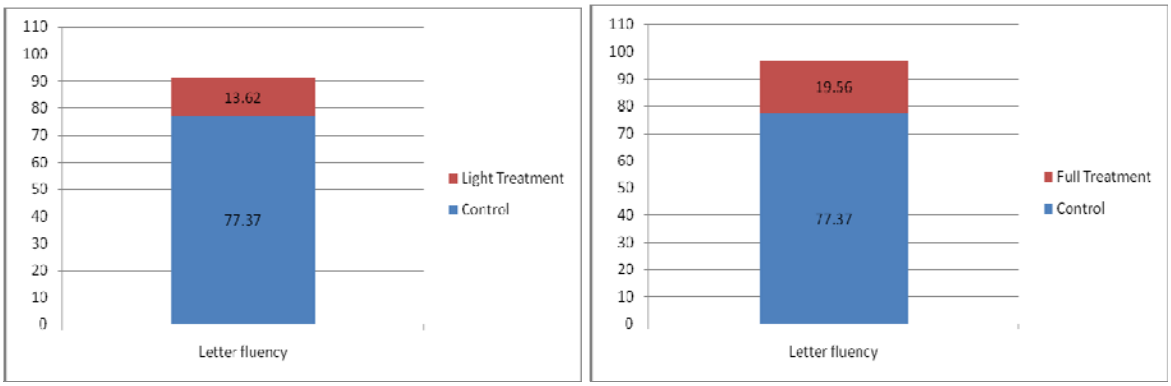
| Outcome | Predictor | Coeffi-cient | Std. error | *T* | Sig. | Confidence interval | | No. of observa-tions | *F* | Sig. | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Lower | Upper | | | | |
| Letters per minute | Post | 7.86 | 1.18 | 6.67 | <.001 | 5.55 | 10.17 | | | | |
| | Light Treatment | -1.24 | 1.12 | -1.11 | .27 | -3.43 | .95 | | | | |
| | **Light Treat*Post** | 13.62 | 1.63 | 8.38 | *<.001* | 10.44 | 16.81 | | | | |
| | Full treatment | -5.00 | 1.131 | -4.42 | <.001 | -7.22 | -2.78 | | | | |
| | **Full Treat*Post** | 19.56 | 1.64 | 11.91 | *<.001* | 16.35 | 22.78 | | | | |
| | Gender (Girl) | .54 | .66 | .81 | .42 | -.77 | 1.84 | | | | |
| | Grade (3) | 12.68 | .66 | 19.14 | <.001 | 11.38 | 13.98 | | | | |
| | **Control Group** | 56.83 | .94 | 60.29 | <.001 | 54.98 | 58.68 | 5590 | 198.36 | <.001 | .20 |

GLH Test (Light Treat*Post - Full Treat*Post = 0), *F* 13.72, *p* <.001.

Therefore, full treatment's impact is statistically significantly greater.

***Figure 20*** below shows the impact of the treatment groups graphically. The graph on the left shows, in blue, the scores of the control children. The additive, causal impact of the light treatment schools is found in red. On the right, using the same color scheme, it is evident that the full treatment group had slightly more impact on letter fluency, by about 6 letters per minute.

**Figure 20. Bar charts comparing impact of light treatment (left) and full treatment (right) programs on letter naming fluency**




### *13.2.3 Phonemic Awareness*

In ***Table 17*** below, the impact of the full and light treatment on student achievement in phonemic awareness is identified. In this subtask, the model shows that the light treatment had no impact on student achievement, since the *p*-value was .28. For the full treatment group, on the other hand, the program increased student achievement by .62

sounds. The pattern is the same as in the letter naming fluency subtask: Girls and boys scored the same (*p*-value .99) and grade 3 more than grade 2 (.76 sounds).

**Table 17. Differences-in-differences regression analysis for phonemic awareness**

| Outcome | Predictor | Coeffi-cient | Std. error | *T* | Sig. | Confidence interval | | No. of observa-tions | *F* | Sig. | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Lower | Upper | | | | |
| Phonemic awareness | Post | .43 | .11 | 3.72 | <.001 | .20 | .65 | | | | |
| | Light Treatment | .20 | .11 | 1.82 | .07 | -.02 | .41 | | | | |
| | **Light Treat*Post** | .18 | .16 | 1.11 | *.27* | -.13 | .49 | | | | |
| | Full treatment | -.02 | .11 | -.18 | .86 | -.24 | .20 | | | | |
| | **Full Treat*Post** | .62 | .16 | 3.92 | *<.001* | .31 | .94 | | | | |
| | Gender (Girl) | -.00 | .06 | -.01 | .99 | -.13 | .13 | | | | |
| | Grade (3) | .76 | .06 | 11.84 | <.001 | .64 | .89 | | | | |
| | **Control Group** | 3.07 | .09 | 33.28 | <.001 | 2.89 | 3.25 | 5669 | 41.34 | <.001 | .05 |

Since light treatment is not significant, GLH test is not relevant: Full treatment effect is larger than light treatment.

### *13.2.4 Familiar Word Fluency*

For familiar words, the analysis shows that light treatment had no statistically significant impact on achievement, but full treatment schools increased achievement by 1.8 familiar words at the .10 level. Again, girls scored the same as boys and grade 3 children better than grade 2 (7.1 words). The $R^2$ is .09, which is larger than for phonemic awareness.

**Table 18. Differences-in-differences regression analysis for familiar word fluency**

| Outcome | Predictor | Coeffi-cient | Std. error | *T* | Sig. | Confidence interval | | No. of observa-tions | *F* | Sig. | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Lower | Upper | | | | |
| Familiar word fluency | Post | 4.54 | .69 | 6.49 | <.001 | 3.17 | 5.91 | | | | |
| | Light Treatment | 1.92 | .66 | 2.91 | <.01 | .63 | 3.22 | | | | |
| | **Light Treat*Post** | 1.00 | .96 | 1.04 | *.30* | -.89 | 2.90 | | | | |
| | Full treatment | .62 | .67 | .94 | .35 | -.69 | 1.95 | | | | |
| | **Full Treat*Post** | 1.79 | .97 | 1.84 | *.07* | -.12 | 3.70 | | | | |
| | Gender (Girl) | -.10 | .39 | -.27 | .79 | -.88 | .67 | | | | |
| | Grade (3) | 7.13 | .39 | 18.15 | <.001 | 6.36 | 7.90 | | | | |
| | **Control Group** | 5.03 | .56 | 8.99 | <.001 | 3.93 | 6.12 | 5551 | 78.85 | <.001 | .09 |

GLH test is irrelevant since light treatment has no impact on familiar words.

### *13.2.5 Unfamiliar Word Fluency*

***Table 19*** shows that the treatments had a statistically significant impact on student achievement, although the light treatment's impact is only significant at the .10 level (rather than .05) level. This is notable since the model shows that the light treatment actually decreased achievement on unfamiliar words. Full treatment increased achievement by 1.6 words per minute.

**Table 19. Differences-in-differences regression analysis for unfamiliar word fluency**

| Outcome | Predictor | Coeffi-cient | Std. error | *T* | Sig. | Confidence interval | | No. of observa-tions | *F* | Sig. | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Lower | Upper | | | | |
| Unfamiliar word fluency | Post | -.09 | .28 | -.31 | .76 | .64 | .47 | | | | |
| | Light Treatment | 1.50 | .27 | 5.60 | <.001 | .97 | 2.02 | | | | |
| | **Light Treat*Post** | -.66 | .39 | -1.70 | *.09* | -1.43 | .10 | | | | |
| | Full treatment | .11 | .27 | .40 | .69 | -.42 | .64 | | | | |
| | **Full Treat*Post** | 1.56 | .39 | 3.95 | *<.001* | .78 | 2.33 | | | | |
| | Gender (Girl) | -.18 | .16 | -1.13 | .26 | -.49 | .13 | | | | |
| | Grade (3) | 1.30 | .16 | 8.20 | <.001 | .99 | 1.61 | | | | |
| | **Control Group** | 1.18 | .23 | 5.21 | <.01 | .73 | 1.62 | 5556 | 20.37 | <.001 | .03 |

GLH Test (Light Treat*Post - Full Treat*Post = 0), *F* 33.25, *p* <.001.

Therefore, full treatment's impact is statistically significantly greater.

### *13.2.6 Oral Reading Fluency*

The analysis for oral reading fluency had some disparate findings depending on whether the baseline scores were included ***(Table 20).*** This model had no problem taking into account the differences in the difficulty of the baseline and midterm assessments. It shows that the light treatment did increase the number of words read per minute by 3.4 (*p*-value .25). It also shows that the full treatment increased student achievement by 7.4 words per minute, which is about the same effect as that of an additional three quarters of a year in school (grade 3=10.8 words). These are impressive results for both the full and light treatment schools. Note that the post-hoc GLH test shows that the impact of full treatment was bigger than that of light treatment on oral reading fluency.

**Table 20. Differences-in-differences regression analysis for oral reading fluency**

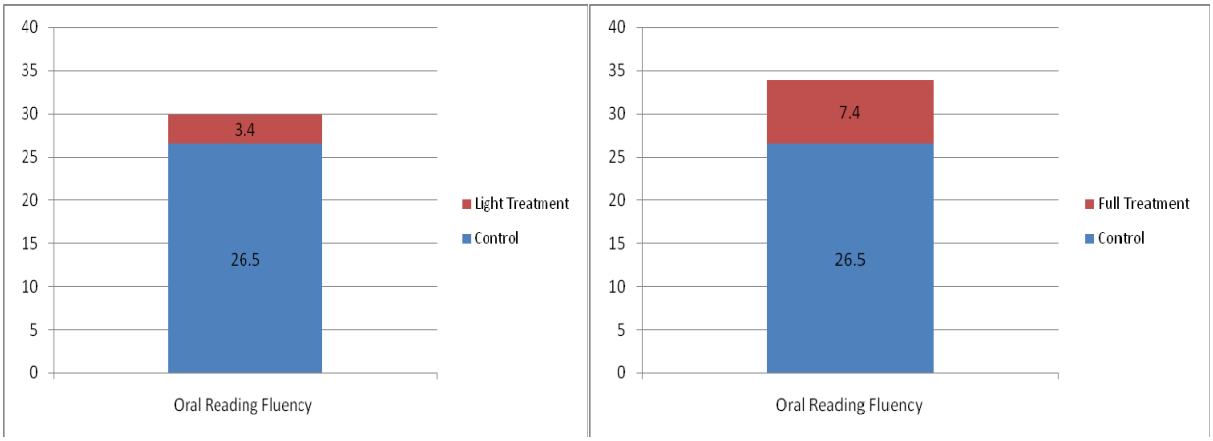
| Outcome | Predictor | Coeffi-cient | Std. error | *T* | Sig. | Confidence interval | | No. of observa-tions | *F* | Sig. | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Lower | Upper | | | | |
| Oral reading fluency | Post | 2.76 | 1.05 | 2.64 | <.01 | .71 | 4.82 | | | | |
| | Light Treatment | 2.80 | .99 | 2.84 | <.01 | .87 | 4.74 | | | | |
| | **Light Treat*Post** | 3.43 | 1.45 | 2.37 | *.02* | .59 | 6.26 | | | | |
| | Full treatment | .96 | 1.00 | .96 | .34 | -1.01 | 2.93 | | | | |
| | **Full Treat*Post** | 7.39 | 1.46 | 5.06 | *<.001* | 4.53 | 10.25 | | | | |
| | Gender (Girl) | .37 | .59 | .63 | .53 | -.79 | 1.53 | | | | |
| | Grade (3) | 10.80 | .59 | 18.32 | <.001 | 9.64 | 11.95 | | | | |
| | **Control Group** | 12.88 | .84 | 15.41 | <.001 | 11.24 | 14.52 | 5507 | 75.20 | <.001 | .09 |

GLH Test (Light Treat*Post - Full Treat*Post = 0), *F* 7.68, *p* <.01.

Therefore, full treatment's impact is statistically significantly greater.

***Figure 21*** below shows, graphically, the impact of full and light treatment on oral reading fluency. This shows that, compared to the control children, pupils in full treatment schools scored 7.4 words higher; pupils in light treatment, 3.4 words higher; and this impact is due to the program.

**Figure 21. Bar charts comparing impact of light treatment (left) and full treatment (right) on oral reading fluency**




### *13.2.7 Reading Comprehension*

For the reading comprehension subtasks, the differences-in-differences model shows that the light treatment increased student achievement by 3.2 percentage points. The full treatment program increased reading comprehension by 8.5 percentage points, more than twice the gender effect, and nearly the same amount as grade (10.0 points).

## Table 21. Differences-in-differences regression analysis for reading comprehension

| Outcome | Predictor | Coeffi-cient | Std. error | *T* | Sig. | Confidence interval | | No. of observa-tions | *F* | Sig. | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Lower | Upper | | | | |
| Reading compre-hension | Post | -7.79 | 1.12 | -6.95 | <.001 | -9.99 | -5.59 | | | | |
| | Light Treatment | .35 | 1.06 | .33 | .74 | -1.73 | 2.42 | | | | |
| | **Light Treat*Post** | 3.21 | 1.55 | 2.07 | .04 | .17 | 6.25 | | | | |
| | Full treatment | -1.89 | 1.07 | -1.76 | .08 | -3.99 | .21 | | | | |
| | **Full Treat*Post** | 8.53 | 1.56 | 5.46 | *<.001* | 5.47 | 11.60 | | | | |
| | Gender (Girl) | .89 | .63 | 1.40 | .16 | -.36 | 2.12 | | | | |
| | Grade (3) | 10.04 | .63 | 15.91 | <.001 | 8.80 | 11.28 | | | | |
| | **Control Group** | .20.18 | .89 | 22.56 | <.001 | 18.43 | 21.94 | 5526 | 47.47 | <.001 | .06 |

GLH Test (Light Treat*Post - Full Treat*Post = 0), *F* 12.09, *p* <.001.

Therefore, full treatment's impact is statistically significantly different from light treatment.

### *13.2.8 Listening Comprehension*

Finally, for listening comprehension, the model shows that both light treatment and full treatment had an impact on student achievement: 9.6 and 9.9 percentage points, respectively ***(Table 22).*** The GLH test reveals no differences between treatment groups on this measure.

## Table 22. Differences-in-differences regression analysis for listening comprehension

| Outcome | Predictor | Coeffi-cient | Std. error | *T* | Sig. | Confidence interval | | No. of observa-tions | *F* | Sig. | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Lower | Upper | | | | |
| Listening compre-hension | Post | 34.27 | 1.20 | 28.66 | <.001 | 31.93 | 36.62 | | | | |
| | Light Treatment | .85 | 1.13 | .75 | .45 | -1.37 | 3.07 | | | | |
| | **Light Treat*Post** | 9.61 | 1.65 | 5.83 | <.001 | 6.38 | 12.85 | | | | |
| | Full treatment | 1.45 | 1.15 | 1.26 | .21 | -.80 | 3.70 | | | | |
| | **Full Treat*Post** | 9.92 | 1.66 | 5.96 | <.001 | *6.65* | 13.18 | | | | |
| | Gender (Girl) | .95 | .67 | 1.41 | .16 | -.37 | 2.27 | | | | |
| | Grade (3) | 7.09 | .67 | 10.56 | <.001 | 5.77 | 8.40 | | | | |
| | **Control Group** | 28.92 | .95 | 30.30 | <.001 | 27.05 | 30.80 | 5598 | 562.94 | <.001 | .41 |

GLH Test (Light Treat*Post - Full Treat*Post = 0), *F* .03, *p* .85.

Therefore, full treatment's impact is no different from light treatment.

## 13.3 Effect Sizes from Differences-in-Differences Analyses

***Table 23*** below takes the parameter estimates of the regression models from the discussion above and summarizes the program effects. The column to the right, Cohen's effect size, takes the estimate and divides it by the standard deviation from the baseline assessment to come up with an effect size, so that the impacts can be compared within this EGRA Plus report and across other analyses. Note that the rows for light treatment in Table 23 are comparing light treatment with control schools, and the rows for full treatment are comparing full treatment with control schools.

It is obvious from a visual inspection that the largest impacts, with respect to effect size, occurred in letter fluency and listening comprehension, with a very large effect size found for each of the light and full treatment groups. Note that for phonemic awareness, unfamiliar words, and oral reading fluency, the light treatment did not have a statistically significant effect on reading achievement. The full treatment had an effect on each of the outcome variables. The smallest effect for full treatment (.13 SD) is found in the familiar word subtask. For oral reading fluency, the full treatment effect was moderate (.38 SD). Similar size effects are found for phonemic awareness (.27 SD) and unfamiliar words (.26 SD).

**Table 23. Differences-in-differences effect sizes and program effects**

| Outcome | Treatment group | Program effect | *p*-value | Effect size (SD) |
|---|---|---|---|---|
| Letter fluency | Light | 13.62 | <.001 | .54 |
| | Full | 19.56 | <.001 | .77 |
| Phonemic awareness | Light | .18 | .27 | *No effect* |
| | Full | .62 | <.001 | .27 |
| Familiar word fluency | Light | 1.00 | .30 | *No effect* |
| | Full | 1.79 | .07 | .13 |
| Unfamiliar word fluency | Light | -.66 | .09 | *No effect* |
| | Full | 1.56 | <.001 | .26 |
| Oral reading fluency | Light | 3.43 | .02 | .17 |
| | Full | 7.39 | <.001 | .38 |
| Reading comprehension | Light | 3.21 | .04 | .13 |
| | Full | 8.53 | <.001 | .35 |
| Listening comprehension | Light | 9.61 | <.001 | .47 |
| | Full | 9.92 | <.001 | .48 |

In short, while the program had limited time to make an impact, and faced many obstacles in doing so, children in full treatment schools in particular learned quite a bit. Comparing effect sizes, though, shows that the greatest need remains in familiar words, unfamiliar words, and oral reading fluency.

## 13.4 Other Predictors

For this section, more regression models were fit to estimate the impact of a variety of student level predictors on reading outcomes. Note that the list of models fit here (see ***Table 24***) was determined by the strength of the Pearson's correlation by the entire set of student background characteristics and student outcomes. Simplified models, combining full and light treatment schools together, were used for parsimony and to save degrees of freedom. In all these models, oral reading fluency ("ORF") is the outcome variable. The other estimates are not shown, but in each case the program is shown to have had a statistically significant impact on student achievement.

**Table 24. Regression analyses by student background predictors**

| Model | Outcome | Predictor | Coeffi-cient | Std. error | *T* | Sig. | Confidence interval | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Lower | Upper |
| I | ORF | Repeated any grade | -.07 | .04 | -2.04 | .04 | -.14 | -.00 |
| II | ORF | Overage | -.15 | .08 | -1.87 | .06 | -.31 | .01 |
| III | ORF | Repeated any grade | -.06 | .03 | -2.05 | .04 | -.14 | -.00 |
| | | Overage | -.15 | .08 | -1.89 | .06 | -.31 | .01 |
| IV | ORF | Days studied | 1.85 | .13 | 14.29 | <.001 | 1.60 | 2.11 |
| V | ORF | Parents did nothing | -4.03 | .46 | -8.75 | <.001 | -4.94 | -3.13 |
| VI | ORF | Parents helped | 2.49 | .35 | 7.16 | <.001 | 1.81 | 3.17 |

These models show several interesting things. In Model I, the main effect for "repeated any grade" is statistically significant, but the estimate is quite small. Model II estimated the impact of being overage, by grade. All children were assigned a number showing whether they were above the expected age of entry for their grade (grade 2 = 8 or 9 years, grade 3 = 9 or 10 years). The estimate shows that being one year overage was related to scoring .15 words less on oral reading fluency. Because this finding might contradict the initial finding about repetition (since repeaters are likely to be older), Model III included both main effects in the model. It shows that the negative impact was not collinear, so that there was an additional barrier placed before repeaters and those who simply entered school late. Note that this was not as a result of repetition per se, since much research in sub-Saharan Africa shows that in some cases, repetition can actually increase achievement; but because children who repeated were different in other ways from those did not, particularly with respect to family and social background. In order to estimate some of these ways, another model (IV) was fit that looked at the number of days studied per week. This showed that children who studied more did better (1.85 words per minute) than those who did not. And note that this was an additive effect, such that if a child studied 2 more days a week, his/her score would be estimated at 3.7 words more per minute. The final two models (V and VI) tested the relationship between parents' responses to negative scores on student achievement. If a parent was informed about a

poor result and did nothing, that child scored 4.0 words per minute less. Conversely, after a poor result, if students' parents helped them with reading, they scored 2.5 words more. Again, we do not argue that this is a causal impact, and that children whose parents help them are actually going to score 6.5 words more than those children whose parents ignore the results. Instead, this is a way to identify and discriminate among particular groups of children in disparate family groups, with those whose parents are working more or are poorer and therefore have less opportunity to help their children, for example.

# 14. Recommendations

In this section, we present recommendations based on the findings from the impact study. Note that some of these recommendations are related to the EGRA Plus program, and others are targeted at reading instruction in general in Liberia.

- **Ensure that adequate emphasis is placed on the achievement of boys in the more complex reading skills.** While boys' and girls' scores increased by nearly the same amount for the more simple EGRA subtasks, a gender gap remained for unfamiliar words, oral reading fluency,[19] and reading comprehension. Note that these are the areas for which reading skills and strategies are applied, rather than the relatively simple areas of letter fluency and familiar word fluency, where children can simply memorize to do better. Since the ultimate goal of EGRA Plus is to increase reading ability and comprehension, and a gender gap appears to be widening, the program should consider responding by heavily involving boys and encouraging reading for fun among young boys.
- **Move past focus on letters and words and expand focus on reading comprehension.** Gains on letter fluency in this midterm assessment were quite substantial. However, while statistically significant, the increases in other portions of reading were comparably smaller. Therefore, further emphasis should be placed on ensuring that teachers in EGRA Plus schools expand the amount of time spent on decoding, reading aloud, reading silently, and learning comprehension strategies. These strategies should be done in parallel: While effort is necessary to encourage accurate decoding, this must not be done at the expense of reading comprehension strategies.
- **Underscore decoding skills, which have relevance to higher-order skills.** While the EGRA Plus program increased students ability to read unfamiliar words, applying the decoding and phonemic awareness skills acquired from the program, the absolute gains were still very small. For the most part, children remained at basic levels of decoding skills, which has relevance with oral reading fluency, and of course, reading comprehension. It appears that children in Liberia are not very skilled (and are very slow) at unpacking new words: 77.0% of children discontinued this subtask. The program was able to allow about 10% of

[19] An unsystematic evaluation of the midterm assessment passage to understand whether it might be gender biased showed that it was not likely to be gender biased. The story was about two boys playing a game, and if anything, it was more likely to be gender biased in the other direction.

nondecoders to start (70.8% of full treatment students discontinued, compared to 81.1% of control students), but this remains insufficient for the type of reading fluency that EGRA Plus hopes to encourage.

- **Maintain test equating and recalibration as an important technical task for EGRA Plus.** It is possible that the modest progress on the reading comprehension subtask was due not to a lack of understanding among children but to the fact that while the reading passage was calibrated, the reading comprehension questions associated with the passage were not. While the technical interest of the EGRA Plus team in calibrating the midterm and baseline passage is laudable, there remain some concerns about the equality of the comprehension questions, which had similarly disparate findings with the effect of being in the midterm being -7.8 percentage points. Likewise, it appears that the listening comprehension questions were much easier than they were at baseline, with the effect of being in the post-assessment alone at 34.3 percentage points. It is highly unlikely that the entirety of this difference was related to the effect of six months of school, and a systematic comparison between these subtasks is critical.
- **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 will help to streamline reading intervention energy, and allow for within-country, rather than cross-country, comparisons.
- **Place considerably more emphasis on within-grade achievement.** While comparisons to international benchmarks are not ideal, it appears that Liberian children's progress within a grade is too modest to allow children to achieve reading fluency by grade 4 when most of 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 (in full treatment schools, 10 words per minute), 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.
- **Move beyond community knowledge of reading achievement to teach the more complex aspects of reading.** The impact of the light treatment on many of the reading outcomes shows that with a simple increase of focus on reading outcomes, student outcomes can increase. This was particularly the case in letter naming fluency. However, for the more technical aspects of reading, dependent 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 among children. In full treatment schools, relatively modest investments in teacher training can pay large dividends; the experience of light treatment

schools shows that these investments remain critical. Attention and focus is not enough. Training and skills are necessary.

- **Emphasize reading comprehension as an important goal for reading acquisition.** It should be noted that even with interventions like EGRA Plus, reading comprehension increases remain relatively modest. More work is necessary to institutionalize reading comprehension and reading fluency skills in in-service and pre-service teacher training, to increase children's ability to access these skills, and to have experience with them early in a child's educational career.

# Appendix A: Comparing the Results of Grade 2 Baseline and Grade 3 Midterm Assessments

It can be argued that a more appropriate comparison between baseline and midterm assessments would be between the grade 2 baseline students (assessed in November 2008) and the grade 3 midterm students (assessed in June 2009). This is because, due to the shifts in the education sector between baseline and midterm, little actual instruction occurred, and this comparison would be better able to capture the true impact of the program, inclusive of the summer loss that occurs in any school. In ***Table A-1,*** the column to the right, "Effect size," shows that by this specification, the program was particularly effective and highly effective in the case of the full treatment schools.

**Table A-1. Comparison of impacts: Baseline assessment for grade 2 and midterm assessment for grade 3**

| Item | School type | Baseline, grade 2 | | | Midterm, grade 3 | | | Program impact | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | Mean | Standard deviation | *N* | Mean | Standard deviation | Gains over baseline | Increase over control | Percent increase over baseline | Effect size (SD) |
| Letter naming fluency | Control | 499 | 57.9 | 23.6 | 404 | 76.7 | 23.3 | 18.8 | | 32.47% | |
| | Full | 494 | 52.27 | 25.31 | 435 | 92.32 | 21.36 | 40.05 | 21.25 | 76.62% | 0.80 |
| | Light | 545 | 55.82 | 25.32 | 438 | 91.09 | 22.32 | 35.27 | 16.47 | 63.19% | 0.62 |
| Phonemic awareness | Control | 499 | 3.26 | 2.33 | 409 | 4.3 | 2.69 | 1.04 | n/a | 31.90% | |
| | Full | 499 | 3.05 | 1.98 | 459 | 4.89 | 2.55 | 1.84 | 0.8 | 60.33% | 0.30 |
| | Light | 547 | 3.22 | 2.25 | 442 | 4.74 | 2.52 | 1.52 | 0.48 | 47.20% | 0.18 |
| Familiar word fluency | Control | 495 | 6.26 | 10.37 | 402 | 15.19 | 14.02 | 8.93 | n/a | 142.65% | |
| | Full | 489 | 4.95 | 9.48 | 434 | 19.63 | 17.45 | 14.68 | 5.75 | 296.57% | 0.32 |
| | Light | 540 | 6.88 | 12.47 | 437 | 21.08 | 19.6 | 14.2 | 5.27 | 206.40% | 0.29 |
| Unfamiliar word fluency | Control | 497 | 1.35 | 4.37 | 404 | 1.92 | 5.14 | 0.57 | n/a | 42.22% | |
| | Full | 492 | 1.23 | 4.38 | 434 | 4.01 | 6.51 | 2.78 | 2.21 | 226.02% | 0.37 |
| | Light | 541 | 2.16 | 6.4 | 433 | 3.26 | 7.92 | 1.1 | 0.53 | 50.93% | 0.09 |
| Connected text fluency | Control | 495 | 14.45 | 15.63 | 396 | 24.21 | 20.65 | 9.76 | n/a | 67.54% | |
| | Full | 488 | 12.97 | 15.81 | 427 | 36.13 | 26.21 | 23.16 | 13.4 | 178.57% | 0.77 |
| | Light | 540 | 16.03 | 18.76 | 425 | 34.25 | 27.76 | 18.22 | 8.46 | 113.66% | 0.49 |
| Reading comprehension | Control | 496 | 21.41 | 23 | 396 | 19.95 | 22.07 | -1.46 | n/a | -6.82% | |
| | Full | 494 | 16.8 | 20.55 | 427 | 29.37 | 25.83 | 12.57 | 14.03 | 74.82% | 0.59 |
| | Light | 544 | 20.48 | 22.65 | 425 | 26.64 | 24.34 | 6.16 | 7.62 | 30.08% | 0.32 |

| Item | School type | Baseline, grade 2 | | | Midterm, grade 3 | | | Program impact | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *N* | Mean | Standard deviation | *N* | Mean | Standard deviation | Gains over baseline | Increase over control | Percent increase over baseline | Effect size (SD) |
| Listening comprehension | Control | 499 | 30.58 | 20.4 | 405 | 70.32 | 33.28 | 39.74 | n/a | 129.95% | |
| | Full | 494 | 30.45 | 19.88 | 441 | 81.9 | 26.51 | 51.45 | 11.71 | 168.97% | 0.39 |
| | Light | 547 | 30.02 | 21.29 | 438 | 82.03 | 25.9 | 52.01 | 12.27 | 173.25% | 0.40 |

# Appendix B: Calibration of Baseline and Midterm Assessments

In order to prevent teaching to the test, or memorization, the midterm assessment used different word lists and passages. Although every effort was made to calibrate the difficulty ex ante using various analyses, such as Spache analysis, ex ante calibration is not good enough, in our experience.

Thus, in addition to the ex ante calibration, we also made an empirical or statistical calibration. This was done using a sample of 80 children[20] that was independent of the sample of children in any of the project schools. Children in both grades 2 and 3 were used, in several schools. Some children were given the previous (2008) passage or set of words first, and then asked to read the new (2009) passage or set of words second, at random; whereas with other children the order was reversed. This approach was intended to prevent a learning effect from biasing the results.

An analysis of the averages showed that the correlation between the two (2008 and 2009) was, in general, excellent. But the analysis also confirmed that the levels of difficulty was slightly different, with the 2009 passage harder than 2008 and the familiar word list easier in 2009 than in 2008, as indicated in ***Table B-1.***

### Table B-1. Analysis of level of difficulty, 2008 compared to 2009

| Subtask | 2008 | 2009 | Adjustment required |
|---|---|---|---|
| Familiar word fluency | 49.33 | 51.52 | 0.93 |
| Oral reading fluency | 68.06 | 54.55 | 1.26 |

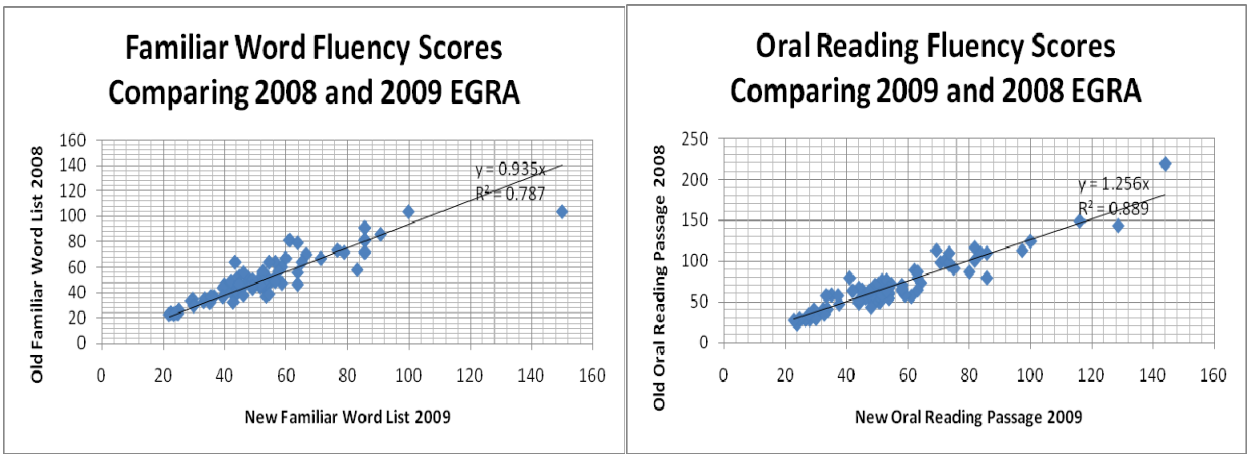
The calibration was carried out by fitting a regression line that would "predict" the score on the 2008 assessment based on performance in the 2009 assessment. The regression lines were forced through the origin, because there was no *a priori* technical reason to suspect that there would be an intercept—i.e., that children reading with 0 fluency in one test would read with positive fluency in the second test.

The results are shown in ***Figure B-1.*** The 2009 results are to be adjusted as follows:

- Connected test fluency in the 2009 passage should be multiplied by 1.26 to make it comparable to fluency in the 2008 passage.
- Familiar word fluency in the 2009 list should be multiplied by 0.93 to make it comparable to fluency in the 2008 list.

[20] Two assessments were excluded from the final data entry and analysis as the data were incomplete, which resulted in an actual sample of 78 rather than 80 as had been intended.

**Figure B-1. Scatter plots comparing scores for familiar word fluency and oral reading fluency, 2008 and 2009**



# Appendix C. Estimating the Impact of Full and Light Treatment on Outcomes, Disaggregated by Gender and Grade

(Extracted from difference-in-difference scores)

This appendix investigates whether there were discrepancies by grade and gender on the impact of both full and light treatment. While grade 2 boys achieved lower scores than expected in letter naming fluency, grade 3 boys had higher scores than expected on familiar word fluency, and grade 3 boys did better on oral reading fluency, there was not a great deal that was that far away from the aggregated findings.

**Table C-1. Analysis of discrepancies by grade and gender between full and light treatment groups**

| | | Grade 2 | | Grade 3 | |
|---|---|---|---|---|---|
| Subtask | Treatment | Boys | Girls | Boys | Girls |
| Letter naming fluency | Light | -2.22 | 12.47*** | 15.24*** | 14.01*** |
| | Full | -6.84*** | 17.06*** | 17.84*** | 21.59*** |
| Phonemic awareness | Light | .55~ | .14 | -.47 | .40 |
| | Full | .53~ | 1.05** | .02 | .83* |
| Familiar word fluency | Light | -1.66 | -.12 | 3.68 | .55 |
| | Full | 1.15 | 2.30 | 2.89 | .29 |
| Unfamiliar word fluency | Light | -.50 | -.38 | -.31 | -1.75~ |
| | Full | 1.81** | .86 | 1.85* | 1.57 |
| Oral reading fluency | Light | .40 | 2.00 | 8.82** | 1.33 |
| | Full | 8.03** | 4.96~ | 10.21** | 5.68~ |
| Reading comprehension | Light | .07 | 3.13 | 5.36 | 2.75 |
| | Full | 9.01** | 7.94** | 7.18* | 9.66** |
| Listening comprehension | Light | 14.90*** | 4.84 | 5.80~ | 12.47*** |
| | Full | 12.20*** | 9.56** | 9.48** | 7.60* |