

How Effective Are Early Grade Reading Interventions?

A Review of the Evidence

Jimmy Graham

Sean Kelly



WORLD BANK GROUP

Education Global Practice Group

January 2018

Abstract

It is imperative that students learn to read in the early grades, yet many fail to do so in developing countries. Early grade reading interventions have emerged as a common means to attempt to address this problem. This paper presents a definition of early grade reading interventions as interventions that employ a combination of five components: at a minimum, they must train teachers to teach reading using simplified instructional techniques and evidence-based curricula. In addition, they typically include in-class coaching and the provision of instructional guidelines, instructional materials, or tools for student assessment. To develop a better understanding of the effectiveness of the interventions, the paper summarizes evidence from 18 early grade reading interventions, occurring across a large variety of contexts, including four World Bank regions and three

World Bank income groups. The study finds that early grade reading interventions are consistently effective, although not infallible. The large majority had highly significant impacts on at least one reading subtask. However, only for a few interventions were effect sizes large enough to equate to more than a year's worth of schooling or create fluent readers on average. The cost of implementation varied widely, but some programs were highly cost-effective. Some programs failed to achieve impact altogether, although these programs were in the minority. In short, early grade reading interventions are not a guaranteed means to improve reading, and they rarely lead to fluency over a short span of time, but they are a mostly reliable means to make significant improvements in literacy over a short period of time.

This paper is a product of the Education Global Practice Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at <http://econ.worldbank.org>. The authors may be contacted at jimmygraham2@gmail.com and sean.kelly1926@gmail.com.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

**How Effective Are Early Grade Reading Interventions?
A Review of the Evidence**

Jimmy Graham
Center for Global Development

Sean Kelly
Management Systems International

JEL Codes

Analysis of Education (I21), Education and Economic Development (I25), Government Policy (I29)

Keywords

Education quality, education policy, literacy, early grade reading

Acknowledgements

We are thankful to Harry Patrinos for providing the initial idea for this paper and for his continual guidance and input; and to Amer Hasan, Eduardo Velez, Thomaz Alvares, Jeff Davis, Amber Gove, Helen Abadzi, Tony Barclay, and Holly Wise for their consultation and support.

I. INTRODUCTION

Despite rising enrollment rates, early grade illiteracy is widespread in the developing world. According to the most recent World Development Report, the average low-income country enrolls students at almost the same rate as the average high-income country. However, millions of students that have completed primary school in low-income countries lack even the most basic literacy skills (World Bank 2017, 5). This illiteracy prevents millions of children from taking advantage of the extensive benefits of education. Because childhood is the ideal time to learn to read, and because reading is a prerequisite for writing, advanced cognitive skills, and the ability to progress through school systems, these children are unable to take advantage of the high private returns to education and increasing demands from labor markets for advanced skills (Abadzi 2017, 8) (Montenegro and Patrinos 2014, 14) (World Bank 2016, 124).

Early grade reading (EGR) interventions have emerged as a potential solution to this problem, as they have become a common means to address illiteracy in developing countries. In this paper, we present a definition of EGR interventions as interventions that employ a combination of five components. At a minimum, they must train teachers to teach reading using simplified instructional techniques and evidence-based curricula. In addition to that foundation they contain a mix of four potential components: provision of instructional guidelines, in-school coaching and monitoring for teachers, provision of supplementary instructional and reading materials, and provision of tools and training for student assessment. While EGR interventions are not limited to including these components, they are the most commonly included.

In order to determine the effectiveness of EGR interventions, this paper summarizes the evidence from 18 evaluations that employed experimental or quasi-experimental methods to determine the impact of EGR interventions as measured by the Early Grade Reading Assessment (EGRA). These evaluations cover four World Bank regions – the Middle East and North Africa, Sub-Saharan Africa, East Asia and the Pacific, and Europe and Central Asia – and three World Bank income groups – low income, lower-middle income, and upper-middle income. The summary concerns results for letter sound recognition (LSR) and letter name recognition (LNR), which measure the basic reading skill of the alphabetic principle, as well as oral reading fluency (ORF) and reading comprehension (RC), which assess fluency and comprehension. Both raw scores and effects sizes when available are presented for these reading subtasks. When possible, the analysis considers evidence for cost-effectiveness, usually in cost per student.

We find that EGR interventions are consistently effective, though not infallible. The large majority had significant and/or substantial impacts on at least one of four reading subtasks. Typically, the impacts are highly significant and occur on various subtasks, but only for a few interventions were effect sizes large enough to equate to over a year's worth of schooling or to create fluent readers on average. Cost of implementation varied widely, but some programs were highly cost-effective. Some programs failed to achieve impact altogether, though these programs were in the minority. In short, EGR interventions are not a guaranteed means to improve reading

and they rarely lead to fluency over a short span of time, but they are a mostly reliable means to make significant improvements in literacy over a short period of time. They are thus a significant contribution to addressing illiteracy, but only a partial solution.

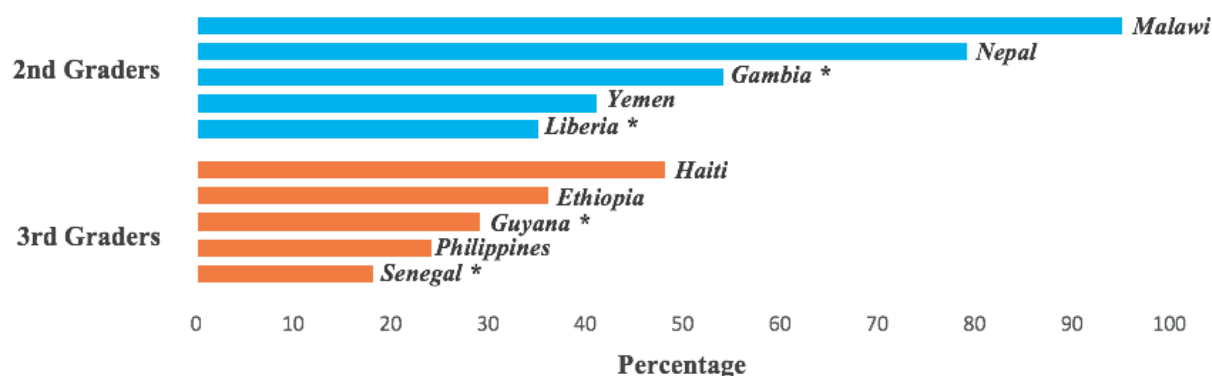
The rest of the paper proceeds as follows: Section II describes the problem being addressed by EGR interventions, Section III defines EGR interventions and summarizes the theoretical evidence for why they should improve literacy, Section IV presents the methodology used in this paper, Section V summarizes and discusses the evidence from the evaluations, and Section VI concludes.

II. LEARNING TO READ EARLY IS CRUCIAL; EARLY GRADE ILLITERACY IS WIDESPREAD

The benefits of literacy are enormous – as are the costs of illiteracy. First, there are high rates of return to education: each additional year of schooling has a 10% private rate of return (Montenegro and Patrinos 2014, 14). Second, labor market demand for advanced skills is rising, while demand for routine and manual skills is falling (World Bank 2016, 124). Since literacy is a foundational skill for learning, people cannot progress to higher levels of schooling (especially tertiary) and take advantage of high rates of return if they are illiterate, and they are not able to learn the more advanced skills for which there is a rising demand. This means that illiterate individuals will become increasingly disadvantaged relative to others, and societies may become increasingly unequal if literacy is not attained for all. Third, shortcomings in reading abilities constrain a country's economic growth because without literacy the labor force lacks the skills necessary to harness technology and enhance productivity (Easterlin 1981, 505-8, 519-21). Finally, research has shown that illiteracy leads to higher societal costs in terms of employment, education, crime, and health (Gross 2009, 6).

In order to achieve high rates of literacy, it is paramount that children learn to read around the age of five or six. Findings from neuroscience research show that the ideal time to develop the basic reading skills that lead to reading fluency is in early childhood (Abadzi 2017, 8). In addition, the costs of addressing reading problems are lower in early primary school compared to upper primary or secondary school (Gross 2009, 6). Ultimately, if a child does not begin learning to read around 1st grade, the negative effects cascade and prevent the person from achieving high levels of fluency, learning more advanced cognitive skills, and progressing to higher levels of education.

Figure 1: Percentage of 2nd and 3rd Graders that Cannot Read a Single Word



*Data taken from national samples. All other data taken from regional samples or program evaluations (Gove and Cvelich 2010, 11). All scores are for the *best-performing* language group from the sample.

Unfortunately, despite the importance of early grade reading, illiteracy is widespread in early primary school in many developing countries. Although enrollment rates have risen markedly in recent years, literacy rates are still very low. UNESCO estimates that 250 million primary school-aged children, out of a total of 650 million, are failing to acquire basic reading skills (Education for All 2014, 31). This epidemic is well illustrated through data collected by the EGRA, which has been used to determine literacy rates using a variety of indicators, including whether or not a student can read at least one single word. This indicator is used in Figure 1 to show illiteracy rates across a variety of countries from five World Bank regions. Although this figure provides only a snapshot of literacy rates across a hand-picked selection of countries, for which only some of the rates are nationally representative, it still provides a clear message: large portions of students in many countries, from a wide variety of regions, are completing several years of schooling while remaining completely illiterate. In Gambia, for example, over half of all 2nd graders cannot read a single word. In Guyana, almost 30% of all 3rd graders are zero-word readers (Gove and Cvelich 2010, 11).

III. A POTENTIAL SOLUTION: EARLY GRADE READING INTERVENTIONS

In recent years, a large number of projects have sought to address these early-grade literacy deficiencies. We propose the term EGR interventions to refer to these projects, and define them as having a combination of five components which have been especially prevalent:

1. Training teachers to teach literacy with simplified instruction and evidence-based curricula (the core component)
2. Providing instructional guidelines
3. Following up with coaching and monitoring
4. Providing supplementary instructional materials
5. Providing tools and training for student assessment

Not every intervention includes all five of these aspects. However, every intervention does include the first component: training teachers and/or providing basic instructional guidelines that teachers can follow – this is the core component. All interventions also included supplementary materials, though this component was not typically the main focus of the programs. Aside from that, interventions employ some mix of the other components and, in some circumstances, additional aspects not listed here. They must also be targeted at the early grades, which we consider to be 1st through 4th grade simply because the vast majority of literacy interventions are targeted at those grades. The remainder of this section explains each element and summarizes the theoretical evidence for why it should improve literacy.

1. Training Teachers to Teach Literacy with Simplified Instruction and Evidence-Based Curricula

Many teachers in developing countries have little-to-no training in teaching literacy, and are often undereducated in general (Pryor et al. 2012, 409-502). According to UNESCO, over a quarter of all teachers in developing countries are not trained according to national standards (Education for All 2014, 6). Given the evidence that reading should be taught in a very specific way, this is an obvious problem. To fill this gap, effective EGR interventions provide capacity development to teachers in multiple ways. First of all, they train teachers. Training should move beyond the large, conference-style, one-off professional development workshops that are so common in many countries (Abadzi 2006, 127-30). Such conferences have proven to be largely ineffective, especially compared to more personal and extensive trainings that focus on practical skills and occur continuously (World Bank 2017, 131-2). Thus, to be considered an EGR intervention, programs must include a series of intensive trainings.

Trainings need to focus very specifically on literacy instruction that is appropriate to the context. Teacher trainings are most effective when they are catered to a specific subject matter (World Bank 2017, 132). And when curricula and learning goals outpace student abilities, learning outcomes tend to be lower (Pritchett and Beatty 2015, 276-88). Even when activities concentrate on reading, if learning objectives are not based in the reality of students' abilities, then instruction tends to become less effective. All too often, approaches to teaching literacy have their basis in curricula formulated in developed countries. Transplanting these curricula to developing countries sets unrealistically high expectations that undermine educational quality (Abadzi 2017, 2-3). Currently, a multitude of curricula are based on overreaching expectations (ASER Center 2011, 6). By utilizing instructional materials and curricula based on students' levels of attainment and literacy, EGR interventions are much more effective at teaching students to read (Kim et al 2016, 9-12).

Related to matching instruction to ability, effective EGR interventions also train teachers to teach reading in ways that are based on evidence from cognitive research. This research shows that reading takes place through a strict sequence, where one skill builds on another (Marinelli et al 2011, 96-98). If this sequence is not followed, literacy acquisition can take much longer, or not occur at all. The literature also shows that the sequence is not necessarily intuitive or the

same for all languages (Abadzi 2013, 9-11). Likewise, the best way to teach literacy depends on the amount of time and resources available (Abadzi 2017, 13). However, curricula in developing countries are often based on curricula developed for western languages and contexts, regardless of their suitability (Abadzi 2013, 9). Thus, EGR interventions often train teachers using literacy curricula that are appropriate to the context and based on sound evidence for how to teach literacy to a specific group.

2. Providing Instructional Guides

Where teacher capacities are low, teachers may need clear instructional guidelines in addition to trainings. These guidelines should allow teachers to develop simple literacy-instruction routines (Abadzi 2013, 34-38). The ideal guide should provide step-by-step instructions without too many words or complex procedures (Allan and Horn 2013). The guidelines may even have interactive content. For an intervention in Jordan, instructional guidelines took the form of videos (Brombacher et al 2014, 8). The best guidelines should also give scripted lessons for every instructional day of the semester. Regardless of the format, teachers should receive training before implementing them (Kim et al 2016, 10-11).

3. Following Up with Coaching and Monitoring

To compound the effectiveness of teacher trainings, EGR interventions can provide ongoing monitoring, feedback, and in-classroom coaching to teachers. Such measures ensure that teachers correctly apply what they have learned in training. Typically, teachers struggle to retain and put into practice new knowledge gained from trainings (Clark-Chiarelli and Louge 2016, 30-40). It is also common for teachers to only partially follow instructional guides (RTI International 2011b, 7). Coaching and continual feedback mollify these problems. Some of the most effective trainings involve modeling from a coach (Abadzi 2006, 127-130). In general, teachers tend to apply more of what they learned when trainings feature in-classroom coaching and feedback (Kim et al 2016, 53-54).

4. Providing Supplementary Instructional Materials

It is difficult to teach children to read without basic materials such as books. Ensuring the classrooms have reading materials is a key component in teaching students to read (Kim et al 2016, 13-15). But the availability of texts in developing country schools is often woefully inadequate (Kim et al 2016, 13-15). While the provision of materials alone does not transform learning outcomes, reading materials are fundamental for instruction once teachers have training to use them in reading instruction (Allan and Horn 2013). However, books should cater to the local culture and to the reading levels of the students (Malik et al 2015, 2). In other words, simply giving books is insufficient. The books need to match the students' abilities and needs, and the teachers need to know how to teach reading effectively with the books.

5. Providing Tools and Training for Student Assessment

Teachers need to understand what their students do not know in order to adjust instruction. Assessments are therefore a critical part of effective literacy instruction (Kim et al 2016, 12-13). Given literacy development's step-by-step nature rooted in a rigid sequence of skill acquisitions, teachers should know where each student lies on the spectrum of development. Without proper assessment, teachers cannot cater instruction. Effective EGR interventions should train teachers and give them the tools to conduct assessments, analyze results, and modify their instruction.

Assessment directly links to the accountability of schools to the community. Education specialists broadly recognize assessment as complementary to accountability and autonomy so that education systems respond to local needs with quality education. Ideally, the presence of autonomy, assessment, and accountability fosters more time-on-task, effective teacher training, and attention to subject matter leading to higher education quality (Patrinos et al 2013, 4-6). Primarily, assessment of students' progress in reading skills presents an opportunity to strengthen accountability. If teachers properly explain assessment results, parents and school administrators will understand students' progression and respond accordingly.

IV. METHODOLOGY

In selecting the evaluations to include in our summary evaluation, we included only projects that could qualify as EGR interventions, i.e. those that included the first component listed above along with some mix of the other components, and which were targeted towards 1st to 4th graders in developing countries. Among the EGR interventions, we only included evaluations that used the EGRA to measure reading performance. We decided to only focus on interventions using the EGRA because it is the most prevalent tool for measuring early-grade literacy in developing countries, it has easy-to-interpret metrics, and focusing on only one measurement tool allowed us to easily compare results. Among these EGRA-evaluated EGR interventions, we only included those that used an experimental or quasi-experimental design for the evaluation. Specifically, we did not keep studies if they did not utilize a control group for comparison. Our search for evaluations was conducted through the document archives of major donors and education initiatives in international development,¹ bibliographic databases,² and discussions with experts

¹ USAID Development Experience Clearinghouse, USAID Early Grade Reading Barometer, USAID EdData, USAID Education Data for All Children Reading, World Bank eLibrary, World Bank Projects and Operations, the Global Partnership for Education Library, the UNESCO Library, the SIDA Unit for Research Cooperation, the World Bank's IE2 Impact Evaluation Repository, the Inter-American Development Bank's Evaluation Reports database, the World Bank's Development Impact Evaluation Initiative database, DFID's Research for Development database, the DAC Evaluation Resource Center, the Asian Development Bank's Evaluation Reports database, the African Development Bank's Evaluation Reports database, 3ie's Registry for International Development Impact Evaluations.

Our search terms included: "EGRA," "Early Grade Reading Assessment," ("Early Grade" OR "Primary School") AND ("Reading" OR "Literacy"), "Early Grade" AND ("Literacy" OR "Reading"), "Reading," ("Reading" OR "Literacy" OR "Primary School") AND ("Intervention" OR "Program" OR "Project").

in the field.³ We also used reference snow-balling: among the papers that our search returned we looked for references to other EGR interventions. The result was a sample of 18 evaluations, 17 of which were publicly available, and one of which (from Papua New Guinea) relied on as-of-yet unpublished documents. These 18 included the Democratic Republic of Congo (Bulat et al 2014; RTI 2011c), the Arab Republic of Egypt (RTI 2014b; Gove et al 2017), Jordan (RTI 2014a; Brombacher et al 2014; Gove et al 2017), Kenya (RTI 2014c; RTI 2014a), the Kyrgyz Republic (AIR and Save the Children 2015); Liberia starting in 2008 (RTI 2011a), Liberia starting in 2011 (King et al 2015), Malawi starting in 2010 (Tilson et al 2013; RTI 2014a), Malawi starting in 2013 (Nagarajan et al 2015), Mali (Spratt et al 2013), Mozambique (Raupp et al 2016), Nigeria (RTI 2015), Papua New Guinea (World Bank 2016a; Macdonald 2016; Williams and de Silva 2015a), the Philippines (EDC 2014; EDC 2015; Social Impact 2016); South Africa (Hollingworth and Gaines 2009; Piper 2009), Tanzania (CAI 2013; CAI 2014; CAI 2015; School-to-School 2015), Tonga (Macdonald et al 2017), and Uganda (NORC 2017).

In examining these evaluations, we focused on a few select indicators of literacy. Primarily, we examined Oral Reading Fluency (ORF) scores. To generate an ORF score, students are given a passage to read. Their score then indicates the number of correct words per minute that they are able to read. We focused on this measure because it is one of the most commonly used metrics in the literature on early grade reading, and because it is one of the most easily understood and crucial measures of literacy. The ability to quickly read words is a core literacy skill, and the ORF score is closely correlated with reading comprehension, which is the ultimate goal of literacy. In other words, if an intervention is improving average ORF scores, it is clearly improving literacy. The ORF score also allowed us to determine the percentage of students that cannot read a single word, i.e. the students that scored zero for ORF. This measure allowed us to see the proportion of students that had achieved even the most basic levels of literacy.

We also looked at two more basic measures of literacy: Letter Sound Recognition (LSR), which measures the number of correct letter sounds per minute that students are able to name, and Letter Name Recognition (LNR), which measures the number of correct letter names per minute that students are able to name. These indicators measure skills that are even more basic than word reading, and thus allowed us to see the impact of interventions on foundational skills. We examined both LNR and LSR because some studies included one indicator and not the other. We refer to LNR and LSR scores in terms of correct letters per minute and correct letter sounds per minute, respectively

To measure more advanced literacy skills, we looked at reading comprehension (RC) scores. This indicator is the percentage of questions (usually out of four to five questions) that a student is able to correctly answer based on the passage that they read for the ORF test. We used this

² JSTOR, SAGE Journals, Google Scholar, International Association for the Evaluation of Educational Achievement, Education Resources Information Center, National Bureau for Economic Research, EconLit, International Development Abstracts.

³ Thomaz Alvares, Jeff Davis, Amber Gove, Harry Patrinos, and Eduardo Velez.

measure because it allowed us to see the impact of interventions on more advanced skills. Changes in RC scores are examined in terms of percentage point changes. By using this set of indicators, we were able to examine three of the main levels of literacy development: partial alphabetic (with LSR and LNR), consolidated alphabetic (with ORF), and automatic (with RC) (Dubeck and Gove 2015).

To summarize the interventions, we examined impact in several ways. We looked at the difference-in-differences (DiD) between treatment and control for the average score for each indicator, the level of significance, and the effect size in terms of standard deviations.⁴ However, for each intervention, a varying degree of data and information is available. Effect sizes permit us to compare results on subtasks across the different EGR interventions using EGRA since raw scores in different languages are not directly comparable. That being said, effect sizes require contextualization in order to understand their meaning. Jacob Cohen asserted that small effect sizes ranged from 0.2 standard deviation to under 0.5 standard deviation, a medium effect from 0.5 standard deviation to under 0.8 standard deviation, and a large effect anything 0.8 standard deviation or greater. Although, Cohen recommended these ranges as general guidelines and offered that relative effect sizes could adjust depending on the area of research (Cohen 1988, 25-27). Within the context of education research, the U.S. Department of Education’s Institute of Education Sciences advises that “effect sizes of 0.25 standard deviations or larger are considered to be substantively important. Effect sizes at least this large are interpreted as a qualified positive (or negative) effect, even though they may not reach statistical significance” (U.S. Department of Education 2014, 23). Carolyn Hill, Howard Bloom, Alison Black, and Mark Lipsey established benchmarks for effect sizes and a year of learning in primary school for reading and math. The effect sizes equivalent with a year of learning in reading vary by grade: Grade 1 equals 0.97 standard deviations, Grade 2 equals 0.6 standard deviations, Grade 3 equals 0.36 standard deviations, and Grade 4 equals 0.4 standard deviations. The data used for calculating the effect sizes in reading came from seven standardized tests in the United States (Hill et al 2008, 173-174). In addition, Hill and others specify a type of effect size benchmarking for an evaluation synthesis in which the effect sizes for similar intervention types are collected and described (Hill et al 2008, 175-177). For our analysis, we consider any effect size over 0.25 standard deviations as substantial. Where possible, we will contextualize the effect sizes according to the grade-level appropriate one year of learning benchmark. Finally, we will present a descriptive analysis of EGR effect sizes delineated by subtask in order to understand the range, mean, median, and standard deviation for the accumulated evidence.

We also summarized impacts by presenting the actual levels of achievement as described by the indicators, rather than only focusing on the impact on the treatment group relative to the control group. Specifically, we present the average ORF scores at baseline and endline for treatment and control, the percentage of students in the treatment group at endline that are zero-word readers, and the average reading comprehension score in the treatment group at endline. By doing so, we show how advanced the average student actually is, and what proportion of students are still

⁴ For studies with DiD analysis, we used the endline difference in scores.

almost fully illiterate. This is important because some interventions have large effects but the students are still largely illiterate. Others have small effects but students in both the control and intervention groups are performing at a relatively high level and improving.

In terms of discussing reading fluency, more than the absolute change in ORF matters. The relative share of students who can fluently read in a language merits attention too. In practice this requires the creation of language-specific and sometimes grade-specific ORF benchmarks. A child who can read at or above the established cut point of correct words per minute is considered fluent and thus more likely to be able to read for comprehension. In general, reading fluency occurs between ORF scores of 45 to 60 correct words per minute (Abadzi 2006, 37). Although fluency levels vary widely for different languages, this range can still be helpful as a very rough guide. Therefore, we will present the mean endline ORF scores from the evaluations relative to the 45 to 60 correct words per minute range to approximate how well EGR interventions move children towards reading fluency.

In determining cost-effectiveness, we looked only at the technical cost per student. These technical costs exclude the costs for EGRA development, intervention development, post intervention, and administration. By looking only at technical intervention costs, the cost figure represents the cost to continue project implementation. To achieve per student costs the total technical EGR costs are divided by the number of students. There are numerous caveats to cost-effectiveness figures. First, not all evaluations had the same time frame. Likewise, some evaluations incurred interruptions, such as strikes in Kenya, that reduced implementation time and costs. Also, the number of students in classrooms, which can vary widely, affects cost-effectiveness. Higher student to teacher ratios, as in Malawi, lowers the cost of implementation per student, but this lower cost partially reflects an obvious deficiency in the education system. Furthermore, different contexts have different needs. In Liberia, for example, basic classroom requirements consumed more resources. In other countries, EGR components account for a lot of costs. Finally, program implementation happens on vastly different scales. By reaching more students, a project may reach a certain economy of scale which reduces costs per student. For these reasons, it is difficult to compare costs across countries and reach general conclusions about how much EGR interventions should cost. However, the data on cost-effectiveness provide a general idea of how much EGR interventions might cost. As a final note, most interventions did not include cost information, so we included it only where possible.

Finally, there are certain aspects of the studies that we will not try to summarize and evaluate. This paper aims to look at the average effect of EGR interventions in different contexts. Thus, although there is information available to do so, we will not summarize the findings in terms of how interventions affect different subgroups, such as language groups, genders, and socioeconomic backgrounds. Although such analyses are important, they are beyond the scope of this paper.

V. SUMMARY OF EVALUATION RESULTS

In this section, we provide a summary of the findings from the evaluations along with a discussion of what can be learned from these findings. We focus on describing the interventions themselves, magnitude and significance of effects, the endline results for treatment groups, cost-effectiveness, and a variety of contextual factors that defined each intervention. This section is broken into seven subsections that summarize various aspects of the interventions.

When considering the analyses below it is important to keep in mind that some evaluations had multiple estimates across different grades or languages (occurring in the same context) of either repeated positive impact (e.g. Mozambique) or negative impact (e.g. the Philippines), which could bias the results upwards or downwards, respectively. In essence, studies with more estimates received more weight than others. To deal with this issue, we not only provide the simple mean of all estimates (i.e. each row in Appendix 1) but also the ‘program mean.’ For the latter, we first take the mean of the indicator of interest for each program, and afterwards take the mean of each program’s respective mean.

Another point to consider is that different data points were missing for different interventions. For example, Uganda failed to report effect sizes for insignificant results, which could bias results upwards, and the study from Mozambique did not report effect sizes despite being highly significant, which could bias results downwards. Therefore, in order to have a single measure that is available for (nearly) all observations and studies, we define each observation and study as having either a ‘substantial impact’ or ‘unsubstantial impact.’ We define a substantial impact observation as having an effect size of at least 0.25 standard deviations or significance at $p < 0.05$ for any of the observed subtasks (see Appendix 1). We define a substantial impact study as having at least 50% substantial observations. Overall, there were 35 substantial observations, 15 unsubstantial, and 3 unknown (the LTTP II study did not report significance or effect sizes), and 14 substantial studies, 3 negative, and 1 unknown (see Table 1).

1. Summary of the Interventions and Their Evaluations

In order to provide context to the results of the evaluations, this sub-section provides a description of the interventions and how they were evaluated. Figure 2 provides a summary of the components that each intervention contained. All of them provided teacher training and instructional materials, most also provided instructional guides and follow-up coaching, and a minority provided tools for student assessment. Most interventions also contained additional components, which are described in Appendix 5.

Figure 2: Components Included in Each EGR Intervention

Country (program)	Training Teachers with Evidence-Based Curricula	Providing instructional guidelines	Following up with coaching and monitoring	Providing instructional materials	Providing tools and training for student assessment	Other components
DRC	x	x	x	x		x
Egypt	x	x	x	x	x	x
Jordan	x	x	x	x		
Kenya	x		x	x		
Kyrgyz Republic	x		x	x		x
Liberia (EGRA Plus)	x		x	x	x	x
Liberia (LTTP II)	x	x	x	x	x	x
Malawi (MTPDS)	x	x	x	x		
Malawi (EGRA)	x	x	x	x		x
Mali	x	x		x	x	
Mozambique (post-endline)	x	x		x	x	x
Nigeria	x		x	x		x
PNG	x	x	x	x		x
Philippines	x	x	x	x		x
South Africa	x	x		x		
Tanzania	x		x	x		x
Tonga	x		x	x		
Uganda	x	x	x	x		x

x = component included

In terms of methodology, there are a few common trends. First, each evaluation involved the random selection of students from participating schools, unless the school was especially small in which case all students were involved. Typically, somewhere between 10 and 20 students were selected per grade from each school. Second, most studies were not longitudinal. For the evaluations that were carried out within a single school year, all of them were longitudinal at the cohort level, such that the same cohort of students was examined at baseline and endline. For the evaluations spanning multiple school years, most were not longitudinal at the cohort level. Rather, they involved assessing the same grade level but a different group of students. The only exception was the program in Uganda. Furthermore, the only study that was longitudinal at the student level was the 2015 evaluation in the Philippines; all others randomly selected students at both baseline and endline. Unless otherwise noted, these random selections at least took place in the same schools.

There are also methodological differences across studies that may affect estimates of impact: most interventions measured a year of impact but some measured less (South Africa), and others measured more (Uganda); some were more rigorous than others (although all were either experimental or quasi-experimental); and there were large variations in the amount of time that elapsed between baseline and endline. Appendices 5 and 6 explain these methodological differences.

2. Summary of Program Impacts

Table 1 shows the average impact and effects for each study, as well as the categorization of each study as substantial or not. It shows that 82.35% of the studies (excluding the unknowns) can be considered to have had substantial impacts. It also shows that effect sizes were substantial (i.e. above 0.25 standard deviations) on average, though they were more substantial for the less advanced subtasks. However, the table also makes clear that, although there were large effect sizes, the absolute DiD averages were rather small, especially for ORF and RC: 6.15 correct words per minute and 5.62%, respectively.

Table 1: Average Impacts and Effects for Each Program

Country (program)	DiD, LSR/ LNR (clspm/ clpm)*	Effect size, LSR/ LNR	DiD, ORF (cwpm)	Effect size, ORF	DiD, RC (%)	Effect size, RC (%)	DiD, % zero word readers	Substantial impact?
Congo, Dem. Rep.	0.17	0.01	3.07	0.13	3	.	-5	No
Egypt, Arab Rep.	19	1.07	12	0.55	.	.	.	Yes
Jordan	11.6	0.48	6.3	0.46	.	.	1.9	Yes
Kenya	.	0.68	.	0.41	.	0.4	.	Yes
Kyrgyz Republic	.	0.23	.	0.27	.	0.19	.	Yes
Liberia (EGRA Plus)	14.8	0.52	21.2	0.8	25.2	0.82	.	Yes
Liberia (LTTP II)	11.1	.	7.03	.	7.67	.	.	.
Malawi (MTPDS)	20.14	.	8.88	.	11	.	-43.7	Yes
Malawi (EGRA)	5.35	.	1.01	.	-0.01	.	-10.8	No
Mali	.	0.63	.	0.33	.	.	.	Yes

Mozambique	8.8	.	3.03	.	3.18	.	.	Yes
Nigeria	6.1	0.78	4.9	0.66	0.1	0.6	.	Yes
PNG	.	.	.	0.3	.	0.12	.	Yes
Philippines	3.12	.	2.16	.	3.96	.	.2	Yes
South Africa	14.32	.	7.21	0.8	9	0.59	1.13	Yes
Tanzania	.	.	0.07	.	-1.5	.	-12.4	No
Tonga	.	0.31	.	0.24	.	0.14	.	Yes
Uganda	3.4	0.55	3.04	0.35	0.24	0.35	.	Yes
Average	9.83	0.53	6.15	0.44	5.62	0.4	-9.8	82.35%

Averages are taken as the simple mean of all of the estimates produced for each study.

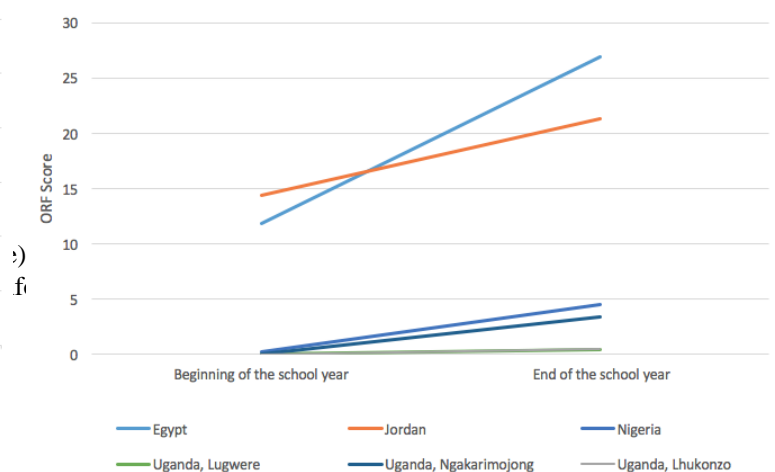
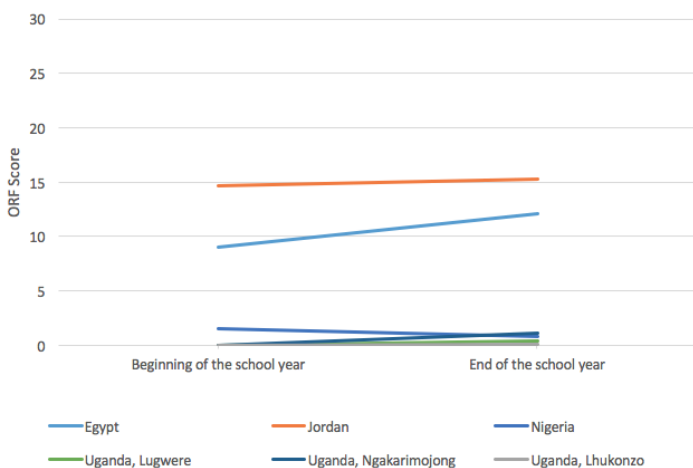
Acronyms: DiD (difference-in-differences), LSR (letter sound recognition), LNR (letter name recognition), ORF (oral reading fluency), RC (reading comprehension), cwpm (correct words per minute), clpm (correct letters per minute), clspm (correct letter sounds per minute).

*If both LSR and LNR is available for a study, the average of the two together is given.

Given the generally low level of learning in these contexts, large effect sizes can clearly still equate to somewhat small gains in absolute terms. This concept is illustrated in Figures 3 and 4, which show the average baseline and endline scores for control and treatment, for all substantial impact observations with endline assessments occurring later in the school year than the baseline.⁵ Figure 3 shows that, in the absence of intervention, the gains that occur are very modest or likely statistically insignificant. Figure 4 shows that gains with the intervention are still too small to bring students to fluency (sometimes far too small), but they at least represent major improvements over the status quo.

Figure 3: Control Group Average ORF Scores, at Baseline and Endline

Figure 4: Treatment Group Average ORF scores, at Baseline and Endline



Aside from the trend of modest-yet-substantial impacts, some programs have in fact had very large impacts. MTPDS, for example, reduced zero-word readers by 43.7 percentage points and improved letter reading by 20.14 letters per minute. EGRA Plus improved ORF scores by 21.2 correct words per minute and RC by 25.2 percentage points.

3. Descriptive Analysis of All Subtask Effect Sizes and Significance

The descriptive analysis of the effect sizes in Table 2 shows that the 18 evaluations for EGR interventions have produced 19 or more estimates for LSR, ORF, and RC.⁶ We do not include LNR in the descriptive analysis because there were only three estimates for effect sizes. While ORF and RC have similar ranges, LSR has a greater minimum value and maximum value than both, meaning the LSR range is substantially different. This similarity carries over to the mean effect size. ORF and RC have similar mean effect sizes, 0.38 standard deviations and 0.34 standard deviations respectively; while LSR’s mean effect size is 0.63 standard deviations. Most strikingly, the mean effect sizes for LSR, ORF, and RC are greater than 0.25 standard deviations (for both the standard mean and unweighted mean), the minimum benchmark for a qualified positive effect. None of these means are distorted by outliers beyond three standard deviations.

Table 2: Descriptive Analysis of Effect Sizes

Subtask	Obs.	Min	Max	Mean	Median	SD
LSR	19	0.26	1.07	0.63	0.63	0.17
ORF	25	-0.02	0.8	0.38	0.36	0.19
RC	22	-0.03	0.82	0.34	0.33	0.21

Includes all reported estimates.

Table 3 describes the percentage of impacts, for various subtasks, that were significant at various levels of significance. One can see that, for all subtasks, the majority of impacts were significant at least at the 10% level. And among the significant impacts, the majority was significant at $p < 0.01$. Furthermore, 73% of observations were significant for at least one subtask. Significant impacts were least common for RC.

⁶ For both Table 2 and Table 3, for the observation for PNG, only the regression that produced the *lowest* level of significance or effect size was used.

Table 3: Percentage of Significant Impacts for Various Subtasks

Subtask	Obs.	p < 0.1	p < 0.05	p < 0.01	Total*
LNR	10	10%	0%	70%	80%
LSR	32	3.1%	6.3%	47.8%	63.5%
ORF	41	12.2%	7.3%	39%	63.4%
RC	38	10.5%	7.9%	28.9%	52.6%
All**	45	.	.	.	73%

Based on all reported estimates.

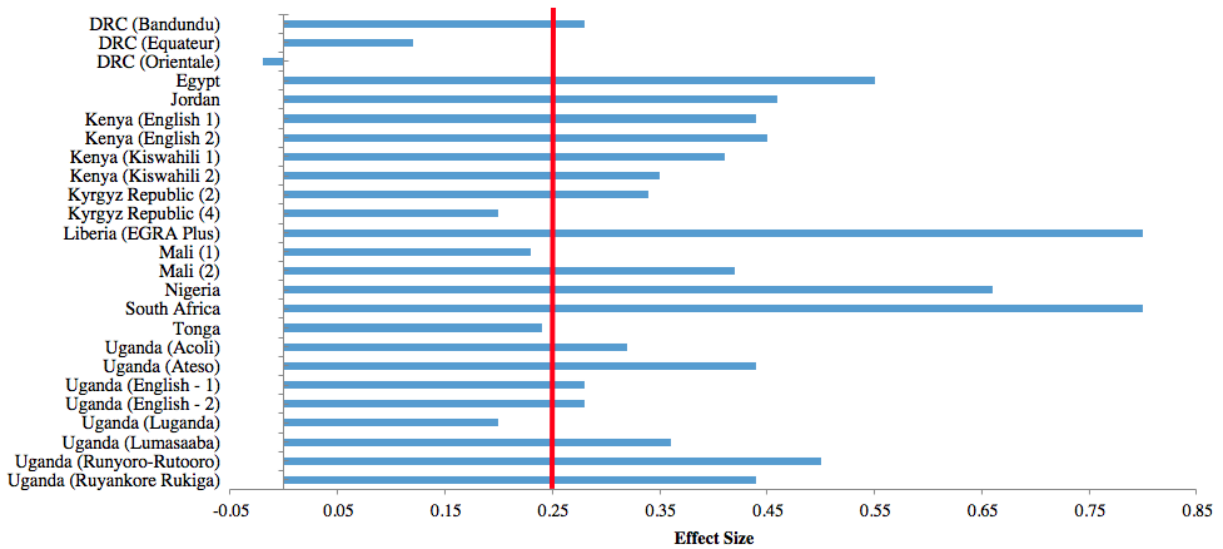
*Percent significant at either p<.1, p<.05, p<.01; includes those which were significant at least p<.1 without determining at which specific level

** “All” shows the percentage of observations that were significant at p<.1 for at least one subtask

4. Summary of ORF Effect Sizes

Pivoting to examine ORF in greater detail, the most basic level for a substantial positive result for an effect size is 0.25 standard deviations. As Figure 5 demonstrates, 19 (76%) of the ORF effect sizes from the evaluations satisfy this basic benchmark of a qualified positive result.

Figure 5: ORF Effect Sizes Compared to the 0.25 SD Benchmark



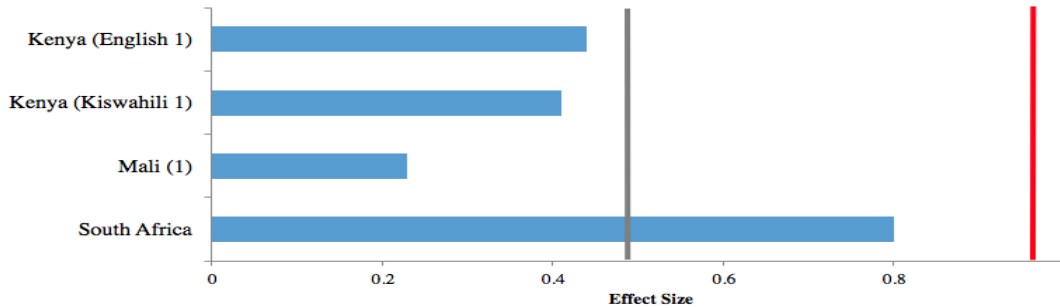
Grades and languages are in parentheses. All reported effect sizes are included.

The red line signifies the .25 SD benchmark, a substantively important effect

Examining ORF by the grade-specific benchmarks reveals an interesting picture. To do so, we only look at evaluations conducted at the baseline and endline for the same grade. As Figure 6

shows, none of the four effect sizes for 1st grade achieve 0.97 standard deviations, one year of learning. However, the students of SMRS in South Africa came close at 0.8 standard deviations and it was the only result where students eclipsed more than one-half a year of learning for 1st grade.

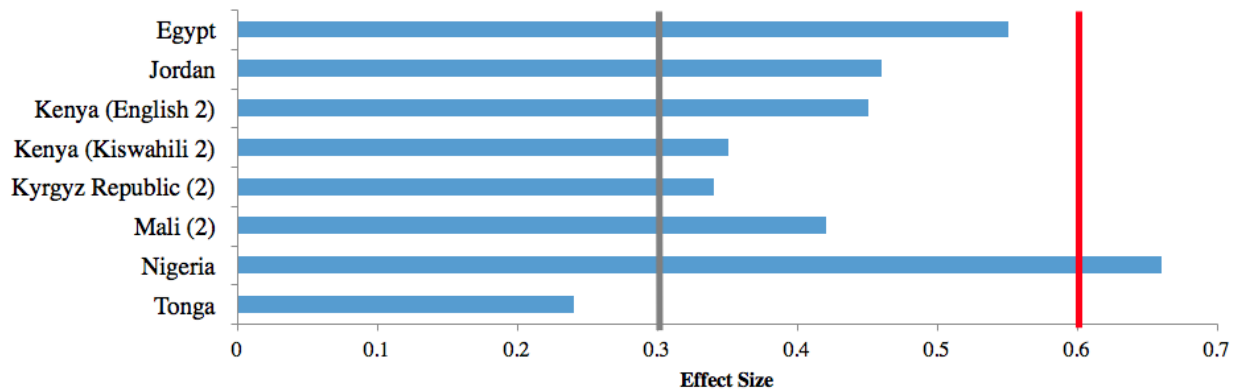
Figure 6: 1st Grade ORF Effect Sizes Compared to the 0.97 SD Benchmark and the 0.485 SD Benchmark



Grades and languages are in parentheses. All reported effect sizes for 1st grade are included. The red lines signify the 0.97 SD Benchmark (one year of learning for 1st grade) and the grey line signifies the 0.485 SD Benchmark (0.5 years of learning for 1st grade)

Most of the ORF effect sizes come from 2nd grade students - eight in total. In this case 0.6 standard deviation is equivalent to one year of learning. Figure 7 shows that students of RARA in Nigeria surpassed the benchmark, while another six ORF effect sizes exceed one-half a year of learning.

Figure 7: 2nd Grade ORF Effect Sizes Compared to the 0.6 SD and the 0.3 SD Benchmark

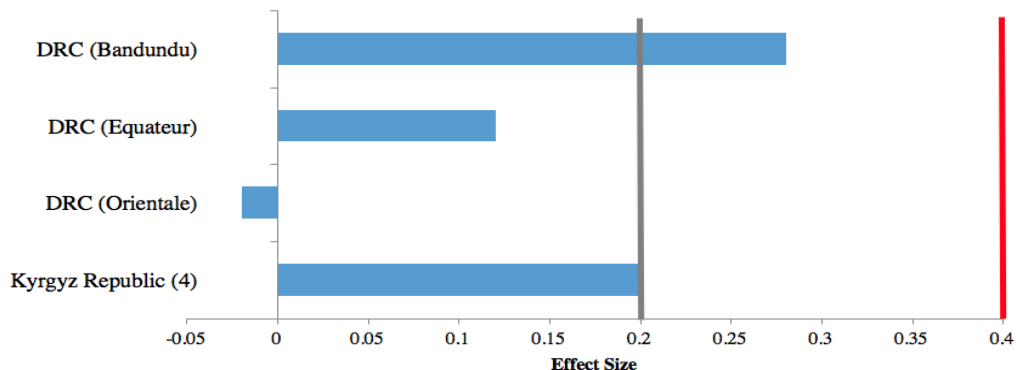


Grades and languages are in parentheses. All reported effect sizes for 2nd grade are included. The red lines signify the 0.6 SD Benchmark (one year of learning for 2nd grade) and the grey line signifies the 0.3 SD Benchmark (0.5 years of learning for 2nd grade)

There were no effect sizes for 3rd grade students; there were four for 4th grade students. One year of learning for these students equates to an effect size of 0.4 standard deviations. Two of the

effect sizes are greater than one-half a year of learning; none of the effect sizes reach one year of learning.

Figure 8: 4th Grade ORF Effect Sizes Compared to the 0.4 SD Benchmark and the 0.2 SD Benchmark



Grades and languages are in parentheses. All reported effect sizes for 4th grade are included.. The red lines signify the 0.4 SD Benchmark (one year of learning for 4th grade) and the grey line signifies the 0.2 SD Benchmark (0.5 years of learning for 4th grade)

The analyses in these first three subsections show that the large majority of EGR interventions have either substantial and/or significant impacts, and the majority has significant impacts on all three reading indicators that we selected. Thus, EGR interventions typically impact everything from fundamental to advanced reading skills. Effect sizes are consistently substantial, across various grades, but many do not cross the half-year threshold, and very few cross the full-year threshold. Likewise, although some interventions have had DiD of a large magnitude relative to fluency benchmarks, most have not. So, while EGR interventions often produce impacts that are substantial and significant (especially relative to the status quo of very slow progress) they are far from a silver bullet approach to early grade literacy. Furthermore, there are several instances, albeit relatively uncommon, where the interventions seem to have little impact. So, the interventions are consistently effective, but not infallible.

These generally positive findings are consistent with the broader literature on education interventions. A recent systematic review of education interventions found that the two interventions with the largest effects were adapting teaching methods to students’ skills and recurrent teacher trainings (Evans and Popova 2015, 12-15). An even larger systematic review asserted that “structured pedagogy programmes have the largest and most consistent positive average effects on learning outcomes” (Snilstveit et al 2015, iv). Another review claimed that “almost all successful instructional interventions... include at least a minimal attempt to develop teachers’ capacity to deliver effective classroom instruction” (McEwan 2015, 280). According to the 2018 World Development Report on Education, teacher trainings are most effective when they teach practical skills, are specific to a subject, and occur continuously - much like and EGR intervention (World Bank 2017, 131-133). While these reviews did not explicitly study literacy, the broader education literature clearly agrees that training teachers on evidence-based

pedagogies is largely effective.

5. ORF Treatment Endline Scores

The endline raw ORF scores for treatment students exhibit variation between 0.44 correct words per minute and 61.2 correct words per minute. Reading fluency occurs somewhere in the range of 45 to 60 correct words per minute. Using that metric, Figure 9 shows that only four groups of students achieved reading fluency: 2nd graders in Kenya for English, 4th graders in the Kyrgyz Republic, students in Liberia, and 3rd grade students in the Philippines reading English and Filipino.

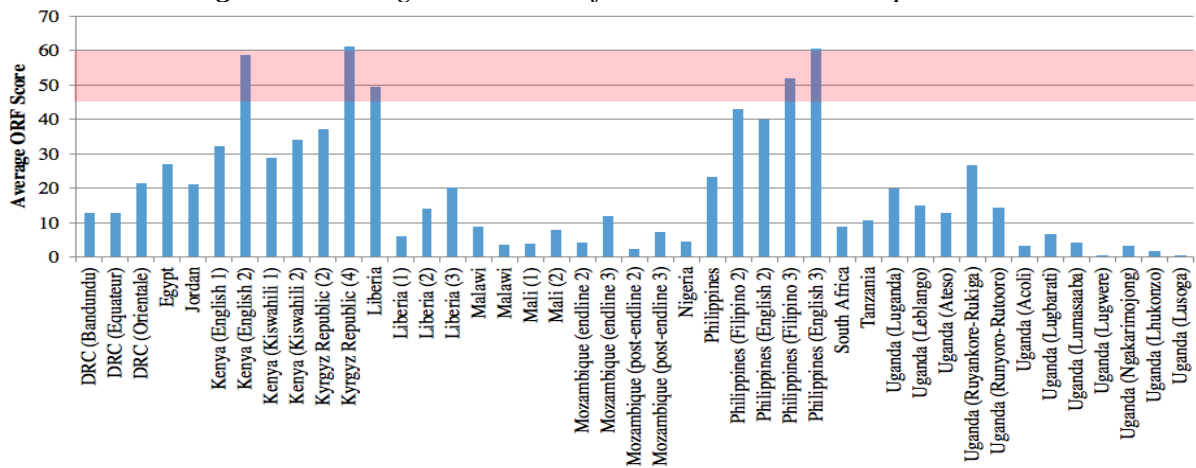
Table 4: Average Endline Score/Percentage for the Treatment Group

Indicator/Subtask	Mean
ORF	19.5 correct words per minute
Zero word readers	34.6%
RC	18.7%

Based on all reported estimates.

Similar to Figure 9, Table 4 shows high rates of zero word readers, low rates of RC, and low ORF scores relative to fluency levels. And as Table 5 indicates, even when ORF scores are grouped by grade, these low averages are not merely due to earlier grades (which one would expect to have lower scores) bringing the average down.

Figure 9: Average ORF scores for the Treatment Group at Endline



Grades and languages are in parentheses. All reported ORF scores included. The red area indicates the approximate level of fluency needed to achieve reading comprehension.

Table 5: Average Endline ORF Scores for the Treatment Group by Grade

1 st	2 nd	3 rd	4 th
17.2	21.2	30.5	13.6

All reported endline ORF scores are included..

Most interestingly, as Table 6 shows, four effect sizes of 0.5 standard deviations or greater - GILO, RARA, SHRP, and SMRS - are associated with students who could not reach reading fluency.⁷ In contrast, EGRA Plus in Liberia had a significant impact of 0.8 standard deviations for ORF and students who read fluently at 49.61 correct words per minute.

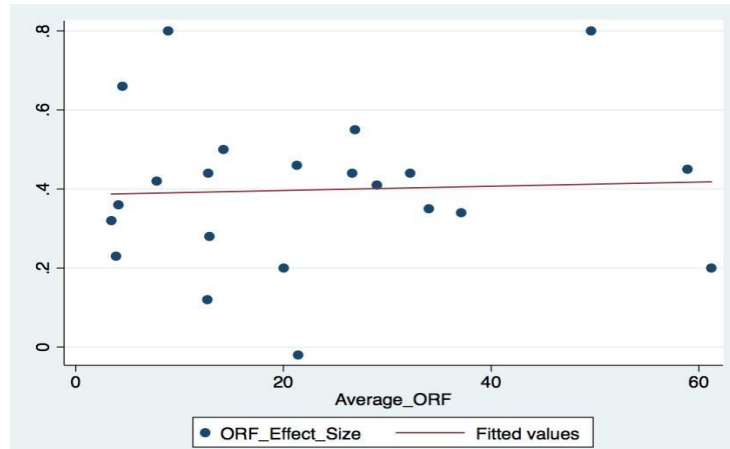
Table 6: Large Effect Sizes Without Fluency - ORF Effect Sizes Compared to the Average ORF Score at Endline for the Treatment Group, for Select Interventions

Program	Country	Effect Size	Raw Endline ORF
GILO	Egypt, Arab Rep.	0.55	26.9
RARA	Nigeria	0.66	4.5
SHRP	Uganda (Runyoro Rutooro)	0.50	14.23
SMRS	South Africa	0.80	8.9

These findings highlight the fundamental fact that students are significantly improving their reading as a consequence of EGR interventions, but far too often they still cannot attain reading fluency. Figure 10 also illustrates that there is not a relationship between ORF effect size and the raw ORF endline score. Essentially, an EGR intervention can produce a statistically significant effect that demonstrates a causal impact, but the practical magnitude of the change does not mean on average students are reading fluently. Only on limited occasions have EGR interventions elevated students on average to reading fluency. This fact demands that evaluators and policy makers consider not only effect sizes, but also how the raw ORF scores compare to benchmarks of reading fluency. It is worth noting, however, that only one of the evaluations studied (that from Uganda) tested the effects of at least two years of intervention; most occurred over a one-year time span. It is possible that the implementation of EGR interventions throughout all of primary school would lead to much higher rates of fluent readers by the end of 3rd or 4th grade.

Figure 10: OLS Unweighted Regression Between the Treatment's Average ORF Score at Endline and the Effect Size on ORF Scores

⁷ All of the students were 2nd graders except those for SMRS which were 1st graders.



There correlation between the two variables is not significant at $p < .1$
 Includes all estimates for which both data points were available

6. Cost-Effectiveness

Slightly less than half of the evaluations detailed cost-effectiveness data. Scholars have pointed out that insufficient data makes it difficult to compare the cost-effectiveness of various interventions (Evans and Popova 2015, 17). However, one study did show that teacher training interventions were among the cheaper and more cost-effective approaches to improving education outcomes. At the very least, they were not among the most expensive approaches, such as computer-based learning interventions (McEwan 2015, 377-379).

As Table 7 shows, the costs of EGR interventions vary widely. As discussed in the methodology section, there are a large number of reasons for variability in cost-effectiveness. To reiterate with specific examples from the interventions, scale matters. EGRA Plus in Liberia and PEARL in Tonga were the most expensive and had the fewest students enrolled, but MTPDS in Malawi and PRIMR in Kenya were among the cheapest and had two of the three largest enrollments. However, large enrollments might entail large class sizes which present challenges for the quality of instruction. Even though MTPDS had the cheapest cost per student, the average class size was 113 students per class (Tilson et al 2013, 33). On top of that, Table 7 also shows that different interventions ran for different lengths of time with different components. The components influence the technical implementation costs of EGR interventions. Some EGR interventions have more components than others, and certain components are more expensive in certain contexts. For these reasons, it is difficult to make sweeping claims about the cost-effectiveness. Despite this difficulty, examples from Kenya and Malawi (both MTPDS and MEGRA) illustrate the potential cost-effectiveness of EGR interventions. Likewise, low costs can facilitate scaling interventions to nationwide programs. Such was the case in both Kenya and Jordan.

Table 7: Cost-effectiveness Information

Country	Cost (USD)	Number of students	Cost per student	Program Length (school years)
Egypt, Arab Rep.	\$ 1,027,506	34,930	\$ 29.42	3
Jordan	\$ 437,500	10,986	\$ 39.82	2
Kenya	\$ 1,575,647	206,151	\$ 7.64	3
Liberia	\$ 1,095,428	2,825	\$ 387.76	2
Malawi	\$ 537,712	264,869	\$ 2.03	3
Malawi	\$ 4,190,168	554,796	\$ 7.55	4
Mozambique	\$ 480,997	52,710	\$ 9.13	2
Tonga	\$ 255,436	1,396	\$ 183.00	0.5

Includes all reported technical cost information.

7. Contextual Factors

The main goal of this paper is to determine the general effectiveness of EGR interventions – not to determine when they work best. This is because the nature of the data collected – which has frequent missing entries, an inconsistent number of observations per study, and relatively few observations overall – makes regression analysis comparing contextual factors and outcomes difficult. Nonetheless, a few conclusions about the factors that determine EGR effectiveness can be drawn from the data we have gathered, and they are worth discussing briefly in this section.

First of all, a proof of concept has been established in certain regards through these studies. First, it has become clear that EGR interventions can be effective in schools teaching in both mother tongue languages (e.g. in Egypt and Uganda) and otherwise (e.g. in Mozambique and Kenya). Second, they can be effective not only as pilot studies (as with EGRA Plus) but also when implemented at large scales (e.g. in Kenya and Kyrgyzstan). Furthermore, large-scale EGR interventions in Zambia and Kenya, despite not using control groups for evaluation, have produced promising results, offering evidence that EGR interventions can be effective at a national, or nearly national, scale (Freudenberger and Davis 2017, 7, 13) (Facloner-Stour et al 2017, x-xiii).

The ApaL program in Mozambique also established proof that program impacts can be maintained even after rigorous implementation has ended. However, given that DiD impacts fell substantially (despite still being significant) following the cessation of implementation, it also offers evidence that sustained implementation (i.e. continued trainings and support on how to use the materials that teachers have been provided) is ideal. Likewise, MTPDS in Malawi showed that a higher level of rigor in implementation leads to much greater impacts. Specifically, it showed that the large conference-style trainings were far less effective than the intensive EGR intervention. These findings suggest that, in order for programs to be effective, they should not be considered short-term interventions, but rather programs that should be sustained and taken over by host-country governments. Finally, programs proved to be effective across various durations of implementation and with the inclusion or exclusion of various EGR intervention elements.

Unfortunately, although certain evaluations have proven when EGR interventions can potentially be effective, there is less evidence that shows why they fail. Certainly, it is clear that not only program elements, but also contextual elements matter. The EGRA program in Malawi, for example, was one of the most intensive with a wide variety of intervention components, but it was also one of the least successful. In Uganda, the same intervention was carried out in a variety of different areas with widely ranging results. Nevertheless, it is not clear which contextual factors matter. For Malawi's EGRA program, failure may have been due to external shocks such as increased class sizes that disrupted implementation. In the Democratic Republic of Congo, it may have been due to a lack of buy-in from teachers. In the Philippines, the fact that another primary school intervention had just started implementation may have lessened the impact of or disrupted the EGR intervention. Elsewhere, as in Tanzania, negative results from the evaluation may have been due at least partially to a lack of a sound comparison group.

VI. CONCLUSION

It is imperative that students learn to read in the early grades, yet many fail to do so in developing countries. EGR interventions have emerged as a possible solution to this problem. In order to develop a better understanding of how effective they are as a solution, we have reviewed the evidence from 18 EGR interventions, which occurred across a large variety of contexts. We found that the large majority of these interventions have a significant and/or substantial impact on at least one of three literacy subtasks, each measuring different skill levels. For most interventions, the impacts are highly significant, and in many cases the effect sizes are equivalent to over half a year's worth of schooling. At times the impacts are remarkable, equating to over a year's worth of schooling, or resulting in massive improvements in scores relative even to high benchmarks of achievement. Interventions also proved to be effective in a wide variety of contexts. The cost of implementation varied greatly, but it is clear that EGR interventions have the potential to be implemented cost-effectively.

Despite consistently substantial effect sizes, impacts in most instances could still be considered moderate by some standards. Specifically, large effect sizes often equate with relatively minor

DiD impacts that do not bring students close to fluency. Regardless of grade level, most students at endline were well below the range of correct words per minute required for fluency. Furthermore, a minority of EGR interventions failed to produce significant or substantial results, and it is not clear why these interventions failed while others succeeded.

These findings make it clear that, while EGR interventions are not a guaranteed means to improve reading and they rarely lead to fluency over a short span of time, they are a mostly reliable means to make significant improvements in literacy, and accelerate learning in contexts where very little learning is taking place. That being said, they seem to be only a partial solution, as they do not typically create fluent readers (at least not over one-year periods). So, improvements to other aspects of education systems – including teacher incentives, school infrastructure, community involvement, early childhood development, and more – are certainly necessary to achieve universal literacy. EGR interventions could be a foundation for additional progress, but they are not a panacea.

To better understand how EGR interventions should be implemented to maximize impact, and to what degree that can be relied upon to improve literacy, there are several avenues for further research that could be pursued. First, more studies that test the long-term effects of EGR interventions, such as that from Uganda, could be carried out in order to determine how large impacts can be if implementation is sustained throughout all of early primary school. Second, more rigorous evaluations of large-scale interventions could be conducted in order to determine if EGR interventions are as reliable when implemented at a nearly national level. Third, more studies should include detailed cost-effectiveness information in order to better illuminate how much the interventions cost on average, and why costs vary so drastically. Finally, evaluations could be carried out with multiple treatment arms that vary not just in intensity (as they have in several studies), but in terms of which EGR components they contain, in order to provide a deeper understanding of which components are most effective.

References

- Abadzi, H. (2017). "Turning a molehill into a mountain? How reading curricula are failing the poor worldwide." *Prospects*, 1-14.
- Abadzi, H. (2013). "Literacy for all in 100 days? A research-based strategy for fast progress in low income countries" (Global Partnership for Education Working Paper Series 7). Washington, DC: Global Partnership for Education.
- Abadzi, H. (2012). *Efficient learning for the poor: Insights from the frontier of cognitive neuroscience*. Washington, DC: World Bank.
- Allan, E.J. and N. Horn. (2013). *Evaluation of the USAID funded textbooks and learning materials programs (TLMP) in Ethiopia, Ghana, Malawi, Senegal, South Africa and Tanzania*. Washington, DC: USAID.
- American Institutes for Research and Save the Children International. (2016). *USAID quality reading project: Midterm impact evaluation report for the Kyrgyz Republic*. Bishkek: USAID/Kyrgyz Republic.
- ASER Centre. (2011). *Annual status of education report 2010*. Mumbai: Pratham Resource Center.
- Brombacher, A., J. Stern, L. Nordstrum, C. Cummiske, and A. Mulcahy Dunn. (2014). *National early grade literacy and numeracy survey - Jordan, intervention impact analysis report*. Amman, JO: USAID/Jordan.
- Bulat, J., A. Brombacher, T. Slade, J. Iriondo-Perez, M. Kelly, and S. Edwards. (2014). *Projet d'amélioration de la qualité de l'éducation (PAQUED): Endline report of early grade reading assessment (EGRA) and early grade math assessment (EGMA)*. Kinshasa: USAID/Democratic Republic of the Congo.
- Clark-Chiarelli, N. and N. Louge. (2016). "Teacher quality as a mediator of student achievement," in *Understanding what works in oral reading assessments*. Montreal: UNESCO, 30-40.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Creative Associates International. (2013). *21st century basic education program: Annual report year 2*. Dar es Salaam: USAID/Tanzania.
- Creative Associates International. (2014). *21st century basic education program: Annual report year 3*. Dar es Salaam: USAID/Tanzania.
- Creative Associates International. (2015). *21st century basic education program: Annual report year 4*. Dar es Salaam: USAID/Tanzania.
- Dubeck, M. and A. Gove. (2015). "The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations." *International Journal of Educational Development* 40, 315-322.
- Easterlin, R. (1981). "Why isn't the whole world developed." *Journal of Economic History* 41 (1), 1-19.
- Education Development Center. (2015). *Basa pilipinas: Transforming reading instruction to boost student learning outcomes*. Manila: USAID/Philippines.

- Education Development Center. (2014). *USAID/Philippines basa pilipinas program: Annual progress report*. Manila: USAID/Philippines.
- Education for All. (2014). *EFA global monitoring report 2013/4*. Montreal: UNESCO.
- Evans, D. and A. Popova. (2015). "What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews" (World Bank Policy Research Working Paper 7203). Washington, DC: World Bank.
- Falconer-Stour, Z., R. Frischkorn, L. M. Franco. (2017). *Time to learn endline evaluation report*. Lusaka: USAID/Zambia.
- Freudenberger, E. and J. Davis. (2017). *Tusome external evaluation - Midline report*. Nairobi: USAID/Kenya.
- Gove, A., A. Brombacher, and M. Ward-Brent. (2017). "Sparking a reading revolution: Results of early literacy interventions in Egypt and Jordan." *New Directions for Childhood and Adolescent Development* 155, 97-115.
- Gove, A. and P. Cvelich (2010). *Early reading: Igniting education for all* (A report by the Early Grade Learning Community of Practice). Research Triangle Park, NC: Research Triangle Institute.
- Gross, J. (2009). *The long term costs of literacy difficulties*. London: Every Child a Chance Trust.
- Hill, C., H. Bloom, A. Black, and M. Lipsey. (2008). "Empirical benchmarks for interpreting effect sizes in research." *Child Development Perspectives* 2 (3), 172-177.
- Hollingsworth, S. and P. Gains. (2009). *The systematic method for reading success (SMRS) in South Africa: A literacy intervention between EGRA pre- and post- assessments*. Pretoria: USAID/Southern Africa.
- Kim, Y. G., H. Boyle, S. Zuilkowski, and P. Nakamura. (2016). *Landscape report on early grade literacy*. Washington, DC: USAID.
- King, S., M. Korda, L. Nordstrum, and S. Edwards. (2015). *Liberia teacher training program: Endline assessment of the impact of early grade reading and mathematics interventions*. Monrovia: USAID/Liberia.
- Macdonald, K., S. Brinkman, W. Jarvie, M. Machuca-Sierra, K. McDonall, S. Messaoud-Galusi, S. Tapueluelu, and B. T. Vu. (2017). *Pedagogy versus school readiness: The impact of a randomized reading instruction intervention and community-based playgroup intervention on early grade reading outcomes in Tonga*. Washington, DC: World Bank.
- Macdonald, K. (2016). *Impact of highly scripted teaching method to improve early grade reading skills: A low cost, randomized intervention in Papua New Guinea* (Draft Working Paper).
- Malik, S., B. Balfour, J. P. Nzabonimpa, S. Cozzolino, G. Dib, and A. J. Dowd. (2015). *Endline evaluation of Rwandan children's book initiative*. Kigali: Save the Children.
- Marinelli, C. V., M. Martelli, P. Praphamontripong, P. Zoccolotti, and H. Abadzi. (2011). "Visual and linguistic factors in literacy acquisition: Instructional implications for beginning readers in low income countries" (GPE Working Paper Series on Learning no 2). Washington, DC: Global Partnership for Education.

- McEwan, P. (2015). "Improving learning in primary schools of developing countries: A meta analysis of randomized experiments." *Review of Educational Research* 85 (3), 353-394.
- Montenegro, C. and H. Patrinos. (2014). "Comparable estimates of returns to schooling around the world" (World Bank Policy Research Working Paper 7020). Washington, DC: World Bank.
- Nagarajan, G., P. Carneiro, M. Chiappetta, B. Fuller, E. Gonzales, B. Sinclair, and A. Mapondera. (2015). *Midline report: Impact evaluation of the early grade reading activity*. Lilongwe: USAID/Malawi.
- NORC at the University of Chicago. (2017). *Performance and impact evaluation of the USAID/Uganda school health and reading program*. Kampala: USAID/Uganda.
- Patrinos, H., E. Velez, and C. Y. Wang. (2013). "Framework for the reform of education systems and planning for quality" (World Bank Policy Research Working Paper 6701). Washington, DC: World Bank.
- Piper, B. (2009). *Impact study of SMRS using early grade reading assessment in three provinces in South Africa*. Pretoria: USAID/Southern Africa.
- Pritchett, L. and A. Beatty. (2015). "Slow down, you're going too fast: Matching curricula to student skill levels." *International Journal of Educational Development* 40: 276-288.
- Pryor, J., K. Akyeampong, J. Westbrook, and K. Lussier. (2012). "Rethinking teacher preparation and professional development in Africa: An analysis of the curriculum of teacher education in the teaching of early reading and mathematics," *The Curriculum Journal* 23 (4), 409-502.
- Raupp, M., B. Newman, L. Revés, C. Lauchande, and E. J. Allan. (2016). *Impact evaluation of the USAID/Aprender a Ler project in Mozambique*. Maputo: USAID/Mozambique.
- RTI International. (2015). *Nigeria reading and access research activity (RARA): Results of an approach to improve early grade reading in Hausa in Bauchi and Sokoto states*. Abuja: USAID/Nigeria.
- RTI International. (2014a). *Costing early grade reading programs: An examination of various costs and issues around costing*. Washington, DC: USAID.
- RTI International. (2014b). *Girls' improved learning outcomes (GILO): Final report*. Cairo: USAID/Egypt.
- RTI International. (2014c). *USAID/Kenya primary math and reading (PRIMR) initiative: Final report*. Nairobi: USAID/Kenya.
- RTI International. (2011a). *EGRA plus: Liberia program evaluation report*. Monrovia: USAID/Liberia.
- RTI International. (2011b). *NALAP formative evaluation report, Ghana*. Washington, DC: USAID.
- RTI International. (2011c). *PAQED: DRC Baseline report, early grade reading assessment*. Kinshasa: USAID/Democratic Republic of the Congo.
- School-to-School International. (2015). *TZ21 endline evaluation report: Evaluation conducted in Zanzibar, Mtwara, and Lindi*. Dar es Salaam: USAID/Tanzania.
- Snilstveit, B., J. Stevenson, D. Phillips, M. Vojtkova, E. Gallagher, T. Schmidt, H. Jobse, M. Geelen, M. G. Pastorello. (2015). *Interventions for improving learning outcomes and access to education in low- and middle-income countries: A systematic review* (Systematic Review 24). Washington, DC: International Initiative for Impact Evaluation.

- Social Impact. (2016). *Basa pilipinas impact evaluation: Midiline report*. Manila:USAID/Philippines.
- Spratt, J., S. King, and J. Bulat. (2013). *Independent evaluation of the effectiveness of Institut pour l'Education Populaire's "Read-Learn-Lead" (RLL) program in Mali*. Research Triangle Park, NC: RTI International.
- Tilson, T., A. Kamlongera, M. Pucilowski, and D. Nampota. (2013). *Evaluation of the Malawi teacher professional development support (MTPDS) program*. Lilongwe: USAID/Malawi.
- U.S. Department of Education. (2014). *What works clearinghouse: procedures and standards handbook version 3.0*. Washington, DC: Institute of Education Sciences.
- Williams, B. and S. de Silva. (2015a). *Pilot reading booster program in Madang province, Papua New Guinea: Evaluation report*. Port Moresby: Papua New Guinea Department of Education.
- Williams, B. and S. de Silva. (2015b). *Pilot reading booster program in Western Highlands province, Papua New Guinea: Evaluation report*. Port Moresby: Papua New Guinea Department of Education.
- World Bank. (2017). *World development report 2018: Learning to realize education's promise*. Washington, DC: World Bank.
- World Bank. (2016a). *Implementation completion report: Reading education (READ-PNG) project*. Washington, DC: World Bank.
- World Bank. (2016b). *World development report 2016: Digital dividends*. Washington, DC: World Bank.

Appendices
Appendix 1: intervention impacts

Country	Program Name	Grade	Language(s) of Assessment	ORF, DiD	ORF, effect size	ORF, sig	% zero word, DiD	LSR, DiD	LSR, effect size	LSR, sig	LNR, DiD	LNR, effect size	LNR, sig	RC DiD (% correct)	RC, effect size	RC, sig	Positive ?
Congo, Dem. Rep. (Bandundu)	PAQUED	2	French	0.06*	0.01*	no	0
Congo, Dem.. Rep. (Equateur)	PAQUED	2	French	0.05*	0.01*	no	0
Congo, Dem. Rep. (Orientale)	PAQUED	2	French	0.05*	0.03*	no	0
Congo, Dem. Rep. (Bandundu)	PAQUED	4	French	4.26	0.28	.	-18	0.14*	0.03*	4.8	.	.	1
Congo, Dem. Rep. (Equateur)	PAQUED	4	French	5.58	0.12	.	5	0.8*	0.12*	3.4	.	.	0
Congo, Dem. Rep. (Orientale)	PAQUED	4	French	-0.64	-0.02	.	-2	-0.44*	-0.12*	0.8	.	.	0
Egypt, Arab Rep.	GILO	2	Arabic	12	0.55	***	.	19	1.07	***	1
Jordan	National Early Grade Literacy and Numeracy Survey	2	Arabic	6.3	0.46	***	1.9	11.6	0.48	***	1
Kenya	PRIMR	1	English	.	0.44	.	.	.	0.68	0.38	.	1
Kenya	PRIMR	2	English	.	0.45	yes	.	.	0.78	yes	0.44	yes	1
Kenya	PRIMR	1	Kishwahili	.	0.41	.	.	.	0.57	0.45	.	1
Kenya	PRIMR	2	Kishwahili	.	0.35	yes	.	.	0.7	yes	0.32	yes	1
Kyrgyz Republic	USAID Reading Quality Project	2	Russian and Kyrgyz	.	0.34	***	0.23	*	.	0.29	no	1

Country	Program Name	Grade	Language(s) of Assessment	ORF, DiD	ORF, effect size	ORF, sig	% zero word, DiD	LSR, DiD	LSR, effect size	LSR, sig	LNR, DiD	LNR, effect size	LNR, sig	RC DiD (% correct)	RC, effect size	RC, sig	Positive ?
Kyrgyz Republic	USAID Reading Quality Project	4	Russian and Kyrgyz	.	0.2	*	0.09	no	1
Liberia	EGRA Plus	2 and 3	English	21.1	0.8	***	14.8	0.52	***	25.2	0.82	***	1
Liberia	LTTP II	1	English	7.2	9.3	.	.	7	.	.	.
Liberia	LTTP II	2	English	6.7	24.4	.	.	8	.	.	.
Liberia	LTTP II	3	English	7.2	-0.4	.	.	8	.	.	.
Malawi	MTPDS	2	Chichewa	8.88	.	***	-43.7	.	.	.	20.14	.	***	11	.	***	1
Malawi	EGRA	2	Chichewa	1.01	.	no	-10.8	.	.	.	5.35	.	.	-0.01	.	.	0
Mali	RLL	1	Bamanankan, Bomu, Fulfulde, and Songhai	0.97	0.23	*	.	4.61	0.59	***	.	.	.	0.6	0.05	no	1
Mali	RLL	2	Bamanankan, Bomu, Fulfulde, and Songhai	3.64	0.42	***	.	9	0.66	***	.	.	.	8.9	0.42	***	1
Mozambique (endline)	ApaL	2	Portuguese	2.5*	.	***	11.8*	.	***	2.2*	.	***	1
Mozambique (endline)	ApaL	3	Portuguese	6.8*	.	***	15.5*	.	***	7*	.	***	1
Mozambique (post-endline)	ApaL	2	Portuguese	.4*	.	***	1.3*	.	***	2*	.	***	1
Mozambique (post-endline)	ApaL	3	Portuguese	2.4*	.	***	6.6*	.	***	1.5*	.	***	1
Nigeria	RARA	2	Hausa	4.9	0.66	**	.	6.1	0.78	***	.	.	.	0.1	0.6	**	1
PNG (MP and WHP)	Reader Booster	3	English	.	0.4*	*	0.26	no	1
PNG (MP only)	Reader Booster	4	English	.	0.2*	no	-0.03	no	0

Country	Program Name	Grade	Language(s) of Assessment	ORF, DiD	ORF, effect size	ORF, sig	% zero word, DiD	LSR, DiD	LSR, effect size	LSR, sig	LNR, DiD	LNR, effect size	LNR, sig	RC DiD (% correct)	RC, effect size	RC, sig	Positive ?
Philippines	Basa Pilipinas	2	Filipino	9.1	.	***	5.51	.	***	23.5	.	***	1
Philippines	Basa Pilipinas	1	Mother tongue	4.31	.	no	2	5.1	.	*	.	.	.	0.03	.	no	0
Philippines	Basa Pilipinas	2	Filipino	-0.03	.	no	-3	0.7	.	no	.	.	.	0.06	.	no	0
Philippines	Basa Pilipinas	2	English	-0.18	.	no	0	-0.4	.	no	.	.	.	0.11	.	*	0
Philippines	Basa Pilipinas	3	Filipino	-0.06	.	no	0	-2.3	.	no	.	.	.	0.02	.	no	0
Philippines	Basa Pilipinas	3	English	-0.18	.	no	2	0.5	.	no	.	.	.	0.02	.	no	0
South Africa	SMRS	1	Sepedi, Zulu, and Setswana	7.21	0.8	***	1.13	14.32	.	***	.	.	.	9	0.59	***	1
Tanzania	TZ21	2	Kiswahili	0.07	.	no	-12.4	.	.	no	.	.	no	-1.5	.	no	0
Tonga	PEARL	2	Tongan	.	.24 - .33	*** (all)	.	.	.59 - .75	*** (all)	.	.02, .05, .07, .15	no, no, no, *	.	.15, .14, .2, .3	no, *, **, ***	1
Uganda	SHRP	1	Luganda	4.5	0.2	*	.	7	0.53	***44'	0.24	*	1
Uganda	SHRP	1	Leblango	0.44	.	no	.	6.27	0.5	***	.	.	.	0.09'	.	no	1
Uganda	SHRP	1	Ateso	6.25	0.44	***	.	2.11	.	no46'	0.45	***	1
Uganda	SHRP	1	Ruyankor e-Rukiga	6.33	0.44	**	.	1.67	.	no59'	0.59	**	1
Uganda	SHRP	1	English	5.04	0.28	*	.	4.23	0.45	***18'	.	no	1
Uganda	SHRP	2	Runyoro-Rutooro	7.25	0.5	***	0.21	2.61	.	no	.	.	.	0.77'	0.53	***	1
Uganda	SHRP	2	Acoli	2.44	0.32	*	0.11	2.63	.	no	.	.	.	0.11'	0.19	*	1
Uganda	SHRP	2	Lugbarati	0.7	.	no	0.02	-1.73	.	no	.	.	.	0.07'	.	no	0
Uganda	SHRP	2	Lumasaaba	2.44	0.36	**	0.12	5.61	0.65	***	.	.	.	0.15'	.	no	1
Uganda	SHRP	2	English	3.37	0.28	***	.	2.37	0.26	**	.	.	.	0.08'	0.25	**	1
Uganda	SHRP	3	Lugwere	0.04	.	no	.	2	0.48	**	.	.	.	-0.01'	.	no	1
Uganda	SHRP	3	Ngakarimjong	2.63	.	*	.	7.63	0.76	***	.	.	.	0.3'	0.33	*	1

Uganda	SHRP	3	Lhukonzo	0.73	.	no	.	0.02	.	no	.	.	.	0.09'	.	no	0
Uganda	SHRP	3	Lusoga	0.39	.	**	.	4.54	0.76	***	.	.	.	0.04'	0.18	*	1
Uganda	SHRP	3	English	4.03	0.6	***	1

Notes:

For the 'Grade' column, longitudinal studies show the grade at baseline. For 'ORF, DiD', 'LNR, DiD', and 'RC DiD', figures with * represent the endline difference not the DiD

For 'ORF, effect size', * represents that the effect size is for familiar word, not ORF for 'LSR, DiD' and 'LSR, effect size', figures with * represent ISI rather than LSR

For 'RC DiD', figures with ' represent the average number correct rather than the average % correct

Sig=significance. For significance columns ***=p<.01, **=p<.05, *=p<.1, yes=p<.1 (at least, it might be more significant), no=p>.1

For all columns, '.'=no data available. For the 'Positive?' column, "1"= yes and "0"=no.

For Tonga, data are presented in one row for all for regression methods - either as a range or, if they vary in significance, as a series

Appendix 2: DiD for ORF

Country	Program Name	Language(s) of Assessment	Grade	Endline later?	ORF, control , before	ORF, control , after	ORF, control , diff	ORF, treatment , before	ORF, treatment , after	ORF, treatment , diff	ORF, DiD
Congo, Dem. Rep. (Bandundu)	PAQUED	French	4	0	3.86	4.28	0.42	8.2	12.88	4.68	4.26
Congo, Dem. Rep. (Equateur)	PAQUED	French	4	0	15.72	12.73	-2.99	10.09	12.68	2.59	5.58
Congo, Dem. Rep. (Orientale)	PAQUED	French	4	0	13.25	18.42	5.17	16.89	21.42	4.53	-0.64
Egypt, Arab Rep.	GILO	Arabic	2	1	9	12.1	3.1	11.8	26.9	15.1	12
Jordan	National Early Grade Literacy and Numeracy Survey	Arabic	2	1	14.7	15.3	0.6	14.4	21.3	6.9	6.3
Liberia	EGRA Plus	English	2 and 3	1	18.14	25.21	7.07	20.83	49.61	28.78	21.71
Liberia	LTTP II	English	1	0	9.5	6.3	-3.2	2.1	6.1	4	7.2
Liberia	LTTP II	English	2	0	6.4	9.1	2.7	4.8	14.2	9.4	6.7
Liberia	LTTP II	English	3	0	18.9	24.3	5.4	7.6	20.2	12.6	7.2
Malawi	MTPDS	Chichewa	2	0	0.38	0.24	-0.14	0.21	8.95	8.74	8.88

Malawi	EGRA	Chichewa	2	1	3.22	1.84	-1.38	3.89	3.52	-0.37	1.01
Mali	RLL	Bamanankan, Bomu, Fulfulde, and Songhai	1	0	0.48	3	2.52	0.39	3.88	3.49	0.97
Mali	RLL	Bamanankan, Bomu, Fulfulde, and Songhai	2	0	1.91	3.96	2.05	2.12	7.8	5.68	3.63
Nigeria	RARA	Hausa	2	1	1.5	0.8	-0.7	0.2	4.5	4	4.7
Philippines	Basa Pilipinas	Mother tongue	1	1	14.43	21.18	6.75	12.28	23.34	11.06	4.31
Philippines	Basa Pilipinas	Filipino	2	1	37.35	45.73	8.38	34.78	43.13	8.35	-0.03
Philippines	Basa Pilipinas	English	2	1	36.01	42.46	6.45	33.69	39.96	6.27	-0.18
Philippines	Basa Pilipinas	Filipino	3	1	47.85	51.1	3.25	48.93	52.12	3.19	-0.06
Philippines	Basa Pilipinas	English	3	1	55.04	61.06	6.02	54.8	60.64	5.84	-0.18
Tanzania	TZ21	Kiswahili	2	1	10.93	15.15	4.22	6.49	10.78	4.29	0.07
Uganda	SHRP	Luganda	1	1	0.14	15.82	15.68	0.17	20.03	19.86	4.18
Uganda	SHRP	Leblango	1	1	0.03	14.22	14.19	0.04	14.86	14.82	0.63
Uganda	SHRP	Ateso	1	1	0.05	6.18	6.13	0	12.76	12.76	6.63
Uganda	SHRP	Ruyankor e-Rukiga	1	1	0.3	20.26	19.96	0.53	26.63	26.1	6.14
Uganda	SHRP	Runyoro-Rutooro	2	1	0.07	7.05	6.98	0.04	14.23	14.19	7.21
Uganda	SHRP	Acoli	2	1	0.02	0.96	0.94	0.01	3.43	3.42	2.48
Uganda	SHRP	Lugbarati	2	1	0.01	6.09	6.08	0	6.56	6.56	0.48
Uganda	SHRP	Lumasaaba	2	1	0	1.72	1.72	0	4.11	4.11	2.39
Uganda	SHRP	Lugwere	3	1	0.01	0.39	0.38	0.01	0.44	0.43	0.05
Uganda	SHRP	Ngakarimojong	3	1	0	1.15	1.15	0	3.39	3.39	2.24
Uganda	SHRP	Lhukonzo	3	1	0.02	1.26	1.24	0	1.82	1.82	0.58
Uganda	SHRP	Lusoga	3	1	0	0.05	0.05	0	0.45	0.45	0.4

Notes:

For the 'Grade' column, longitudinal studies show the grade at baseline.

Some DiD figures in this table are slightly different than those in Appendix 1 - this is because some of the DiD impacts in Appendix 1 were calculated using regressions, which affected the DiD in various ways, for example by including controls. The DiD figures here are simple differences in differences. Observations without detailed DiD data were omitted.

"Diff" = difference. The 'Endline later' column describes whether the endline was conducted later in the school year than the baseline. '1' indicates that it was, '0' that it was not.

Appendix 3: endline scores

Country	Program name	Grade	Language(s) of Assessment	ORF, treatment, after	% zero word, treatment endline	RC, treatment endline
Congo, Dem. Rep. (Bandundu)	PAQUED	4	French	12.88	40	6.8
Congo, Dem. Rep. (Equateur)	PAQUED	4	French	12.68	52	1.6
Congo, Dem. Rep. (Orientale)	PAQUED	4	French	21.42	26	13.6
Egypt, Arab Rep.	GILO	2	Arabic	26.9	20.7	.
Jordan	National Early Grade Literacy and Numeracy Survey	2	Arabic	21.3	4	45
Kenya	PRIMR	1	English	32.2	.	21.1
Kenya	PRIMR	2	English	58.9	.	48.4
Kenya	PRIMR	1	Kishwahili	29	.	25.6
Kenya	PRIMR	2	Kishwahili	34	.	46.6
Kyrgyz Republic	USAID Reading Quality Project	2	Russian and Kyrgyz	37.12	.	.
Kyrgyz Republic	USAID Reading Quality Project	4	Russian and Kyrgyz	61.2	.	.
Liberia	EGRA Plus	2 and 3	English	49.61	5.7	59.38
Liberia	LTTP II	1	English	6.1	.	10
Liberia	LTTP II	2	English	14.2	.	11
Liberia	LTTP II	3	English	20.2	.	24
Malawi	MTPDS	2	Chichewa	8.95	48.4	11.4
Malawi	EGRA	2	Chichewa	3.52	73.6	0.012
Mali	RLL	1	Bamanankan, Bomu, Fulfulde, and Songhai	3.88	.	2.6
Mali	RLL	2	Bamanankan, Bomu, Fulfulde, and Songhai	7.8	.	12.4

Mozambique (endline)	ApaL	2	Portuguese	4.2	.	3
Mozambique (endline)	ApaL	3	Portuguese	12	.	10.8
Mozambique (post-endline)	ApaL	2	Portuguese	2.4	.	5.5
Mozambique (post-endline)	ApaL	3	Portuguese	7.4	.	9
Nigeria	RARA	2	Hausa	4.5	.	0.1
Philippines	Basa Pilipinas	1	Mother tongue	23.34	16	28
Philippines	Basa Pilipinas	2	Filipino	43.13	18	31
Philippines	Basa Pilipinas	2	English	39.96	9	21
Philippines	Basa Pilipinas	3	Filipino	52.12	4	18
Philippines	Basa Pilipinas	3	English	60.64	5	31
South Africa	SMRS	1	Sepedi, Zulu, and Setswana	8.9	23	11
Tanzania	TZ21	2	Kiswahili	10.78	55.5	14.8
Uganda	SHRP	1	Luganda	20.03	.	.
Uganda	SHRP	1	Leblango	14.86	.	.
Uganda	SHRP	1	Ateso	12.76	.	.
Uganda	SHRP	1	Ruyankor e-Rukiga	26.63	.	.
Uganda	SHRP	2	Runyoro-Rutooro	14.23	37	.
Uganda	SHRP	2	Acoli	3.43	82	.
Uganda	SHRP	2	Lugbarati	6.56	66	.
Uganda	SHRP	2	Lumasaaba	4.11	71	.
Uganda	SHRP	3	Lugwere	0.44	.	.
Uganda	SHRP	3	Ngakarimojong	3.39	.	.
Uganda	SHRP	3	Lhukonzo	1.82	.	.
Uganda	SHRP	3	Lusoga	0.45	.	.

Notes:

For the 'Grade' column, longitudinal studies show the grade at baseline.

For all columns, '.'=no data available. Observations with no data for all three measures were omitted.

Appendix 4: contextual information

Country	Program name	Program start year	Length of program, school years	Evaluation start year	Evaluation length, school years	Number of schools in the program	Implementers and evaluators	Donor	Additional components
Congo, Dem. Rep.	PAQUED	2012	1	2012	1	618	EDC, RTI	USAID	Weekly lessons for students from external educators on reading, math, and life skills.
Egypt, Arab Rep.	GILO	2008	3	2009	3	166	RTI	USAID	Expanding access for girls, improving school management, and strengthening decentralized institutions.
Jordan	National Early Grade Literacy and Numeracy Survey	2013	1	2013	1	43	RTI	USAID	.
Kenya	PRIMR	2012	3	2012	2	310	RTI	USAID	.
Kyrgyz Republic	USAID Reading Quality Project	2013	4	2014	1	616	StC, AIR	USAID	Facilitated policy reforms such as the creation of teacher training standards and national standards for assessments.
Liberia	EGRA Plus	2008	2	2008	2	60	RTI	USAID	Provision of information on student reading performance to parents.
Liberia	LTTTP II	2011	4	2011	2	792	RTI, FHI 360	USAID	Provision of assessments that became part of school report cards that were presented to parents
Malawi	MTPDS	2010	3	2010	2	.	Social Impact	USAID	.
Malawi	EGRA	2013	4	2013	2	1603	RTI, Social Impact	USAID	Rewards for high-performing teachers, reading fairs, invitations for parents to participate more in schools, trainings to foster community support for schools, an extension of the school day by one hour, and reduced class sizes.
Mali	RLL	2009	3	2009	3	51	Institut pour l'Éducation Populaire, RTI	Hewlett Foundation	.
Mozambique	ApaL	2013	1.5	2013	2	60	World Education	USAID	School management improvement.
Nigeria	RARA	2014	1	2014	1	60	RTI	USAID	Teacher cluster meetings twice per year and activities with parents and communities to encourage reading at home.
PNG (WHP)	Reader Booster	2014	1	2013	2	23	World Bank	GPE	Remote text message support to teachers.

PNG (MP)	Reader Booster	2013	0.5	2011	3	14	World Bank	GPE	Remote text message support to teachers.
Philippines	Basa Pilipinas	2013	4	2013/2014	1	.	EDC, Social Impact	USAID	The creation of 'national literacy day' and training for administrators to support literacy.
Philippines	Basa Pilipinas	2013	4	2015	1	.	EDC, Social Impact	USAID	The creation of 'national literacy day' and training for administrators to support literacy.
South Africa	SMRS	2009	0.5	2009	0.5	29	RTI	USAID	.
Tanzania	TZ21	2012	.	2013	1	~902	CAI, School-to-School, WB	USAID	Events designed to increase community engagement in education.
Tonga	PEARL	.	.	2015	1	.	World Bank	GPE	.
Uganda	SHRP	2013	4	2013	4	166	RTI, NORC	USAID	Leadership training for head teachers and district officials

Notes:

School year durations are approximations. The column 'Number of schools in the program' shows the total number of schools receiving the intervention at the time of the endline evaluation.

For all columns, '.'=no data available. Observations with no data for all three measures were omitted. For the 'Positive?' column, "1"= yes and "0"=no.

Appendix 5: methodological information

Country	Program name	# of treatment schools at endline	# of control schools at endline	RCT ?	DiD ?	Notes (for those not using an RCT and DiD methodology)	Longitudinal at the cohort level?	Longitudinal at the student level?
Congo, Dem. Rep.	PAQUED	36	36	0	1	Treatment selected randomly from eligible schools; Control selected randomly from schools deemed to be "comparable" to treatment schools	1	0
Egypt, Arab Rep.	GILO	28	30	1	1	.	0	0
Jordan	National Early Grade Literacy and Numeracy Survey	41	110	1	1	.	1	0
Kenya	PRIMR	310	101	1	1	.	0	0
Kyrgyz Republic	USAID Reading Quality Project	30	30	1	1	.	1	0
Liberia	EGRA Plus	59	57	1	1	.	0	0
Liberia	LTTP II	~60	~30	1	1	.	0	0

Malawi	MTPDS	8	16	0	1	The random selection of treatment and control schools did not take place in the same counties	0	0
Malawi	EGRA	59	21	1	1	.	0	0
Mali	RLL	38	37	1	1	.	0	0
Mozambique	ApaL	60	60	1	0	Because a baseline was conducted which showed no statistical difference between schools in control and treatment, results were not compared in terms of DiD but rather in terms of mean differences at endline.	1	0
Nigeria	RARA	60	60	1	1	.	1	0
PNG (WHP)	Reader Booster	10	10	1	1	.	0	0
PNG (MP)	Reader Booster	7	6	1	1	.	0	0
Philippines	Basa Pilipinas	80	40	.	1	Information on school sampling for the evaluation is not available.	1	0
Philippines	Basa Pilipinas	122	122	0	1	Control schools were selected using propensity score matching	1	1
South Africa	SMRS	29	15	1	1	.	1	0
Tanzania	TZ21	80	40	0	1	Control schools were selected using propensity score matching	1	0
Tonga	PEARL	37	36	1	0	No baseline assessment was conducted. Impacts were thus calculated using class random effects, class random effects with covariates, and class random effects with weighting propensity score, in addition to OLS.	1	0
Uganda	SHRP	14	12	1	1	.	1	0

Notes:

For schools that used both an RCT and DiD methodology, schools were randomly assigned treatment across a single sample frame, and scores were measured and compared at baseline and endline.

School year durations are approximations. For the 'RCT?' column, '.'=info not available. For the 'Notes' column, '.'=not applicable.

For the 'RCT?' column, '1'=an RCT methodology was used with random assignment of both treatment and control status across a single population, '0'=this degree of rigor was not used

For the 'DiD?' column, '1'=impacts were calculated with a DiD methodology, '0'= they were not.

For the 'Notes' column, the methodology is explained in further detail if it was not both an RCT and DiD