USAID | KYRGYZ REPUBLIC
FROM THE AMERICAN PEOPLE

USAID Quality Reading Project:

Midterm Impact Report for the Kyrgyz Republic

March 2016

A partnership with
American Institutes for Research and Save the Children International

Contract No.: AID-176-C-13-00001

USAID Quality Reading Project

Midterm Impact Report for the Kyrgyz Republic

Submitted by:
American Institutes for Research

March 2016

This midline impact report of the USAID Quality Reading Project is made possible by the support of the American people through the United States Agency for International Development (USAID). The contents are the sole responsibility of American Institutes for Research and Save the Children International, and do not necessarily reflect the views of USAID or the United States Government.

CONTENTS

Foreword v
Acknowledgments vi
Executive summary 1
1. Introduction 2
2. Intervention 2
2.1 Teacher training 2
2.2 Mentoring 3
2.3 Reading materials 3
2.4 Out-of-school activities 3
2.5 Policy level 4
3. Theory of change 4
4. Literature review 6
5. Research design 8
5.1 Benefits of randomization 8
5.2 Statistical model to estimate impacts 9
5.3 Main outcome: Early grade reading assessment 10
5.4 Sampling 14
6. Descriptive statistics 14
7. Results 17
7.1 Impact on reading outcomes 17
7.2 Subgroup analysis 21
7.3 Longitudinal subsample results 25
7.4 Impact on intermediate outcomes 27
7.5 Language of instruction and home language: associations with reading outcomes 29
8. Limitations 30
9. Conclusion 31
Works cited 33
Appendix A: Balance tables 1
Appendix B: EGRA Subtask correlations 1
Appendix C: EGRA Factor analysis 1
Appendix D: Reading outcome regression tables 1
Appendix E: Other outcome regression tables 1

LIST OF TABLES

Table 5.1. EGRA in the Kyrgyz Republic..... 11
Table 5.2. Correlation Matrix of Grade 2 Russian EGRA Subtasks 13
Table 6.1. Reading Outcome Summary Statistics by Treatment Arm 15
Table 6.2. Background Characteristic Summary Statistics..... 16
Table 7.1. Decoding Effect Sizes, by Grade..... 18
Table 7.2. Language Comprehension Effect Sizes, by Grade..... 19
Table 7.3. Effects by Student Gender, Grade 2 22
Table 7.4. Effects by Student Gender, Grade 4 22
Table 7.5. Effects by Language of Instruction at School, Grade 2 23
Table 7.6. Effects by Language of Instruction at School, Grade 4 23
Table 7.7. Effects by School Location, Grade 2 24
Table 7.8. Effects by School Location, Grade 4 24
Table 7.9. Effects on Grade 2 Longitudinal Sample 25
Table 7.10. Baseline Balance, Longitudinal Subsample Including Students Who Dropped Out 26
Table 7.11. Baseline Balance, Longitudinal Subsample Excluding Students Who Dropped Out 26
Table A.1. Grade 2, Full 130-School, Baseline Balance..... 1
Table A.2. Grade 4, Full 130-School, Baseline Balance..... 1
Table A.3. Grade 2, 60-School Midterm Subsample, Baseline Balance 1
Table A.4. Grade 4, 60-School Midterm Subsample, Baseline Balance 2
Table B.1. Correlation Matrix of Kyrgyz Grade 2 EGRA Subtasks 1
Table B.2. Correlation Matrix of Russian Grade 4 EGRA Subtasks 2
Table B.3. Correlation Matrix of Kyrgyz Grade 4 EGRA Subtasks 2
Table C.1. Kyrgyz Grade 2 Factor Analysis 3
Table C.2. Russian Grade 2 Factor Analysis 3
Table C.3. Kyrgyz Grade 4 Factor Analysis 4
Table C.4. Russian Grade 4 Factor Analysis 4
Table D.1. Reading Outcome Subskills, Grade 2..... 1
Table D.2. Reading Outcome Subskills, Grade 4..... 1
Table E.1. Impact on teaching practices 1
Table E.2. Impact on ParentAL Attitudes to Reading 1
Table E.3. Impact on Student Reading at Home 2
Table E.4. Impact on Student Participation in Reading Activities 2

LIST OF FIGURES

Figure 3.1. USAID/Quality Reading Project Theory of Change 5
Figure 5.1. Difference-in-difference model 10
Figure 6.1. Percent of Students by Household Size 16
Figure 6.2. Percent of Students by Mother's Education 16
Figure 7.1. Impact on Decoding Subtasks, by Grade 19
Figure 7.2. Impact on Language comprehension Subtasks, by Grade 20
Figure 7.3. Impact on Dictation, by Grade 20
Figure 7.4. Impact on Reading Comprehension, by Grade 21

FOREWORD

The main goal of the United States Agency for International Development (USAID) Quality Reading Project (QRP) is to improve reading skills for primary grade students. This Midterm Impact Report uses data from the midterm early grade reading assessment (EGRA) to provide an analysis of the impact the program has had on students' reading skills. This report is written from a research perspective; it states clearly what has the statistical significance to be called an impact. The report is based on 1 year of in-service teacher training for Cohort 1. The second half of the Midterm Impact Report will use Cohorts 2 and 3 EGRA results, which will be collected in April/May 2016.

The Midterm Impact Report shows that the USAID Quality Reading Project *is* having an impact. "We find that beneficiary students know seven more words per minute in Grade 2 than control students, and five more words per minute in Grade 4 than control students. The impacts on reading fluency for familiar non-connected words are larger: 9 words for Grade 2, and 12 words for Grade 4."[1] It is exciting to see that the program has a noteworthy effect on students who learn in a language of instruction different from their mother tongue.

Even though it is exciting to see where the program has had an impact, there are areas of no significant change. For example, in Grade 4, in general there was no significant impact in the single year of program implementation. This report raises questions: Does the program need different activities for the higher primary grades? Are factors beyond the program's scope such as curriculum affecting Grade 4? Are enough hours provided for reading instruction? Are there age- and level-appropriate books for this reading level? Do we need additional years of implementation to effect changes in Grade 4? Most likely it is a combination of factors, and an issue to be discussed by teachers, parents, and policy makers.

The USAID Quality Reading Project looks forward to continued partnership and collaboration with the Ministry of Education and Science of the Kyrgyz Republic, the National Testing Center, the Kyrgyz Academy of Education, teacher training institutes, regional and district education offices, school directors, teachers, National Children's Library, and parents.

Barbara A. Greenwood, MS, MPA

Chief of Party, USAID Quality Reading Project

[1] Results section of this report.

ACKNOWLEDGMENTS

The impact results using early grade reading assessment (EGRA) have been successfully analyzed because of the contributions of many organizations and individuals. The Ministry of Education and Science of the Kyrgyz Republic, National Testing Center (NTC), Pedagogy University, and Kyrgyz Academy of Education provided vital guidance and oversight for the whole assessment process. Also gratefully acknowledged is Bakirov Artur, director of NTC, for his constant involvement and follow-up.

This report would have been impossible without the generosity of the American people through the United States Agency for International Development (USAID).

The EGRA was implemented with the technical and logistical assistance of American Institutes for Research (AIR) and Save the Children International. AIR planned and executed its administration with the support of staff members in both its Washington, DC–based headquarters and the USAID Quality Reading Project office in Kyrgyzstan. Many thanks go to Project Manager Jerome Mindes, based in Washington, DC, for his overall role in supporting these activities; Chief of Party Barbara Greenwood; Deputy Chief of Party Kathryn Fleming; Saule Khamzina, monitoring and evaluation manager for technical and administrative leadership; and all USAID Quality Reading Project staff for their support. We also thank Amy Todd; Dr. Todd Drummond; Dr. Pooja Nakamura; Alvaro Ballarin Cabrera; and Paul Sirma in the International Development, Evaluation and Research Division at AIR for their collaborative endeavors in producing this midterm impact report.

EXECUTIVE SUMMARY

This report presents the midterm results of an impact evaluation conducted by American Institutes for Research for the United States Agency for International Development (USAID) Quality Reading Project for the Kyrgyz Republic. The USAID Quality Reading Project's main goal is to increase reading levels of Grades 1–4 students in Tajikistan and the Kyrgyz Republic through teacher training, mentoring, increased availability of reading materials, out-of-school activities, and policy changes. The findings presented cover the one-year impacts of the intervention on students in regions where the intervention was rolled out in year one. Our midterm sample covers 2,318 students in Grade 2 and 4 in 60 schools in these regions.

We used a randomized controlled trial (RCT) design to determine the impact of the program on student reading outcomes and intermediate outcomes that could lead to changes in student reading outcomes. A well-designed and well-implemented RCT is the most powerful research design for drawing conclusions about the impacts of an intervention on specific outcomes, such as reading. Our baseline analysis finds that randomization created equivalence across treatment and control groups at baseline, and for the midterm subsample of30 treatment and 30 control schools. We use an Early Grade Reading Assessment (EGRA) to gauge student reading outcomes. Decoding, language comprehension, and reading comprehension are the three main outcome measures. In addition, we will measure the impact of the program on dictation.

The program is having an impact on many student reading outcomes, but in most cases we do not find statistically significant effects on language comprehension. Beneficiary students are better able to decode written text, a fundamental reading construct, than control students. We find positive impacts of 0.30 standard deviations for Grade 2 and 0.28 for Grade 4 on decoding. These effects are both statistically and practically significant. To understand the meaning of the impact, we also examine the impacts on each component of the composite decoding measure. We find that beneficiary students know 7 more words per minute in Grade 2 than control students and 5 more words per minute in Grade 4 than control students on the reading passage section. Although we find some evidence for positive effects on language comprehension (Grade 2 only), the effect sizes are small and not statistically significant in most cases. However, we find evidence for positive statistically significant effects on students' ability to identify initial letter sounds, a subtask given only to Grade 2 students, where the program adds an additional 3.5 sounds per minute over the 93.6 for control group students. We do not find evidence of an impact on reading comprehension scores.
To understand how different groups may be impacted differently, we also conducted analyses for various subgroups of interest, including gender, language of instruction, and school location. We do not find evidence for statistically significant effects on the decoding or language comprehension skills of girls. However, our results show that the program has reasonably large and statistically significant effects on the decoding and language comprehension skills of boys. Interestingly, we find a 0.28 standard deviation impact on reading comprehension for Grade 4 boys, one of the few subgroups for whom we see impacts on this overall reading skill. The program also has positive effects on decoding and dictation on students who study in Kyrgyz in Grade 2, but no effect on language or reading comprehension. We do not find evidence for positive impacts on the reading

outcomes of students who study in Russian, with the exception of an impact on Grade 4 decoding skills. We find mixed results when looking at the results by school location, in terms of rural and urban schools. We see impacts on many skills in Grade 2 for rural students, including decoding, dictation, and reading comprehension. The program also appears to positively affect decoding skills for urban in Grade 2 and 4. However, we do not find evidence for positive effects on other skills for urban students.

1. INTRODUCTION

This report presents the midterm results of an impact evaluation conducted by American Institutes for Research (AIR) for the United States Agency for International Development (USAID) Quality Reading Project for the Kyrgyz Republic. We present a summary of the intervention, as well as a theory of change to lay out the conceptual framework that links project activities to reading outcomes. Next, we present a literature review to situate our project and midterm impact findings within the greater context of similar research findings in the early grade reading community internationally. We then discuss the research design section, which includes an explanation of the cluster randomized control trial (RCT) and difference-in-difference model we used to determine the impact of the program. This section also includes a discussion about balance (or equivalence) between treatment and control groups. We then explain the main outcome measures, reading skills tested through the Early Grade Reading Assessment (EGRA), and how they are presented in this report. Before presenting results of the midterm study, we provide descriptive statistics for both reading outcomes and background factors. The results section includes overall results, subgroup analysis, results for a longitudinal subsample of students, impact estimates on intermediate outcomes, and an associated factors analysis of language match (between school language of instruction [LOI] and home language) as a predictive factor for reading success. The report ends with sections on limitations and conclusion, including a discussion of recommendations.

2. INTERVENTION

The main goal of the USAID Quality Reading Project's is to increase reading levels of Grades 1–4 students in Tajikistan and the Kyrgyz Republic. It is a four-year project implemented by AIR and Save the Children International (SC). The project aims to improve reading skills among primary grade students by drawing on existing school system structures in both countries. The USAID Quality Reading Project is building capacity from the national level down to the classroom, all of which is aimed at supporting the common goal of improving student reading skills. The following descriptions (and findings in this report, in general) reflect the intervention and figures for the Kyrgyz Republic only.

2.1 TEACHER TRAINING

The main activities for the USAID Quality Reading Project in the Kyrgyz Republic focus on In-Service Teacher (IST) Training at the school level, community awareness about the

importance of reading, out-of-school reading activities, local development of reading material, and government partnership. The IST package is a locally adapted version of SC's Literacy Boost, which is focused on developing five reading skills during the lessons through the use of effective teaching methods, reading assessment, and enrichment of the reading environment in the classroom. These reading skills include phonemic awareness, phonics, vocabulary, reading fluency, and reading comprehension. The trainings follow a cascade model. The first level encompasses project staff training national trainers. The training is 72 hours long and conducted over 10 days. It focuses on providing skills and knowledge to train others to conduct comprehensive school-based training on teaching reading on the basis of the five skills. In the second level of training, national trainers train teachers and mentors over a 5-day period for a total of 40 hours. Mentors are the school deputy director, head of the methodological unit, and a selected advanced-level teacher from each school. The second level training focuses on a deep understanding of the five skills and associated national reading standards. Finally, training is conducted at a third level in project schools. Mentors lead teachers in weekly trainings and provide mentoring support (detailed below) for 32 hours over approximately 10 weeks. These school-based trainings focus on the application and practice of teaching reading. The program reaches approximately 60% of public primary schools in the country, for a total of 10,500 teachers.

2.2 MENTORING

In addition to providing direct training to teachers on early grade reading instruction, the USAID Quality Reading Project provides training in mentoring support around the new methods. Mentors in each school are given an additional day of training that focuses solely on how to provide mentorship support to all teachers. Mentorship creates a feedback loop for teaching reading where the teachers are able to reflect on their weekly lessons and mentors can provide ongoing support. To support mentors, the USAID Quality Reading Project, together with District Education Departments, conduct mentoring and monitoring visits to schools, organize district-level methodological meetings, and arrange planned and routine district-level meetings to share information and consultation on implementing the program during the academic year.

2.3 READING MATERIALS

The USAID Quality Reading Project also plans to increase the number of available low-cost reading materials by 500,000 over the life of the project. These materials will include grade-appropriate books, reading texts for teachers to use during the lessons, and more. Some materials have been developed and distributed already as a part of IST materials, but the majority of material distribution will be during the last two years of the project.

2.4 OUT-OF-SCHOOL ACTIVITIES

USAID Quality Reading Project staff are also conducting trainings for summer reading camps, following a cascade model. Project staff have trained one representative per school. Reading camps engage primary grade students in reading activities for 30 minutes to 2 hours

daily, for 21 days. The first camps were held in the summer of 2015, after the 2015 round of data collection reported on here.

2.5 POLICY LEVEL

Regional USAID Quality Reading Project offices have been established within the government structure. The locations of the regional teams are flexible based upon where space can be provided by the government; they have been established in the Regional Teacher Training Institutes, Regional Education Bureaus, District Education Offices, and at schools. The USAID Quality Reading Project regional teams are building relationships with the Regional Ministry of Education and Science (MOES) structures to support the level-three IST and out-of-school activities, and follow up on mentoring activities. Furthermore, USAID Quality Reading Project worked closely with Kyrgyz Academy of Education (KAE) on the development of Reading and Writing Standards and the IST packages. Competency-based Reading and Writing Standards were developed and approved by the KAE Academic Council in 2015. These standards are now being used as a base for State subject standards. The IST packages were approved by the KAE and the MOES collegium in May 2014.

ROLLOUT

Because of the nation-wide scale of the project, rollout of each of the above activities was divided into multiple waves. The country was divided by region based on the number of schools in each region and location of regions. Cohort 1 consists of Chui, Bishkek, Talas, and Jalalabad regions, which total 616 schools. Cohort 1 school-based training was completed in May 2015. Mentoring at the school level continues throughout the life of the project. During the August 2014 to April 2015 data collection period, 34 out-of-school reading camp and book-day activities were rolled out. Because of this phased roll-out design, the data collection has been planned accordingly to ensure that we are collecting data only in relevant areas. Baseline data collection was completed in the full sample of 130 schools (see Sampling section below for a full description of the sample size and power calculation). The first midterm data collection, which is the basis of this report, was conducted in a subsample of 60 schools that were in Cohort 1 regions. The second half of the midterm data will be collected during 2016 in the remaining regions. Endline data collection in 2017 will again cover the full nationally representative sample of 130 schools.

3. THEORY OF CHANGE

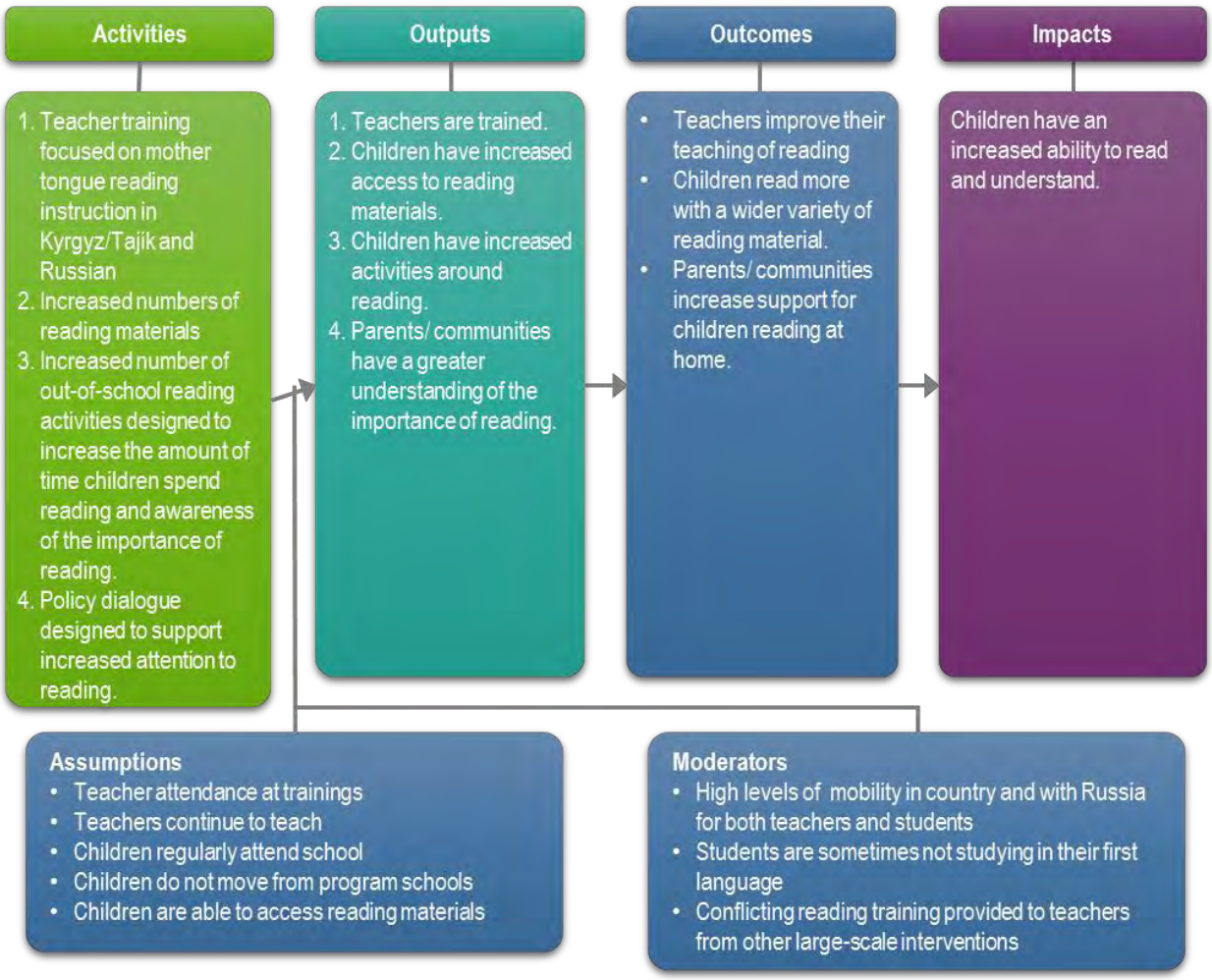
We believe that policy-relevant research and evaluation should be built on a theory of change that maps out the causal links among inputs, activities, outputs, outcomes, and impacts, as well as the underlying assumptions. Therefore, we begin this section with a description of the theory of change that underpins the teacher training program in order to ground our impact evaluation.

The theory of change we present is based on the design of the intervention, which focuses on training primary-grade teachers at level three of the training model in every cohort. The

training focuses on changing the way reading is taught in the primary grades and has a strong practical emphasis on instructional practices, as well as the use of assessment. The curriculum is based on SC's Literacy Boost, which focuses reading instruction on the five skills of reading (phonemic awareness, phonics, vocabulary, reading fluency, and reading comprehension). In the Kyrgyz Republic, the aim is to train primary-grade teachers at level two of the training model in every cohort in addition to level three. The intervention is assigned at the school level so that each of the teachers in the beneficiary schools receives training within the context of the USAID Quality Reading Project. Control schools are schools where none of the teachers receive any training (either at level two or three of the training model) in the context of the USAID Quality Reading Project. The control schools also do not receive the other program components.

We hypothesize that the teacher training will have a significant impact on teacher knowledge which, in turn, will significantly improve teacher practices in the classroom and student reading outcomes. Figure 3.1 presents the theory of change.

FIGURE 3.1. USAID/QUALITY READING PROJECT THEORY OF CHANGE



Several assumptions underpin the theory of change, each of which may need to be fulfilled for the teacher training to have a positive effect on teacher knowledge and behavior, and

student outcomes. First, implementation fidelity requires that the teachers are able to take part in the training, gain new knowledge and skills, and be motivated to apply what they learn during the training in the classroom. Second, students need to have the necessary inputs and incentives (such as textbooks and motivation to go to school) in order to achieve positive reading outcomes. Students must also continue to attend program schools in order to achieve positive impacts of the teacher training.

Moderators are factors outside the influence of the program that could influence the program's effectiveness. Expected moderators in this case include mobility of teachers and students, which may limit exposure of students to trained teachers and to other program activities. The multilingual environment of the Kyrgyz Republic is also a potential moderator, as students who do not have strong language skills in the language in which they study and are tested could be at a disadvantage. Finally, another expected moderator consists of the prior experience of teachers with teacher training or other messages about how to teach reading. We do not know whether other interventions would positively or negatively impact student reading outcomes, in interaction with our program's training.

Our evaluation design will examine the validity of the theory of change by analyzing the impact of the intervention, along with the causal chain among inputs, outputs, and intermediary and final outcomes. Intermediary outcomes will be tracked in the context of the monitoring and evaluation (M&E) system. The impact evaluation will estimate the impact of the intervention on the reading outcomes of students in Grades 1–4 and present some preliminary results about the impact on teacher behavior, student reading practices and parent's attitudes towards reading.

4. LITERATURE REVIEW

This midterm report presents evidence from the one-year midterm impacts of the USAID Quality Reading Project on student reading outcome using the EGRA as the measurement tool. There are many other examples of similar interventions. Here, we present the findings of impact studies of comparable interventions in order to put the USAID Quality Reading Project in context.

What do studies that use the EGRA measurement tool say about improving early grade reading? Although the multiple features of most early grade reading interventions make it difficult to attribute effects to particular aspects of the programs, the conclusions of several studies using the EGRA point to teacher training, curriculum design, and teacher accountability as being important factors. In addition, the overall literature on early grade reading, that is, beyond the set of studies that use the EGRA measurement, suggest that interventions should be designed in such a way that teaching can be matched to the student learning style and learning pace.

An early grade reading program using the EGRA that has highlighted the importance of teacher training is the "Girls' Improved Learning Outcomes" (GILO) in Egypt. The GILO is a program that focused on improving teaching quality through capacity building and

coaching, as well as by expanding equitable educational access for girls. The results of the intervention showed that children in treatment schools in 2011 had a score of 21.1 in oral reading fluency, whereas children from the same schools in 2009, before the program had been implemented, had an average score of 11.09. By contrast, the oral reading fluency average score of children in control schools changed from 8.92 in 2009 to only 10.93 in 2011. Other early reading interventions that have not used the EGRA measurement tool have also found similar effects of teacher training on early grade reading improvement (Conn, 2014; He, Linden, & MacLeod, 2009; McEwan, 2015 Murnane & Ganimian, 2014). Teacher training is also a core component of the USAID Quality Reading Project.

Although teacher training appears to be important, EGRA programs that do not focus exclusively on this area suggest that there may be other ways to improve early grade reading. One example is the "Read-Learn-Lead" (RLL) Program in Mali, which involved both teacher training and improving the curriculum for early grade children. The intervention offered a set of structured and systematic lessons and activities for instruction and practice on early reading skills. However, the curriculum was not designed to be adapted to the child's level, an element that has been found to be important for making pedagogical interventions effective (Kremer, Brannen, & Glennerster, 2013). In spite of this potential shortcoming, a difference-in-difference evaluation showed a positive impact on children's reading skills overall, for both Grade 1 and 2 students. Although children who participated in the treatment group did not improve their language skills and phonetic awareness relative to the control group, they did obtain significantly better scores in actual reading skills, such as letter recognition (effect size [ES]=0.40), decoding (ES=0.32), and reading familiar words (ES=0.33).

Similarly, an RCT of another intervention, the Integrated Education Program (IEP), finds evidence that supports the conclusions found in the RLL. The program also entailed coaching teachers and establishing a new curriculum based on a program called Systematic Method for Reading Success (SMRS). The idea behind the SMRS is that people learn to read best when they do so first in their mother tongue and when they begin by learning to read familiar words such as their name. However, unlike the RLL program, the IEP is a teacher-led adaptive instruction program. Students had to take a diagnostic exam, and the instructors were trained on how to assess student progress. The results of an RCT showed that children in the treatment school learned to read two to three times faster on average than children in the control schools. The strong effect sizes of the program are consistent with broader evidence in early grade reading that indicates that adaptive instruction is key for making teacher training programs effective. Conn (2014) reviewed a set of pedagogical interventions in Africa and found that among the interventions that used teacher training, the pooled effect size of programs with adaptive instruction was 0.42 standard deviations, whereas those without adaptive instruction only had an effect of 0.12 standard deviations.

Another factor that may boost the effectiveness of teacher training programs is making instructors accountable for student performance. The EGRA Plus program was grounded on this hypothesis, and created two different types of treatment groups to specifically measure

whether teacher accountability had any independent effects of teacher training. In the "full" treatment group, teachers had both the teacher training and accountability measures imposed on them, by making student performance data available to parents and family members. In the "light" treatment group, teachers were also held accountable by the same system but there was no teacher training component. The study found that students in the full treatment group had higher scores in all reading skills than their counterparts in the other groups. However, students in the "light" treatment group, where increased teacher accountability was implemented without accompanying teacher training, also significantly outperformed children in the control group. Both of these results combined suggest that both teacher training and accountability contribute to improving early grade reading. As with teacher training, other accountability-boosting interventions not using the EGRA measurement tool have also found similar results (Duflo, Hanna, & Ryan, 2012; Kremer et al., 2013; McEwan, 2015 Muralidharan & Sundararaman, 2009).

Several other studies demonstrate that there is only limited evidence for positive effects of other programs on early grade reading outcomes. For example, Evans (2015) finds no evidence for positive effects of cash grants, fee reductions, and deworming to improve learning outcomes. However, there is still a knowledge gap about what works in improving reading outcomes. Thus, it remains important to conduct rigorous evaluations with an emphasis on early grade reading outcomes.

5. RESEARCH DESIGN

We will measure impacts along the causal chain of inputs, activities, outputs, outcomes, and impacts. We rely on the tracking of progress in the M&E plan to assess the inputs and outputs of the intervention. In addition, we use the RCT design to determine the impact on student reading outcomes, as well as intermediate outcomes such as teacher behavior and attitudes towards reading that we theorize lead to change in student reading outcomes.

5.1 BENEFITS OF RANDOMIZATION

A well-designed and well-implemented RCT is the most powerful research design for drawing conclusions about the impacts of an intervention on specific outcomes. An RCT draws from a pool of comparable schools and then randomly assigns some to a treatment group that receives the intervention and others to a control group that does not receive the intervention, against which comparisons can be made. An RCT permits us to directly attribute any observed differences between the treatment and control students to the intervention; otherwise, other unobserved factors, such as motivation, could have influenced members of a group to move into a treatment or a control group (Campbell & Stanley, 1963). Randomization helps ensure that both observed and unobserved characteristics that may affect the outcomes are similar between the treatment and control conditions of the sample. In a randomized experiment, treatment and control groups are expected to be comparable (with possible chance variation between groups) so that the average differences in outcome

between the two groups at the end of the study can be attributed to the intervention. Our analysis of comparison and treatment groups finds that randomization created equivalent groups at baseline for the evaluation of the program.

Balance between treatment and control schools: At baseline, we found that the treatment and control groups are statistically equivalent in observable characteristics. Treatment and control schools are balanced on all the EGRA subtasks with the exception of Grade 2 listening comprehension. Treatment schools performed, on average, almost five percentage points higher than control schools in listening comprehension. One statistical difference between the two groups at baseline is, however, not a cause for concern about randomization. Thus, our randomization can be considered successful.

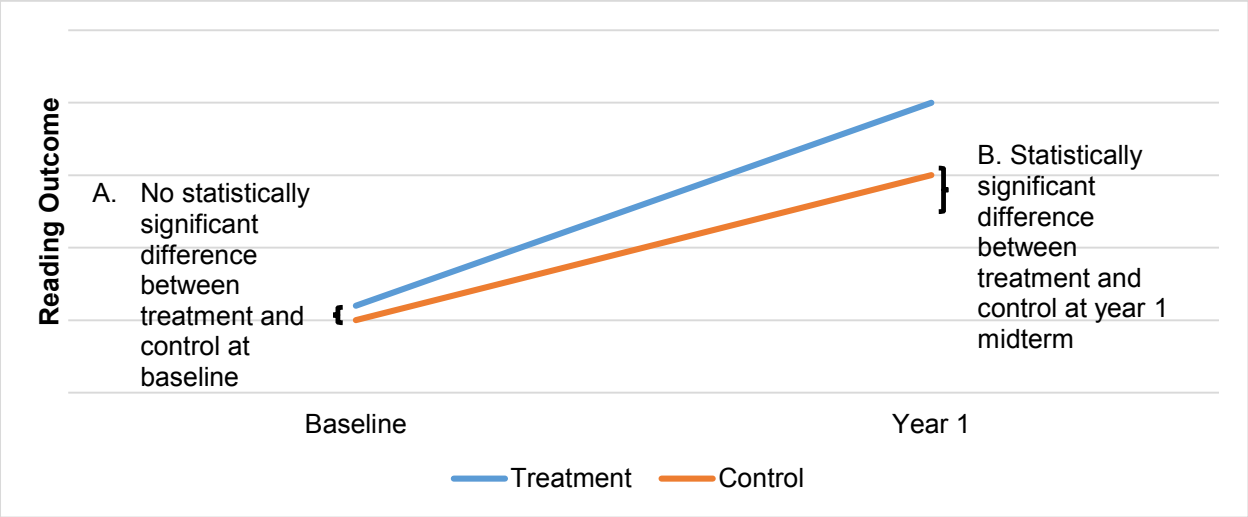
We also tested whether the treatment and control groups are statistically equivalent across the variables of interest in the subsample of 60 schools that we discuss in this report, because this midterm reports only the one-year impacts in Bishkek, Chui, Jalal-Abad, and Talas. The baseline scores for these 60 schools are balanced between treatment and control schools for all the EGRA subtasks. There are no statistically significant differences between treatment and control school EGRA performance. See Appendix A for balance tables of the full, 130-school sample, and the 60-school, 2015, midterm subsample.

5.2 STATISTICAL MODEL TO ESTIMATE IMPACTS

The USAID Quality Reading Project team uses a cluster RCT using a difference-in-difference model to determine the impact of the program on teacher behavior and reading outcomes. The difference-in-difference model is a statistical technique that allows for estimating the causal effect of the teacher training program by comparing changes in outcomes (EGRA scores and teacher behavior) over time across the treatment group and the control group. The difference-in-difference model is the most rigorous method to estimate the impact of the program because the model controls for unobservable factors that do not change over time by subtracting the changes in outcomes for the control group from the changes in outcomes of the beneficiaries.[2] Figure 5.1 depicts the difference-in-difference model through a graph. The figure shows that the impact of the program is the value of B minus the value of A.

[2] Results below are shown without control variables, as results are roughly the same with or without covariates.

FIGURE 5.1. DIFFERENCE-IN-DIFFERENCE MODEL



5.3 MAIN OUTCOME: EARLY GRADE READING ASSESSMENT

Initiated in 2013, the USAID Quality Reading Project has introduced reading interventions at the school and community levels, as well as rolled out an EGRA. Though EGRA does not serve as a complete assessment of the state of reading in the republic, EGRA diagnostic results can serve to focus on the contextual, resource, and pedagogical obstacles to the development of early grade reading skills. Research conducted on primarily monolingual, alphabetic language learners shows that there are five important components of early grade reading: phonemic awareness, phonics, oral reading fluency, vocabulary, and reading comprehension (NICHD, 2006). The Kyrgyz Republic EGRA incorporates an assessment of these areas, but the assessment has also been carefully customized to be linguistically and culturally appropriate for use in the republic. The project administered a baseline EGRA in April, 2014. Data was collected in four regions of the republic on reading in Grades 1, 2 and 4; and Kyrgyz and Russian languages. In April, 2015, a follow-up midterm assessment was conducted for Grades 2 and 4, and the results of that assessment were the focus of the recently released *2015 EGRA Midterm Analytic Report*.

In 2015, the EGRA in the Kyrgyz Republic had nine total subtasks.[3] The 2014 section, *Unfamiliar Words* (2014), was implemented as *Nonsense Words* in 2015.[4] Table 5.1 below presents the nine total tasks (second grade) and seven tasks (fourth grade), and the demands of the tasks for the 2015 assessment in both the Kyrgyz and Russian languages. Four of the

[3] One subtest from 2014, *Initial Word Sound,* was dropped from the 2015 EGRA because there was no difference in the results between the *Initial Word Sound* and *Initial Letter Sound* subtasks, and Russian and Kyrgyz language experts recommended excluding this subtask.

[4] In the *Unfamiliar Words* task, pupils were asked to read actual words, but ones that were judged to be above the pupils' grade level. Because it is difficult to control exposure to such words, it cannot be determined if decoding skills or memorization/familiarity is being used in the accomplishment of the task. In the *Nonsense Words* task, items are not actual words but ones that resemble the structure of common one- and two-syllable grade-level words. This task represents a more uncontaminated measure of decoding. The objective of this task is to determine how well pupils can associate written letter combinations with their spoken forms (words) without relying on sight reading (Hirsch, 2003).

nine subtasks for second graders were timed and three of the seven subtasks for fourth graders were timed. Pupils were given a maximum of 120 seconds to complete each timed subtask. These timed tasks were reading *Letter Name Recognition*, *Familiar Word Recognition, Nonsense Words,* and *Reading Passage.* A more detailed description of each subtask is presented in the *2015 EGRA Midterm Analytic Report*.

TABLE 5.1. EGRA IN THE KYRGYZ REPUBLIC

2015 Midterm Subtasks		
Subtask (Grade)	**Reading Skills**	**Pupils were asked to:**
1. Letter Name Recognition (2)	Letter identification	*Identify* correctly and *read* aloud letters of the alphabet in lower and upper cases in a 2-minute period (TIMED)
2. Initial Letter Sound (2)	Phonemic awareness, letter–sound correspondence	*Sound out* 10 commonly used letters, randomly arranged, repeating after administrator
3. Familiar Word Recognition (2,4)	Word recognition and decoding	*Read aloud* 40 familiar[5] 1- and 2-syllable words in a 2-minute period (TIMED)
4. Nonsense Word Recognition (2,4)	Letter–sound correspondence, decoding	*Decode* 40 1- and 2-syllable nonsense words in a 2-minute period (TIMED)
5. Oral Vocabulary (2,4)	Basic vocabulary, listening in context	*Identify* 10 objects in pictures after listening to a list of objects read by the administrator using the validated PPVT-R format
6a. Passage Reading (2,4)	Reading fluency	*Demonstrate* oral reading of 1 short passage in a 2-minute period (TIMED)
6b. Reading Comprehension (2,4)	Reading comprehension of texts	*Demonstrate* reading comprehension by answering 4–5 oral questions from the administrator about the reading passage just read aloud
7. Listening Comprehension (2,4)	Oral language comprehension, vocabulary knowledge	*Demonstrate* listening comprehension by answering 4–5 questions based on a short paragraph read by the administrator
8. Dictation (2,4)	Oral language comprehension, writing skills	*Listen* to a sentence and *reproduce* it correctly in written form

CORRELATIONS BETWEEN SUBTASKS

As part of our analyses of the EGRA subtasks, we explored the relationships between subtasks. In Table 5.2 below, we present the correlation matrix of the Russian Grade 2 EGRA subtasks as an example (the other three matrices can be found in Appendix B). As might be expected, *Initial Letter Sound*, which taps into phonological skills, was not highly correlated with any other subtask. The correlations between *Oral Vocabulary, Listening Comprehension*, and *Reading Comprehension* were reasonably high at .61 and .54, respectively. Note that *Reading Passage*, an indicator of fluency and predictor of comprehension, was correlated at moderately high or high levels with all the other subtasks except *Initial Letter Sound*. *Dictation* was another subtask that correlated moderately with

[5] "Familiar" words were identified through an EGRA protocol that requires word analyses on grade-level textbooks to derive counts of the most commonly encountered words.

most all of the subtasks. The subtasks that all loaded highly on the construct "decoding" in the factor analysis (see following section) were highly correlated: *Familiar Words* and *Nonsense Words* (.83), *Nonsense Words* and *Reading Passage* (.87), and *Reading Passage* and *Familiar Words* (.88).

TABLE 5.2. CORRELATION MATRIX OF GRADE 2 RUSSIAN EGRA SUBTASKS

Russian Grade 2		Initial Letter Sound	Oral Vocab	Passage Comp	Listening Comp	Dictation	Letter Name	Nonsense Words	Familiar Words	Reading Passage
Correlation	Initial Letter Sound	1.000								
	Oral Vocabulary	.069	1.000							
	Passage Comprehension	.028	.541	1.000						
	Listening Comprehension	.011	.610	.562	1.000					
	Dictation	.160	.300	.463	.330	1.000				
	Letter Name	.131	.235	.293	.148	.275	1.000			
	Nonsense Words	.161	.142	.307	.159	.415	.340	1.000		
	Familiar Words	.199	.176	.330	.173	.444	.294	.832	1.000	
	Reading Passage	.171	.237	.390	.255	.480	.260	.869	.877	1.000

UNDERLYING READING SKILL FACTORS IN EGRA DATA AS OUTCOME MEASURES

We conducted a factor analysis to determine what underlying constructs relating to reading are being tested by the EGRA subtasks. The goal was to consider combining subtasks as outcome measures for looking at the impact of the program. We are interested in measuring the impact of the program on children's *reading ability*, which is arguably best represented by the underlying skill factors in the EGRA subtasks. The detailed results of our factor analysis are presented in Appendix C. In short, we have averaged the standardized values of the familiar word, nonsense word, reading passage fluency, and letter-naming subtasks to a construct that represents *decoding*. Listening comprehension, oral vocabulary, and initial letter sound standardized values have been averaged to create a construct for *language comprehension*. Dictation did not clearly load with these existing factors, so it has been left separate. Finally, reading comprehension remains a separate outcome of interest, as it is the end goal of the other reading skills.

Decoding, language comprehension, and reading comprehension (with dictation also examined separately, although a less important outcome in looking at the impact of the program on reading) are the three main constructs we will be using as outcome measures. Impact will also be measured for each subtask separately, but those measurements have less power than the measurements with a greater number of data points per child.

5.4 SAMPLING

This study relies on a full sample of 130 schools, 65 of which are treatment schools receiving the intervention, and 65 of which are control schools that would be eligible to receive the intervention but were not randomly assigned treatment status. The sample size should allow us to detect a small but relevant effect of the intervention of approximately 0.16–0.2 standard deviations at a minimum. These effect sizes are well within the normal accepted range for education interventions.

Nonetheless, because of the staggered rollout of the intervention, described above in the Intervention section, our midterm sample is limited to 60 schools in regions where the intervention rolled out in the first year. These 60 schools are evenly balanced between treatment and control schools. The remaining 70 schools will be the subject of 2016 data collection, and the second half of the midterm impact report. As demonstrated in the above section on balance, and in Appendix A, these 60 schools are balanced at baseline on the outcomes of interest.

6. DESCRIPTIVE STATISTICS

The following statistics present the EGRA reading results by subtask, broken down by treatment and control groups. For a full description of the instrument, please see the earlier section entitled "Main outcome: early grade reading assessment."

TABLE 6.1. READING OUTCOME SUMMARY STATISTICS BY TREATMENT ARM

Grade 2	Treatment (n=481)		Control (n=501)		Total (n=982)	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Letters Named per Minute	67.98	18.50	66.41	19.69	67.62	18.78
Initial Letter Sounds per Minute	95.51	10.43	93.63	12.18	95.09	10.87
Familiar Words per Minute	59.67	26.15	54.87	27.23	58.58	26.46
Nonsense Words per Minute	27.41	11.25	24.78	11.05	26.81	11.25
Oral Vocabulary	87.80	12.92	87.33	13.48	87.69	13.04
Reading Passage Words per Minute	42.05	20.46	36.55	19.18	40.81	20.30
Reading Passage Comprehension	57.95	31.08	50.12	33.85	56.18	31.88
Listening Comprehension	73.22	29.60	65.80	28.30	71.54	29.46
Dictation	75.44	23.37	71.19	25.73	74.48	23.97

Grade 4	Treatment (n=506)		Control (n=481)		Total (n=987)	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Familiar Words per Minute	90.84	33.29	85.67	33.76	89.74	33.44
Nonsense Words per Minute	34.15	12.80	32.25	13.03	33.74	12.86
Oral Vocabulary	90.47	11.27	89.89	12.71	90.34	11.59
Reading Passage Words per Minute	61.67	23.67	56.05	25.30	60.47	24.13
Reading Passage Comprehension	77.41	26.49	72.60	27.60	76.38	26.79
Listening Comprehension	68.16	28.94	65.75	29.69	67.65	29.11
Dictation	81.83	20.12	80.02	19.94	81.45	20.09

The summary statistics in Table 6.1 show that children, overall, are best at oral vocabulary and dictation in both grades, and in Grade 4, children also have high scores in reading passage comprehension. The statistics also show that, for both grades, the average score in the treatment group is higher than in the control group in every category. In Grade 2, these differences are especially large in reading passage comprehension, listening comprehension, reading passage words per minute, familiar words per minute, and dictation. In Grade 4, the differences are largest for reading passage words per minute, familiar words per minute, and reading passage comprehension. The differences in mean scores between treatment and control, overall, are larger for Grade 2 than for Grade 4, though Grade 4 has a greater difference in some categories, such as reading passage words per minute, familiar words per minute, and oral vocabulary.

TABLE 6.2. BACKGROUND CHARACTERISTIC SUMMARY STATISTICS

Grades 2 and 4	Total (n=2318)	
	Mean	Standard deviation
Attended Kindergarten	42.60%	0.49
Attended Preschool	38.35%	0.49
Study in Kyrgyz at School	79.34%	0.40
Speak in Kyrgyz at Home	92.04%	0.27
Student's Family Owns Television	99.62%	0.61
Student Lives in Bishkek Region	11.22%	0.32
Student Is Male	52.51%	0.50

The mean values of the background characteristics at baseline show that slightly less than half of students have attended kindergarten in both grades, and over a third have attended preschool. Most children in both grades speak Kyrgyz both at home and at school, and only 10%–20% speak Russian. The gender breakdown is almost even in both grades, but there are slightly more boys than girls.

FIGURE 6.1. PERCENT OF STUDENTS BY HOUSEHOLD SIZE

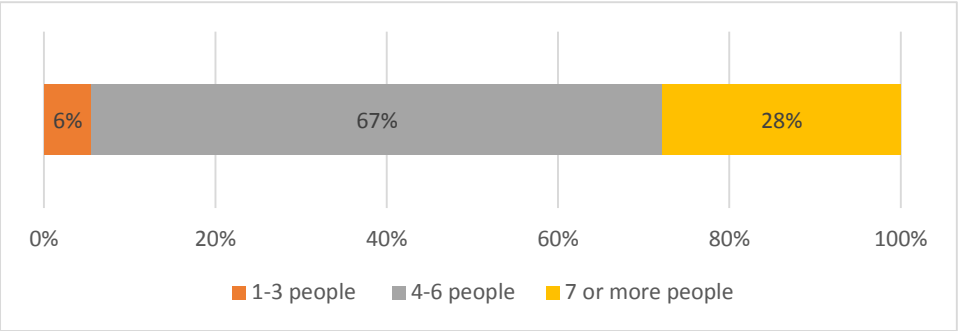


FIGURE 6.2. PERCENT OF STUDENTS BY MOTHER'S EDUCATION

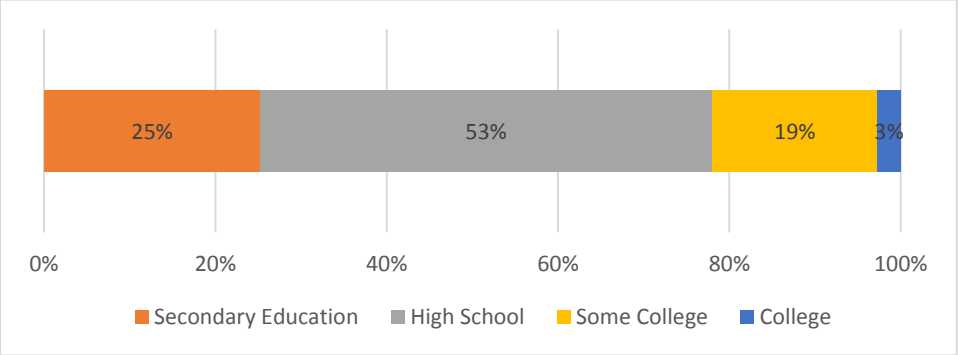


Figure 6.1 and 6.2 show descriptive statistics at the household level. The first figure also shows demonstrates that the majority of children come from homes where the household size is between 4 and 6 people. With respect to socioeconomic indicators, the vast majority of children come from families that own a television. And with respect to educational achievement in their families, about Furthermore, 52% of children come from families where their mothers have completed high school, about 25% come from families where mothers

only haveonly completed some secondary education, 20% come from families where mothers completed have some college education, and about 3% are part of households where mothers have completed college.

7. RESULTS

This section presents the impact of the intervention on reading outcomes and other intermediate outcomes. We first present a repeated cross-sectional analysis to determine the impact of the program on variables measuring reading outcomes, including composite variables of decoding and language comprehension. Second, we present separate analysis for different subgroups because our theory of change indicates that the impact of intervention may vary depending on the characteristics of the population of interest, such as gender and language. Third, we explore the impact results for the longitudinal sample of students now in Grade 2. Fourth, we analyze the impact of the program on other outcomes along the causal chain of the theory of change, including teacher behavior, parental attitudes, and children's participation in reading activities and at-home reading practices, as our theory of change shows that the program can result in improvements in reading outcomes as a result of these outcomes. Finally, we also present the results of an analysis that looks at the LOI and home language as factors that predict reading outcomes.

It is key to understand how to read the results, which are presented in table and graphical form. The tables presented here show the standardized scores for each subtask, so that results can be interpreted as the impact expressed in standard deviations. The charts here show the average score that students receive in the absence of the program, which is the average score of the control group in 2015, as well the score achieved with the program, shown in the same bar but in a different shade. The score achieved with the program is attributable to the program only in cases where the label has asterisks (*, **, or ***). The asterisks denote statistical significance, which refers to the probability of finding that the program has an impact due to chance alone when there is no true impact. A p-value measures this probability and is grouped into 3 categories: less than 0.10 or 10% has 1 asterisk (*), less than 0.05 or 5% has 2 asterisks (**), and less than 0.01 or 1% has 3 asterisks (***). The lower the p-value, the more certain we are in the results we are reporting. However, it is worth noting that a p-value of 0.05 is generally accepted as adequate to call the results statistically significant. Additional regression tables can be found in Appendix D. A full set of tables is available upon request.

7.1 IMPACT ON READING OUTCOMES

The program is having an impact on student reading outcomes. Our theory of change suggests that the project activities could have a positive effect on reading outcomes through several mechanisms. We present the impact on the composite measures of decoding and language comprehension based on the results of the factor analysis, as well as on dictation and reading comprehension separately.

Impact on decoding: Our results demonstrate that the project has meaningful effects on variables that are associated with decoding. Beneficiary students are better able than control

students to decode written text, which is a fundamental reading construct. Table 7.1 below shows the impact on decoding, differentiated by grade. An impact of 0.30 standard deviations for Grade 2 and 0.28 for Grade 4 are both statistically significant and meaningful.

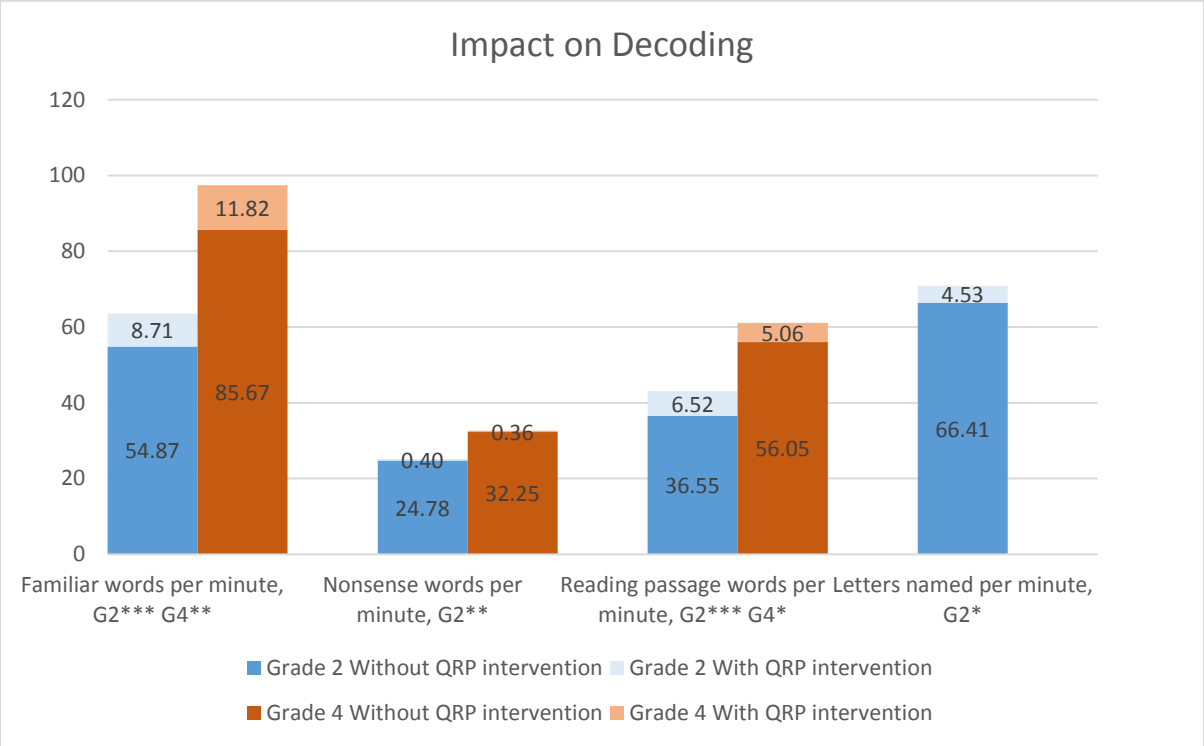
TABLE 7.1. DECODING EFFECT SIZES, BY GRADE

Dependent Variable	12-Month Impact (1)	Baseline Mean (2)
Decoding, Grade 2	0.30***	0.08
	(3.21)	
N	1,957	482
Decoding, Grade 4	0.28**	0.02
	(2.15)	
N	2,009	520

Notes: Estimations use difference-in-difference modeling. Robust t-statistics clustered at the school level are in parentheses. **=p<0.05, ***=p<0.01

In order to understand the meaning of the impact, we also examine the impacts on each component of the composite decoding measure. We find that beneficiary students know 7 more words per minute in Grade 2 than control students, and 5 more words per minute in Grade 4 than control students. The impacts on reading fluency for familiar nonconnected words are larger: 9 words for Grade 2 and 12 words for Grade 4. Figure 7.1 below shows the impacts on decoding subtasks. The results remain similar after we control for a range of control variables, including gender, mother's education, home language, kindergarten attendance, household size, an asset index used as a measure of socioeconomic status, and whether the student attends school in Bishkek.

FIGURE 7.1. IMPACT ON DECODING SUBTASKS, BY GRADE



Note: *=p <0.10, **=p<0.05, ***=p<0.01

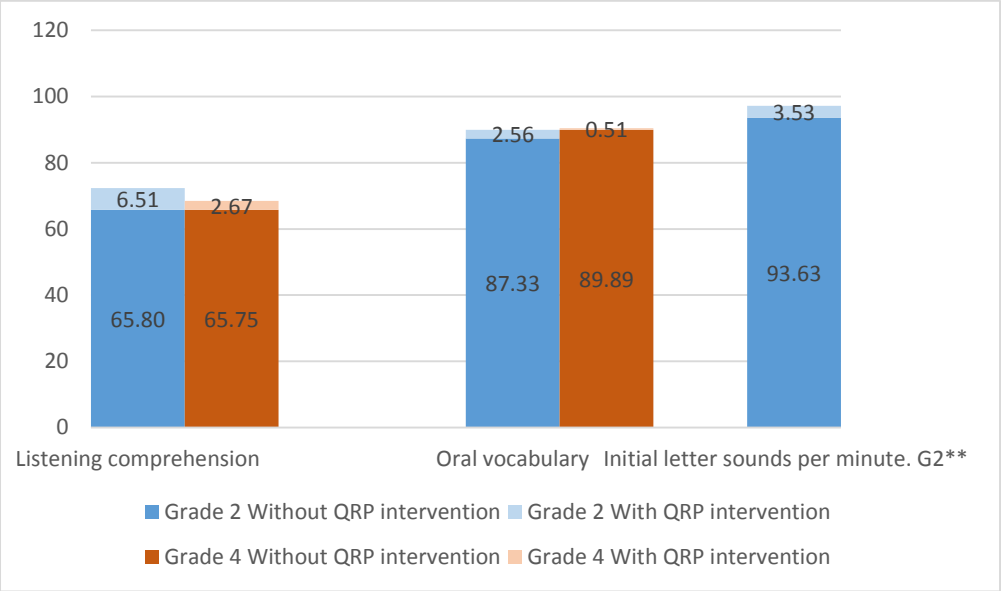
Impact on language comprehension: Although we find some evidence for positive effects on language comprehension (Grade 2 only), the effect sizes are small and not statistically significant in most cases. Table 7.2 below shows these results by grade. Our results show no evidence for positive and statistically significant effects of the program on listening comprehension or oral vocabulary. However, we find evidence for positive statistically significant effects on students' ability to identify initial letter sounds, a subtask given only to Grade 2 students, where the program leads to an additional 3.5 sounds per minute over the 93.6 sounds per minute for control group students. Figure 7.2 depicts the effects of the program on subtasks that make up language comprehension.

TABLE 7.2. LANGUAGE COMPREHENSION EFFECT SIZES, BY GRADE

Dependent Variable	12-Month Impact (1)	Baseline Mean (2)
Language comprehension, Grade 2	0.25*	0.18
	(1.85)	
N	1,957	482
Language comprehension, Grade 4	-0.04	0.06
	(-0.31)	
N	2,009	520

Notes: Estimations use difference-in-difference modeling. Robust t-statistics clustered at the school level are in parentheses. *=p<0.10

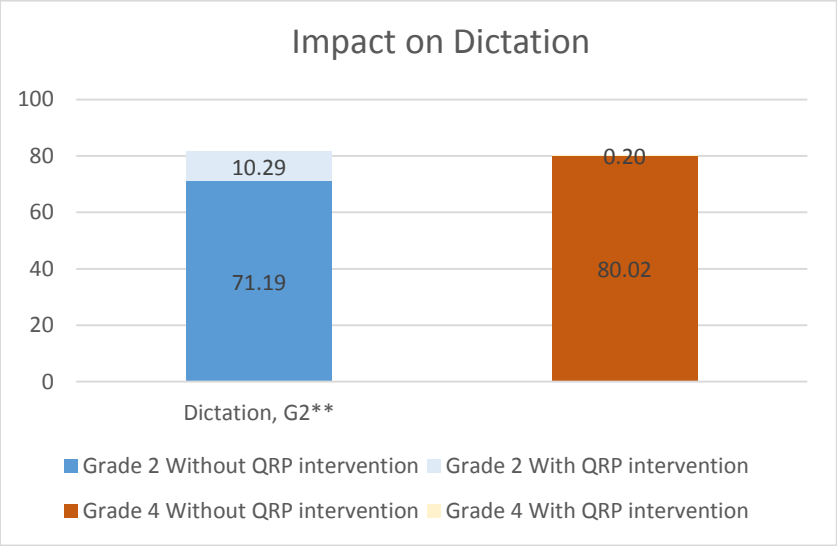
FIGURE 7.2. IMPACT ON LANGUAGE COMPREHENSION SUBTASKS, BY GRADE



Note: **=p<0.05.

Impact on dictation: The program has a significant impact on dictation skills for Grade 2 students, improving their ability to write down dictated text by 10 percentage points. We do not find program impacts for Grade 4 students. Values are shown in Figure 7.3 below.

FIGURE 7.3. IMPACT ON DICTATION, BY GRADE

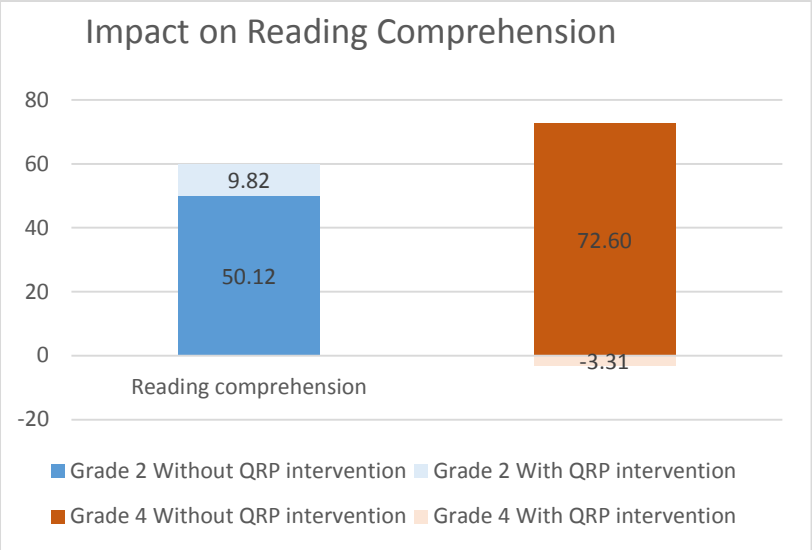


Note: **=p<0.05.

Impact on reading comprehension: We do not find evidence of a positive average e impact on reading comprehension scores. There are several explanations for this lack of a positive effect, but a lack of evidence for positive impacts does not necessarily imply that there is no program impact. We measure reading comprehension through 4-5 oral questions of a student after s/he reads a short grade-level appropriate text. Because the reading comprehension measure relies on five or fewer questions there is only limited variation in the

reading comprehension measure. This reduces the statistical power to detect a positive impact. At the same time, however, it may well be possible that the program does not have a positive effect on reading comprehension for the overall sample. Our heterogeneous impact estimates show that in certain sub-samples we do have sufficient statistical power to detect small but meaningful effects of the intervention. Reading comprehension results are expressed graphically below, in Figure 7.4.

FIGURE 7.4. IMPACT ON READING COMPREHENSION, BY GRADE



Our evidence indicates that the program is having an impact on reading outcomes, but we do not find clear evidence for improvements in reading comprehension as there are clear results that student skills in early reading skills relating to decoding are improving because of the program. However, these impact do not extend to comprehension, either as a composite measuring language comprehension, or when looking at reading comprehension. The implications of these findings are discussed in the Conclusion below.

7.2 SUBGROUP ANALYSIS

The results presented above are average treatment effect for the entire population. In order to understand how different groups may be impacted differently, we also conducted impact estimations on various subgroups of interest. These include looking at impact by gender, LOI, and school location.

Impact by gender: Perhaps surprisingly, we find only little evidence for positive effects of the program on girls' reading outcomes. We do not find evidence for statistically significant effects on the decoding or language comprehension skills of girls. Table 7.3 demonstrates that effect sizes are also relatively small (<0.25 standardized mean differences), which indicates that the program has not been successful in improving the reading outcomes of girls. Full results for each subtask are available in Appendix D. Possibly, the program does not result in positive effects for girls because of the relatively strong reading skills of girls at baseline. The program may be more effective for marginalized students with low reading

skills. Alternatively, our sample size may be too small to detect small but meaningful effects of the program on girls' outcomes.

Our results show that the program has reasonably large and statistically significant effects on the decoding and language comprehension skills of boys. Tables 7.3 and 7.4 show that the program has positive effects on both decoding, language comprehension, and dictation for Grade 2, and decoding only for Grade 4. Interestingly, we find a 0.28 standard deviation impact on reading comprehension for Grade 4 boys, one of the few subgroups for whom we see impacts on this overall reading skill. For the reading passage subtask, we find that male beneficiary students know 8 more words per minute in Grade 2 than male control students, and 8 more words per minute in Grade 4 than male control students. We find similarly large effects on reading comprehension for Grade 4 boys, though no effects for Grade 2 boys.

TABLE 7.3. EFFECTS BY STUDENT GENDER, GRADE 2

Dependent Variable	Girls (1)	Boys (2)	Test Girls=Boys (3)	Mean Girls (4)	Mean Boys (5)
Decoding	0.20	0.37***	0.28	0.13	-0.19
	(1.66)	(3.14)			
Language Comp.	0.13	0.35**	0.19	-0.03	-0.01
	(0.87)	(2.11)			
Dictation	0.20	0.57***	0.08*	0.20	-0.19
	(1.26)	(2.75)			
Reading Comp.	0.13	0.45	0.18	0.09	-0.12
	(0.64)	(1.45)			
N		1,957		498	486

Notes: Estimations use difference-in-difference modeling. Robust t-statistics clustered at the school level are in parentheses. Column (3) presents the p-value of an F-test for the equality of the impacts in Columns (1) and (2). Columns (4) and (5) show the means at baseline for the corresponding subgroup. *=p<0.10, ***=p<0.01

TABLE 7.4. EFFECTS BY STUDENT GENDER, GRADE 4

Dependent Variable	Girls (1)	Boys (2)	Test Girls=Boys (3)	Mean Girls (4)	Mean Boys (5)
Decoding	0.20	0.35**	0.34	0.22	-0.24
	(1.46)	(2.20)			
Language Comp.	-0.12	0.04	0.42	-0.01	0.02
	(-0.68)	(0.26)			
Dictation	-0.10	0.12	0.26	0.23	-0.22
	(-0.92)	(0.72)			
Reading Comp.	-0.10	0.28*	0.06*	0.06	-0.07
	(-0.61)	(1.79)			
N		2,009		490	497

Notes: Estimations use difference-in-difference modeling. Robust t-statistics clustered at the school level are in parentheses. Column (3) presents the p-value of an F-test for the equality of the impacts in Columns (1) and (2). Columns (4) and (5) show the means at baseline for the corresponding subgroup. *=p<0.10, **=p<0.05

Impact by school LOI: The program has positive effects on Grade 2 students who study decoding and dictation in Kyrgyz but no effect on language or reading comprehension. These same effects are not found in Grade 4, with the exception of a weaker effect on decoding. Students who study in Russian are not seeing an impact of the program, with the exception of an impact on Grade 4 decoding skills. These differences between groups are not statistically significant. The lack of impact on students who study in Russian is not surprising, and could have a number of reasons. First, the program may not affect the decoding and language comprehension skills of Russian speakers because their skills are already well-developed at baseline. Second, our sample size may be too small to detect small but meaningful effects of the program on Russian speakers.

TABLE 7.5. EFFECTS BY LANGUAGE OF INSTRUCTION AT SCHOOL, GRADE 2

Dependent Variable	Russian (1)	Kyrgyz (2)	Test Russian=Kyrgyz (3)	Mean Russian (4)	Mean Kyrgyz (5)
Decoding	0.21	0.32***	0.53	-0.03	-0.03
	(1.37)	(3.07)			
Language Comp.	0.24	0.25	0.94	-0.01	-0.02
	(1.28)	(1.60)			
Dictation	0.37	0.41**	0.89	-0.01	0.01
	(1.20)	(2.28)			
Reading Passage Comp	0.33	0.29	0.92	-0.01	-0.01
	(1.14)	(1.05)			
N		1,957		325	659

Notes: Estimations use difference-in-difference modeling. Robust t-statistics clustered at the school level are in parentheses. Column (3) presents the p-value of an F-test for the equality of the impacts in Columns (1) and (2). Columns (4) and (5) show the means at baseline for the corresponding subgroup. **=p<0.05, ***=p<0.01

TABLE 7.6. EFFECTS BY LANGUAGE OF INSTRUCTION AT SCHOOL, GRADE 4

Dependent Variable	Russian (1)	Kyrgyz (2)	Test Russian=Kyrgyz (3)	Mean Russian (4)	Mean Kyrgyz (5)
Decoding	0.39**	0.26*	0.58	-0.02	-0.01
	(2.13)	(1.84)			
Language Comp.	0.00	-0.04	0.86	0.02	-0.01
	(0.02)	(-0.24)			
Dictation	-0.05	0.03	0.73	0.01	-0.00
	(-0.24)	(0.28)			
Reading Passage Comp.	0.13	0.09	0.91	-0.01	0.00
	(0.53)	(0.61)			
N		2,009		311	676

Notes: Estimations use difference-in-difference modeling. Robust t-statistics clustered at the school level are in parentheses. Column (3) presents the p-value of an F-test for the equality of the impacts in Columns (1) and (2). Columns (4) and (5) show the means at baseline for the corresponding subgroup. *=p<0.10, **p<0.05

Impacts by school location: We find mixed results when looking at the results by school location, differentiated by rural and urban schools. We see impacts on many skills in Grade 2 for rural students, including decoding, dictation, and reading comprehension. In particular,

the impact on reading comprehension—something we have not often found when looking at other subgroups or average treatment effects—is substantial at 0.39 standard deviations. This translates to a 13-percentage-point improvement on the reading comprehension score for Grade 2 rural students. We do not find large or statistically significant effects for rural students in Grade 4.

Urban students in both Grades 2 and 4 have improved decoding skills as a result of the program. We are unable to find effects on other skills for urban students. The decoding impacts for Grade 4 urban students are quite large, at 0.61 standard deviations. Beneficiary students are able to read 14 more words per minute on the reading passage subtask, one of the subtasks groups under decoding.

TABLE 7.7. EFFECTS BY SCHOOL LOCATION, GRADE 2

Dependent Variable	Rural (1)	Urban (2)	Test Urban=Rural (3)	Mean Rural (4)	Mean Urban (5)
Decoding	0.28**	0.24*	0.80	-0.04	0.02
	(2.54)	(1.84)			
Language Comp.	0.18	0.25	0.80	-0.04	0.05
	(1.26)	(0.98)			
Dictation	0.50***	0.20	0.43	-0.03	0.13
	(3.21)	(0.57)			
Reading Passage Comp.	0.39*	0.03	0.48	-0.08	0.20
	(1.87)	(0.07)			
N		1,942		738	246

Notes: Estimations use difference-in-difference modeling. Robust t-statistics clustered at the school level are in parentheses. Column (3) presents the p-value of an F-test for the equality of the impacts in Columns (1) and (2). Columns (4) and (5) show the means at baseline for the corresponding subgroup. *=p<0.10, **=p<0.05, ***=p<0.01

TABLE 7.8. EFFECTS BY SCHOOL LOCATION, GRADE 4

Dependent Variable	Rural (1)	Urban (2)	Test Urban=Rural (3)	Mean Rural (4)	Mean Urban (5)
Decoding	0.05	0.61***	0.03**	-0.06	0.13
	(0.39)	(2.72)			
Language Comp.	0.04	-0.27	0.24	-0.01	0.05
	(0.25)	(-1.27)			
Dictation	-0.04	0.05	0.72	-0.06	0.18
	(-0.28)	(0.24)			
Reading Passage Comp.	0.02	0.18	0.52	-0.04	0.12
	(0.13)	(0.86)			
N		1,991		738	249

Notes: Estimations use difference-in-difference modeling. Robust t-statistics clustered at the school level are in parentheses. Column (3) presents the p-value of an F-test for the equality of the impacts in Columns (1) and (2). Columns (4) and (5) show the means at baseline for the corresponding subgroup. **=p<0.05, ***=p<0.01

7.3 LONGITUDINAL SUBSAMPLE RESULTS

This section serves to assess the impact of the program on student decoding, language, and reading comprehension outcomes for a subsample of students for which we collected longitudinal data. The collection of longitudinal data refers to collecting data for the same student over different years. We managed to collect these longitudinal data for 404 beneficiary students and 404 control students who were in Grade 1 during the baseline data collection and in Grade 2 during the midterm data collection. This sample of 808 students covers 71% of the sample of 900 students who we hoped to reach in our longitudinal sample; thus, our attrition rate is 29%. This attrition was caused by students moving to different schools or geographies and by students who did not attend school on the EGRA date.

The results of the impact of the program on student decoding, language, and reading comprehension in our longitudinal sample in the first instance seem somewhat disappointing. Table 7.9 indicates that the program has a negative effect on letter-naming fluency in the longitudinal sample; there appears to be no evidence for positive effects of the program on the other indicators. These results are somewhat surprising considering the fact that we find positive effects of the program on decoding, language and, in some cases, reading comprehension in our cross-sectional sample.

TABLE 7.9. EFFECTS ON GRADE 2 LONGITUDINAL SAMPLE

Dependent Variable	12-Month Impact (1)	Baseline Mean (2)
Letter-Naming Fluency, Standardized	-0.33**	-0.18
	(-2.07)	
Initial Letter Sound Identification, Standardized	0.03	-0.01
	(0.18)	
Familiar Word Fluency, Standardized	-0.08	-0.16
	(-0.45)	
Oral Vocabulary, Standardized	-0.11	-0.11
	(-0.60)	
Reading Passage Fluency, Standardized	-0.11	-0.16
	(-1.12)	
Listening Comprehension, Standardized	-0.11	-0.05
	(-0.29)	
Reading Passage Comprehension, Standardized	-0.11	-0.12
	(-0.52)	
Dictation, Standardized	-0.09	-0.12
	(-0.52)	
Decoding, Standardized	-0.17	-0.16
	(-1.39)	
Language Comprehension, Standardized	-0.06	-0.06
	(-0.47)	
N	808	194

Notes: Estimations use difference-in-difference modeling. Robust t-statistics clustered at the school level are in parentheses. **=p<0.05

Further analyses indicate that the impact estimates on our longitudinal sample are most likely biased because of selective attrition. We analyzed balance at baseline separately for the longitudinal sample, both including and excluding the students who dropped out from our sample for various reasons. Our results indicate that beneficiary students and control students in our longitudinal sample are roughly equivalent in observable characteristics when we include students who dropped out from our sample in the analysis. These results are presented in Table 7.10. This table shows that there are no statistically significant differences between beneficiary students and control students when we include students who dropped out from our sample in the analyses. However, our findings also demonstrate that when we exclude students who dropped out from our sample, beneficiary students have statistically significantly higher scores than control students in letter-naming fluency, reading comprehension, and dictation. These results, which are depicted in Table 7.11, demonstrate that we have to be very cautious in interpreting the results from our longitudinal sample.

TABLE 7.10. BASELINE BALANCE, LONGITUDINAL SUBSAMPLE INCLUDING STUDENTS WHO DROPPED OUT

Variables, standardized	Control		Treatment		Balance Test			
	Mean	N1	Mean	N2	Diff	SE	p-value	ES
Letter-Naming Fluency	-0.17	292	0.08	281	0.19	0.16	0.23	0.22
Initial Letter Sound Identification	-0.07	292	0.08	281	-0.12	0.22	0.60	-0.12
Familiar Word Fluency	-0.14	292	0.04	281	0.19	0.18	0.31	0.21
Oral Vocabulary	-0.04	292	0.04	281	0.16	0.16	0.33	0.16
Reading Passage Fluency	-0.14	292	-0.03	281	0.12	0.16	0.46	0.17
Listening Comprehension	-0.06	292	0.06	281	0.06	0.23	0.79	0.06
Reading Passage Comprehension	-0.04	292	0.04	281	0.17	0.15	0.25	0.17
Dictation	-0.11	292	0.11	281	0.16	0.16	0.33	0.16
Decoding	-0.15	292	0.03	281	0.17	0.16	0.29	0.23
Language Comp.	-0.06	292	0.06	281	0.03	0.12	0.77	0.05

TABLE 7.11. BASELINE BALANCE, LONGITUDINAL SUBSAMPLE EXCLUDING STUDENTS WHO DROPPED OUT

Variables, standardized	Control		Treatment		Balance Test			
	Mean	N1	Mean	N2	Diff	SE	p-value	ES
Letter-Naming Fluency	-0.18	194	0.11	210	0.29	0.13	0.03	0.34
Initial Letter Sound Identification	-0.01	194	0.15	210	0.09	0.09	0.32	0.11
Familiar Word Fluency	-0.16	194	0.05	210	0.24	0.17	0.18	0.27
Oral Vocabulary	-0.11	194	0.01	210	0.20	0.15	0.21	0.18
Reading Passage Fluency	-0.16	194	-0.03	210	0.17	0.13	0.21	0.24
Listening Comprehension	-0.05	194	0.08	210	0.20	0.27	0.47	0.21

Reading Passage Comprehension	-0.12	194	-0.01	210	0.21	0.12	0.09	0.22
Dictation	-0.12	194	0.14	210	0.27	0.11	0.02	0.27
Decoding	-0.16	194	0.04	210	0.23	0.14	0.11	0.32
Language Comp.	-0.06	194	0.08	210	0.16	0.11	0.14	0.25

The lack of balance at baseline may have resulted in an underestimate of the impact estimates in the difference-in-difference analysis. Possibly, the relatively weak control students were able to catch up with the relatively strong beneficiary students despite the lack of teacher training for their trainers. This catch-up mechanism should not be confused with a negative or no effect of the intervention on the decoding, language, and reading comprehension outcomes on the beneficiary students in our longitudinal sample. Thus, we suggest interpreting the findings on the impact of the program in the longitudinal sample with caution because of the selective attrition. Unfortunately, this selective attrition prevents us from making causal claims about the impact of the program in the longitudinal sample. We will instead rely on the cross-sectional sample for deriving conclusions about the impact of the program.

7.4 IMPACT ON INTERMEDIATE OUTCOMES

This section serves to determine how the program achieves its impact on reading outcomes by measuring the impact of the program on intermediate outcomes for teachers, parents, and students. The results of the significance testing are in Appendix E. Results are presented without control variables, as results with and without control variables differ very little. A complete set of results tables can be made available upon request.

TEACHER PRACTICES

The theory of change suggests that the program can influence reading outcomes through positive effects on teacher practices. The teacher training program aims to improve the pedagogical skills of teachers by training teachers to teach reading according to the five skills of reading. There is also ongoing school-based mentoring support. We measure the outcome of these project inputs through Indicator 4 of the project's performance M&E plan. We construct a two-part measure of teacher practices. Teachers must be observed doing at least 7 of 10 essential best practices to pass part 1. Using the data from classroom observations, as well as teacher interviews, we construct an index that reflects teaching best practices. Teachers must score above a certain threshold to pass part 2. (For more information on the index, please see *USAID Quality Reading Project Performance Monitoring and Evaluation Annual Reports for Indicator 4*.)

We estimate the impact of the program on the following outcomes indicators:

1. An overall binary measure based on whether the teacher passed both parts 1 and 2
2. The number of essential practices the teacher demonstrates, on a scale of 0–10
3. The index value for part 2
4. Each of the essential practices separately, as a binary outcome

Although the theory of change indicates that the program could positively influence teacher practices, we find little evidence for positive effects. We find only small positive effects that are marginally statistically significant. Full results can be found in Appendix E. We see small differences between treatment and control teachers, but very few are statistically significant. The one significant outcome measurement is the aggregate number of essential practices observed, although none of the individual practices shows significance. The program increases teachers' overall usage of essential best practices by 5%, which translates to half of an essential practice.

We identified four possible reasons for the lack of positive effects on teacher practices. First, our teacher sample size may be too small to detect small but meaningful effects of the program on teacher practices. Our power calculations suggest that our sample size is sufficient to detect small but meaningful effects on students. However, our teacher sample size is smaller (4 teachers per school compared to 40 students), which may result in imprecise impact estimates. Second, our measures of teacher behavior may not include the right measures of teacher practices, as there is limited consensus as to what constitutes good teacher practice measures. Thus, it is possible our measures do not include the right measures of teacher practices. Third, our measures of teacher practices may be too noisy to detect small but meaningful effects of the program on these indicators. Measuring teacher practices is challenging and, as a result, standard deviations may be high. Fourth, the program may lead to program impacts through different channels such as increased access to reading materials, increased activities around reading, and increased emphasis on reading in the communities where the beneficiary children reside.

PARENTAL ATTITUDES TO READING

We do not find evidence for an impact of the program on parental attitudes towards reading. Parental attitudes are measured through an index created from their responses to a survey. (For more information on the index, please see *USAID Quality Reading Project Performance Monitoring and Evaluation Annual Reports for Indicator 15*.) When scores are standardized, and treatment and control scores are statistically compared using a two-sample t-test, we see evidence of a program impact. Parents of students who attend treatment schools have a 0.01 standard deviation increase in the index that measure attitudes. However, when scores are clustered in schools and tested using ordinary least squares regression, the impact is no longer significant. Parents in program communities/schools still score higher on the index than parents in control communities/schools. The reasons for no strong evidence of an impact are similar to those explained above in the section on teachers: limited sample size during the midterm makes it harder to detect an impact, measuring attitude change is difficult, and there is a slight ceiling effect with this index.

STUDENT READING BEHAVIOR

We also see impacts on student behavior, although only for at-home reading activity and not participating in reading activities. Students in treatment schools self-report a statistically higher level of reading non-textbooks at home, compared to control schools. The program

causes a 0.04 percentage point increase in a child's likelihood of reading at home. Likely, we were unable to detect a program impact on participating in reading activities because the data were collected in April 2015, which was before the intervention out-of-school reading activities started, as noted in the intervention description in this report.

7.5 LANGUAGE OF INSTRUCTION AND HOME LANGUAGE: ASSOCIATIONS WITH READING OUTCOMES

Several variables, such as home print environment (Senechal & Lefevre, 2002), community print environment (Raz & Bryant, 1990), and various cognitive and linguistic factors (e.g., Snow, Burns, & Griffin, 1990) are known to predict reading outcomes. Another critical predictor of reading success is the LOI (Alidou et al., 2006). If a child goes to school and learns in a language they do not speak and understand well, they are unlikely to be able to effectively acquire literacy skills in that language.

Almost 30% of the participants in this study had a mismatch between home language and the LOI in the school. We examined the relationship between this mismatch and various reading outcomes. We hypothesized that language and comprehension reading subskills are more susceptible to a language mismatch than sound- (phonology-) or script- (orthography-) dependent reading subskills. In other words, if a child is in a Kyrgyz-LOI school but does not speak Kyrgyz at home, or if a child is in a Russian-LOI school but does not speak Russian at home, he or she is more likely to lag behind his or her matched-language counterparts in most reading scores, but especially in the comprehension scores.

We find some evidence that the hypotheses about language mismatches are correct. As predicted, the results showed that the language mismatch group did significantly worse than the language match group on oral vocabulary knowledge $t\ (1969)=-13.42$, $p<.001$; reading comprehension $t\ (1969)=-9.47$, $p<.001$; familiar word decoding $t\ (1969)=-3.0$, $p<.01$; oral reading fluency $t\ (1969)=-5.0$, $p<.001$; and letter name knowledge $t\ (1969)=-7.77$, $p<.001$. The mismatch group also had slightly, but significantly lower scores on nonsense words $t\ (1969)=-2.13$, $p<.05$. There were no significant differences between the two groups on initial letter sound scores (phonological awareness) and dictation.

The significant differences on all the readings skills clearly indicate that children who are not learning to read in the language of the home are likely to fall behind their peers whose LOI is the same as the language of the home. This is line in with most research on LOI and learning to read globally (e.g., Alidou et al., 2006). The lack of a significant difference in initial letter sound scores is also predictable and interesting. Research on biliteracy acquisition shows that phonological awareness is significantly correlated in both languages (August & Shanahan, 2006) and is often conceptualized as a "sharable" and "transferable" skill (i.e., tapping into the same general cognitive construct in both languages) (Koda & Reddy, 2008). Therefore, LOI may not matter as much for this skill, but its influence on reading outcomes will manifest in the other skills. The slightly smaller gap between LOI match and mismatch groups on nonsense words is also in line with the research in that nonsense word reading is reliant on the ability to apply sound–symbol correspondence rules, and in transparent

languages like Kyrgyz and Russian, this skill may not draw heavily from oral language resources. The finding that the two groups performed similarly on dictation might suggest that skills other than language, including working memory and application of sound–symbol rules, are at play here as well.

In summary, the importance of LOI match to child's home language is a very important significant predictor of reading outcomes, especially for comprehension-based reading outcomes. Children who go to school with LOIs that are different from their home language are being set up for a host of reading difficulties in the future.

8. LIMITATIONS

One of the major limitations of this midterm impact report is the reduced sample size in cohort 1 regions only. The study is adequately powered to detect a small but relevant effect of approximately 0.16–0.2 standard deviations at a minimum with a sample size of 130 schools. These effect sizes are well within the normal accepted range for education interventions. Because the midterm sample is a subsample of 60 schools, we have a reduced ability to detect positive effects of the program. Note that this does not mean that there aren't effects, but merely that we lack that statistical power to detect them. Therefore, any time we are reporting that we cannot detect an effect on a certain outcome (such as on teacher behavior or reading comprehension), this does not necessarily mean that there is no effect.

The findings should also be cautiously interpreted because of imperfections in the instruments used for data collection. For example, the lack of impact on teacher instructional practices may be caused by a result of the limitations of the classroom observation and teacher interview forms. Proper measurement of "good teaching" is a hotly contested issue in the education research community globally, so any designed measure is sure to have many limitations.

Third, contamination may influence the results to a certain extent. Schools/communities that are not receiving the intervention may still benefit from neighboring schools. For example, at district education meetings, intervention school directors share learning experiences with control school directors. Community reading events are also open to all, so students from control schools are free to attend.

Fourth, imperfect take-up of the intervention may have resulted in lower impact estimates. Some teachers in schools meant to receive the intervention may not have participated in the training. The number of schools in that situation is very low. We do not have reason to believe that there are high levels of contamination, but the effect is still possible. Because contamination moves in both directions (where treatment schools act like controls or control schools are exposed to treatment), the effect on impact estimates could go in either direction.

9. CONCLUSION

The major findings of our midterm impact study allow us to conclude the following four major points. We make accompanying recommendations below.

1. **The program is having an impact on student reading outcomes.**

The USAID Quality Reading Project is clearly positively impacting student reading outcomes. Because of our research design, it is hard to isolate the exact elements of the program that are having the biggest impact. However, we do have adequate evidence to recommend a continuation of the existing model of the program. This report provides evidence of the one-year midterm impacts of the program on schools in the first regions where the intervention was rolled out.

2. **However, the results are on decoding and early reading skills, not on comprehension.**

Given the context of the intervention and the linguistic circumstances of Kyrgyz and Russian, a possible explanation for these results follows. The Cyrillic script is transparent (i.e., one-to-one correspondence between sounds and letters). This makes acquisition of decoding skills relatively quick and easy, compared to more opaque orthographies, such as English, where oral reading fluency may take longer to acquire, and may need more phonological and decoding-focused pedagogical interventions. Therefore, if a reading program focuses on phonological skills, decoding, and oral reading fluency for several years at the cost of comprehension-related pedagogical interventions, it is likely that decoding-related skills will be enhanced more than comprehension-related reading subskills. In addition, building comprehension abilities requires intense time and focus on several aspects of vocabulary (e.g., usage in different contexts, morphological awareness), grammar, inference making, story prediction, background knowledge, and other skills. Without time committed to these varied skills around understanding text, comprehension skills may suffer. The focus of the intervention was on these early reading skills. Furthermore, the traditional practices of reading instruction in the Soviet context relate to reading fluency and speed, so teachers are likely more comfortable focusing reading efforts on improving those skills.

To address this finding, we have the internal recommendation to shift some of the focus of the mentoring and any potential follow-up training to more comprehension activities than simply decoding or decoding-related activities. We propose working with teachers to show them the data and findings on reading fluency and comprehension, which are sometimes more weakly correlated in our data as one might expect, given the transparency of the scripts. This is perhaps the result of an overemphasis on reading fluency without attention to comprehension. Note that this recommendation (and others) is a hypothesis and how to best address this finding and has not been tested.

3. **The program is impacting more reading subskills of those whose scores are lower: boys are improving in more reading skills than girls, students who study in Kyrgyz over Russian, rural students over urban in Grade 2.**

The program's results are helping to improve the subgroups with lower scores. This phenomenon is more likely because there is potentially more room for improvement with those with lower skill levels. However, it is still an important finding and supports the program's work tremendously. The gender gap in performance (see *EGRA Midterm Report for Kyrgyz Republic*) between boys and girls is significant, with girls strongly outperforming boys in almost all subtasks. Our impact findings suggest that teachers need to demonstrate adapted instruction to take into account the specific needs of students in the classroom. We have no additional recommendations relating to this finding.

4. **We do not find evidence for impacts on intermediate outcomes in teachers, parents, or students.**

Although our lack of evidence for findings for positive impacts on the outcome pathways intermediate outcomes in the theory of change may be associated with imperfect measurement tools or lack of statistical power, is not evidence that there are no impacts, but suggests we may have limitations with our data or sample size. However, it does raise the issue that it remains important to understand what is going on with teachers in the classroom. We recommend follow-up qualitative research with teachers, and the collection of more in-depth longitudinal data on classroom observation and reading instruction. Also, we propose disseminating the midterm impact results to teachers through data fact sheets. Finally, we propose improving the classroom observation, teacher, parent, and student survey instruments in an effort to better understand the pathways through which the program is having an impact. This may help improve our understanding on a lack of evidence for positive impacts on intermediate outcomes.

5. **For children to successfully gain text comprehension skills, they must be taught literacy in the languages of their home (first).**

Although only correlational, our findings are in line with many other studies (e.g., Alidou et al., 2006; Shanahan & August, 2006) that show the importance of learning to read in a language you understand and speak well (most often the language of the home). Although certain decoding related scores, perhaps oral reading fluency, may develop without much oral language proficiency, comprehension is likely to remain challenging to acquire.

WORKS CITED

Alidou, H., Boly, A., Brock-Utne, B., Diallo, Y. S., Heugh, K., & Wolf, H. E. (Eds.). (2006). Optimizing learning and education in Africa—the language factor: A stock-taking research on mother tongue and bilingual education in Sub-Saharan Africa. Paris, France: Association for the Development of Education in Africa.

August, D., & Shanahan, T. (Eds.). (2006). Developing literacy in second-language learners: Report of the National Literacy Panel on language-minority children and youth. Washington, DC: Center for Applied Linguistics.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Hopewell, NJ: Houghton Mifflin.

Conn, K. (2014). *Identifying effective education interventions in Sub-Saharan Africa: A meta-analysis of rigorous impact evaluations*. Columbia University.

Connor, C. M., Morrison, F. J., & Slominski, L. (2006). Preschool instruction and children's emergent literacy growth. *Journal of Educational Psychology*, *98*(4), 665–689.

Duflo, E., Dupas, P., & Kremer, M. (2015). School governance, teacher incentives, and pupil–teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics*, *123*, 92–110.

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. The American Economic Review, 1241-1278.

Evans, D., & Popova, A. (2015). What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. An Analysis of Divergent Findings in Systematic Reviews (February 26, 2015). World Bank Policy Research Working Paper, (7203).

Fitzgerald, J., & Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist, 35,* 39-50.

Friedman, W., Gerard, F., & Ralaingita, W. (2010). *International independent evaluation of the effectiveness of Institut pour l'Education Populaire's "Read-Learn-Lead" (RLL) program in Mali, mid-term report.* Research Triangle Park, NC: RTI International.

He, F., Linden, L. L., & MacLeod, M. (2009). A better way to teach children to read? Evidence from a randomized controlled trial. Unpublished manuscript. New York, NY: Columbia University.

Hirsch, E. D. (2003). Reading comprehension requires knowledge—of words and the world. American Educator, 27(1), 10-13.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing (2), 127–160.*

Joshi, R. M., & Aaron, P.G. (2000). The component model of reading: Simple view of reading made a little more complex. *Reading Psychology, 21,* 85–97.

Koda, Keiko, and Pooja Reddy. "Cross-linguistic transfer in second language reading." *Language Teaching* 41.04 (2008): 497-508.

Kremer, M., Brannen, C., & Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science*, *340*(6130), 297–300.

Linden, L. L., & MacLeod, M. (n.d.). *How to teach English in India: Testing the relative productivity of instruction methods within the Pratham English Language Education Program.*

McEwan, P. J. (2015). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research*, *85*(3), 353–394.

Muralidharan, K., & Sundararaman, V. (2009). *Teacher performance pay: Experimental evidence from India*. National Bureau of Economic Research.

Murnane, R. J., & Ganimian, A. J. (2014). *Improving educational outcomes in developing countries: Lessons from rigorous evaluations*. National Bureau of Economic Research.

National Institute of Child Health and Human Development (NICHD) (2006. *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction. Report of the National Reading Panel.* Washington, DC: U.S. Department of Health and Human Services.

Piper, B. (2009). *Impact study of SMRS using Early Grade Reading Assessment in three provinces in South Africa*. Research Triangle Park, NC: Integrated Education Programme, RTI International.

Piper, B., & Korda, M. (2010). *Early Grade Reading Assessment (EGRA) Plus: Liberia*. *Program evaluation report prepared for USAID/Liberia*. Research Triangle Park, NC: RTI International.

Raz, I. S., & Bryant, P. (1990). Social background, phonological awareness, and children's reading. *British Journal of Developmental psychology, 8,* 209-225.

RTI International (2014). *Girls' improved learning outcomes: Final report. Program evaluation report prepared for USAID/Egypt*. Research Triangle Park, NC: RTI International.

Senechal, M., & LeFevre J. A. (2002). Parental involvement in the development of children's reading skill: A five year longitudinal study. Child Development, 73(2), 445–460.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). Preventing reading difficulties in young children. Washington, DC: National Academy Press.

APPENDIX A: BALANCE TABLES

TABLE A.1. GRADE 2, FULL 130-SCHOOL, BASELINE BALANCE

Variables	Control		Treatment		Mean	Diff	
	Mean	N1	Mean	N2	Diff	SE	p-value
Letter-Naming Fluency	59.562	1,130	59.192	1,099	-0.370	2.068	0.858
Initial Letter Sound Identification	90.150	1,130	90.055	1,099	-0.096	1.107	0.931
Familiar Word Fluency	47.950	1,130	51.065	1,099	3.115	2.166	0.153
Unfamiliar Word, 2014 Only	24.139	1,130	25.062	1,099	0.923	1.074	0.392
Oral Vocabulary	91.310	1,130	91.611	1,099	0.301	1.156	0.795
Reading Passage Fluency	32.807	1,130	34.802	1,099	1.995	1.639	0.226
Listening Comprehension	69.292	1,130	74.204	1,099	4.912	2.414	**0.044****
Reading Passage Comprehension	46.711	1,130	51.153	1,099	4.442	3.128	0.158
Dictation	58.186	1,130	58.821	1,099	0.635	2.508	0.800

Notes: Cluster at the school level. **=p<0.05

TABLE A.2. GRADE 4, FULL 130-SCHOOL, BASELINE BALANCE

Variables	Control		Treatment		Mean	Diff	
	Mean	N1	Mean	N2	Diff	SE	p-value
Familiar Word Fluency	67.234	1,160	70.404	1,108	3.170	2.157	0.144
Unfamiliar Word, 2014 Only	31.820	1,160	33.289	1,108	1.469	1.253	0.243
Oral Vocabulary	95.034	1,160	96.110	1,108	1.076	0.713	0.134
Reading Passage Fluency	65.415	1,160	68.920	1,108	3.505	2.186	0.111
Listening Comprehension	73.263	1,160	74.188	1,108	0.925	2.393	0.700
Reading Passage Comprehension	65.810	1,160	67.473	1,108	1.663	2.482	0.504
Dictation	80.152	1,160	81.347	1,108	1.196	1.301	0.360

Notes: Cluster at the school level.

TABLE A.3. GRADE 2, 60-SCHOOL MIDTERM SUBSAMPLE, BASELINE BALANCE

Variables	Control		Treatment		Mean	Diff	
	Mean	N1	Mean	N2	Diff	SE	p-value
Letter-Naming Fluency	62.910	482	61.378	491	-1.532	3.103	0.624
Initial Letter Sound Identification	93.361	482	92.994	491	-0.367	1.189	0.759
Familiar Word Fluency	52.106	482	50.166	491	-1.940	2.894	0.505
Unfamiliar Word, 2014 Only	26.737	482	25.311	491	-1.426	1.305	0.279
Oral Vocabulary	93.402	482	92.831	491	-0.572	1.073	0.596
Reading Passage Fluency	36.318	482	34.664	491	-1.654	2.023	0.417
Listening Comprehension	74.274	482	76.986	491	2.712	3.254	0.408
Reading Passage Comprehension	51.089	482	53.462	491	2.373	4.158	0.570
Dictation	66.805	482	63.431	491	-3.374	3.196	0.296

Notes: Cluster at the school level.

TABLE A.4. GRADE 4, 60-SCHOOL MIDTERM SUBSAMPLE, BASELINE BALANCE

Variables	Control		Treatment		Mean	Diff	
	Mean	N1	Mean	N2	Diff	SE	p-value
Familiar Word Fluency	70.822	520	70.510	502	-0.312	3.172	0.922
Unfamiliar Word, 2014 Only	34.422	520	35.872	502	1.450	1.881	0.444
Oral Vocabulary	95.135	520	96.614	502	1.479	0.958	0.128
Reading Passage Fluency	67.988	520	71.009	502	3.021	3.116	0.336
Listening Comprehension	77.952	520	80.159	502	2.207	3.158	0.487
Reading Passage Comprehension	70.317	520	74.781	502	4.464	3.088	0.154
Dictation	81.749	520	84.098	502	2.349	1.674	0.166

Notes: Cluster at the school level.

APPENDIX B: EGRA SUBTASK CORRELATIONS

TABLE B.1. CORRELATION MATRIX OF KYRGYZ GRADE 2 EGRA SUBTASKS

Kyrgyz Grade 2		Initial Letter Sound	Oral Vocabulary	Reading Comprehension	Listening Comprehension	Dictation	Letter Name	Nonsense Words	Familiar Words	Reading Passage
Correlation	Initial Letter Sound	1.000	.255	.245	.186	.241	.264	.213	.208	.223
	Oral Vocabulary	.255	1.000	.249	.187	.242	.139	.030	.065	.070
	Reading Comprehension	.245	.249	1.000	.464	.437	.299	.441	.343	.450
	Listening Comprehension	.186	.187	.464	1.000	.289	.110	.209	.138	.204
	Dictation	.241	.242	.437	.289	1.000	.508	.553	.573	.537
	Letter Name	.264	.139	.299	.110	.508	1.000	.637	.646	.579
	Nonsense Words	.213	.030	.441	.209	.553	.637	1.000	.847	.918
	Familiar Words	.208	.065	.343	.138	.573	.646	.847	1.000	.886
	Reading Passage	.223	.070	.450	.204	.537	.579	.918	.886	1.000

TABLE B.2. CORRELATION MATRIX OF RUSSIAN GRADE 4 EGRA SUBTASKS

Russian Grade 4		Oral Vocabulary	Reading Comprehension	Listening Comprehension	Dictation	Nonsense Words	Familiar Words	Reading Passage
Correlation	Oral Vocabulary	1.000	.544	.484	.404	.253	.231	.331
	Reading Comprehension	.544	1.000	.589	.471	.353	.411	.419
	Listening Comprehension	.484	.589	1.000	.353	.235	.228	.257
	Dictation	.404	.471	.353	1.000	.415	.511	.430
	Nonsense Words	.253	.353	.235	.415	1.000	.684	.697
	Familiar Words	.231	.411	.228	.511	.684	1.000	.601
	Reading Passage	.331	.419	.257	.430	.697	.601	1.000

TABLE B.3. CORRELATION MATRIX OF KYRGYZ GRADE 4 EGRA SUBTASKS

Kyrgyz Grade 4		Oral Vocabulary	Reading Comprehension	Listening Comprehension	Dictation	Nonsense Words	Familiar Words	Reading Passage
Correlation	Oral Vocabulary	1.000	.336	.334	.267	.168	.172	.148
	Reading Comprehension	.336	1.000	.521	.569	.459	.440	.390
	Listening Comprehension	.334	.521	1.000	.342	.274	.256	.310
	Dictation	.267	.569	.342	1.000	.577	.584	.517
	Nonsense Words	.168	.459	.274	.577	1.000	.838	.821
	Familiar Words	.172	.440	.256	.584	.838	1.000	.749
	Reading Passage	.148	.390	.310	.517	.821	.749	1.000

APPENDIX C: EGRA FACTOR ANALYSIS

Based on the National Reading Panel report *Teaching Children to Read* (National Institute of Child Health and Human Development, 2000) and other research, most Early Grade Reading Assessments (EGRAs) measure five reading skills: phonics, phonemic awareness, vocabulary, reading comprehension, and fluency. Evidence from research on early grade reading suggests there are multiple dimensions underlying the latent construct of *reading ability* (i.e., the ability to comprehend a passage or more of written text). The Simple View of Reading (e.g., Hoover & Gough, 1990; Joshi & Aaron, 2000) holds that reading comprehension is a result of two main skills: language comprehension and decoding (and their corresponding subskills). According to the Simple View of Reading, decoding ability is the ability to *sound out* written letters, words, and texts, while language comprehension is described as the ability to understand spoken language. Each of these constructs may have subconstructs such as phonological skills, vocabulary, and background knowledge. The relative importance of these constructs varies by language type and level of reading ability. As such, research suggests that phonological skills, decoding skills (letters, words, and texts), and language comprehension skills are all possible underlying dimensions of reading ability (Good & Jefferson 1998; Kaminski & Good, 1998; NICHD, 2000).

The purpose of factor analysis is to identify the number and nature of these dimensions (or factors) that reading assessments such as the EGRA are tapping into. In this study, EGRA subtask scores at each grade level and language were analyzed to determine the underlying dimensions of the data using the SPSS "data reduction" function. Subscores were analyzed using *principal axis factor analysis* to test hypotheses about the number of factors that could be extracted. Based on the stated purpose(s) of the subtasks, a 3-factor model for Grade 2 and a 2-factor model for Grade 4 were hypothesized initially, based on theoretical propositions about what skills the various subtasks assessed. The results of the Grade 2 analyses indicated that a 2-factor model was a better fit for the data structure for both language groups. A 2-factor model for both Grade 4 languages was confirmed by the analysis.

In line with the Simple View of Reading and other research on the underlying constructs of reading ability, the two EGRA dimensions identified in this study were decoding (letters, words, and text) and language comprehension. Because research from other EGRAs has indicated that the subtasks were likely to be correlated, oblique rotation was used to aid in interpretation of results. The results indicate that the two factors identified were indeed correlated on each of the 4 versions (K2=.38, R2=.36, K4=.44, and R4=.48), but not correlated highly enough to indicate the presence of just a single factor. The resulting pattern matrices allowed interpretation of the overall simple structure of the data by examining how factors clustered on the matrix.[6] High factor loadings indicate which subtasks are tapping into which common dimensions. Interpretation of results by each of the grade level EGRAs are presented below.

[6] See tables for pattern matrix with factor loadings for both grade levels and languages.

The factor analysis results indicated that familiar word, nonsense word, reading passage fluency, and letter-naming subtasks all loaded onto one construct that represents *decoding* skills. The decoding construct served as a dependent variable in our analyses for the impact study. The outcome measure is an average of the standardized values of the subtasks that loaded on this single construct. Listening comprehension, oral vocabulary, and initial letter sound all loaded on the factor *language comprehension*. Similarly, the language comprehension construct employed in the impact study is the average of the standardized values of the previously mentioned three subtasks. In line with existing literature that writing requires similar but distinct skills from reading (Fitzgerald & Shanahan, 2000), dictation did not clearly load with these existing factors, so as a subtask, it has been left separate. Finally, reading comprehension was treated as a separate outcome variable, as it is the end goal of the other reading skills.

Decoding, language comprehension, dictation, and reading comprehension were the main constructs employed as outcome measures in the impact study. Impact was measured for each subtask separately, as well, but those measurements have a higher likelihood of type II errors, with are false positives, because of the number of statistical tests run for each subgroup. Therefore, we put a greater emphasis on the combined measurements in the interpretation of our results. The paragraphs that follow present the individual factor analyses for each grade and language.

KYRGYZ GRADE 2

High factor loadings on the *Nonsense Words, Reading Passage, Familiar Words, and Letter Name* subtasks represent the construct of decoding skills, or factor 1 (letter, word, and text). Decoding can be understood as the ability to map sounds with their respective letters (or letter clusters) in order to be able to *sound out* written information (letters, words, or passages). The high loadings by these subtasks indicate that each of them tap into this latent construct. Note that the *Dictation* subtask tapped into both factor 1 (decoding) and factor 2 (comprehension) but not strongly on either. The *Oral Vocabulary* and *Listening Comprehension* tasks represent the construct of language comprehension (factor 2), though *Listening Comprehension* for Kyrgyz Grade 2 loads weakly at .350. Language comprehension is understood as the ability to understand oral language information without having to rely on decoding. In other words, a person who is completely illiterate, but has complete knowledge of his or her spoken language, has high language comprehension ability (and low or no decoding ability).

TABLE C.1. KYRGYZ GRADE 2 FACTOR ANALYSIS

Pattern Matrix[a]		
	Factor	
	1	2
Familiar Words	.973	
Nonsense Words	.952	
Reading Passage	.941	
Letter Name	.645	
Dictation	.474	.354
Reading Comprehension		.620
Listening Comprehension		.576
Oral Vocabulary		.464
Initial Letter Sound		.344

Notes: Extraction Method—Principal Axis Factoring.
Rotation Method—Oblimin with Kaiser Normalization.
[a] Rotation converged in four iterations.

RUSSIAN GRADE 2

According to factor loadings on the subtasks, the prediction of a 2-factor model was also borne out by the Russian Grade 2 data. We see clustering and high loadings of *Nonsense Words, Familiar Words*, and *Reading Passage*, which are the constituent parts of factor 1, decoding. The subtask *Letter Name* did not load on factor 1 or factor 2 for the Russian group as it did for the Kyrgyz group (not above .3, which was the minimum criteria selected for reporting loadings). *Oral Vocabulary* and *Listening Comprehension* both loaded highly on a single factor (factor 2, comprehension), as predicted, considering the nature of the subtasks. The Subtask, *Dictation*, loaded on factor 1, decoding, though weakly.

TABLE C.2. RUSSIAN GRADE 2 FACTOR ANALYSIS

Pattern Matrix[a]		
	Factor	
	1	2
Nonsense Words	.944	
Familiar Words	.940	
Reading Passage	.928	
Dictation	.376	.343
Letter Name	.267	.200
Initial Letter Sound	.208	
Listening Comprehension		.791
Oral Vocabulary		.780
Passage Comprehension	.130	.712

Notes: Extraction Method—Principal Axis Factoring.
Rotation Method—Oblimin with Kaiser Normalization.
[a] Rotation converged in five iterations.

KYRGYZ GRADE 4

A 2-factor model was predicted considering the subtasks employed for EGRA Grade 4. The results for Kyrgyz Grade 4 indeed indicated a 2-factor solution with clear clusters of factor loadings on decoding (factor 1) and language comprehension (factor 2). The subtask, *Dictation,* loaded on factor 1, though at a lower level than the other three subtasks.

TABLE C.3. KYRGYZ GRADE 4 FACTOR ANALYSIS

Pattern Matrix[a]		
	Factor	
	1	2
Nonsense Words	.972	
Familiar Words	.890	
Reading Passage	.850	
Dictation	.440	.389
Reading Comprehension	.136	.737
Listening Comprehension		.651
Oral Vocabulary		.491

Notes: Extraction Method—Principal Axis Factoring.
Rotation Method—Oblimin with Kaiser Normalization.
[a] Rotation converged in four iterations.

RUSSIAN GRADE 4

The results for Russian Grade 4 indicate a 2-factor solution with distinct factor loadings on decoding and language comprehension. Note that unlike for Kyrgyz Grade 4, the *Dictation* task loaded on both tasks, but at just under 4, a relatively low level.

TABLE C.4. RUSSIAN GRADE 4 FACTOR ANALYSIS

Pattern Matrix[a]		
	Component	
	1	2
Nonsense Words	.918	
Familiar Words	.883	
Reading Passage	.831	
Dictation	.457	.401
Listening Comprehension		.870
Oral Vocabulary		.818
Reading Comprehension		.772

Notes: Extraction Method—Principal Component Analysis.
Rotation Method—Oblimin with Kaiser Normalization.
[a] Rotation converged in six iterations.

APPENDIX D: READING OUTCOME REGRESSION TABLES

TABLE D.1. READING OUTCOME SUBSKILLS, GRADE 2

Dependent Variable	12-Month Impact (1)	Baseline Mean (2)
Letter-Naming Fluency, Standardized	0.23*	0.10
	(1.98)	
Initial Letter Sound Identification, Standardized	0.29**	0.25
	(2.42)	
Familiar Word Fluency, Standardized	0.32***	0.08
	(2.96)	
Oral Vocabulary, Standardized	0.19	0.17
	(1.21)	
Reading Passage Fluency, Standardized	0.34***	0.05
	(3.01)	
Listening Comprehension, Standardized	0.23	0.12
	(1.03)	
Reading Passage Comprehension, Standardized	0.29	0.11
	(1.25)	
Dictation, Standardized	0.40**	0.26
	(2.42)	
N	1,957	482

Notes: Estimations use difference-in-difference modeling. Robust t-statistics clustered at the school level are in parentheses. *=p<0.10, **=p<0.05, ***=p<0.01

TABLE D.2. READING OUTCOME SUBSKILLS, GRADE 4

Dependent Variable	12-Month Impact (1)	Baseline Mean (2)
Familiar Word Fluency, Standardized	0.35**	0.06
	(2.16)	
Oral Vocabulary, Standardized	0.04	0.02
	(0.22)	
Reading Passage Fluency, Standardized	0.20*	-0.01
	(1.74)	
Listening Comprehension, Standardized	-0.12	0.11
	(-0.56)	
Reading Passage Comprehension, Standardized	0.09	0.15
	(0.74)	
Dictation, Standardized	0.01	0.01
	(0.12)	
N	2,009	520

Notes: Estimations use difference-in-difference modeling. Robust t-statistics clustered at the school level are in parentheses. *=p<0.10, **=p<0.05

APPENDIX E: OTHER OUTCOME REGRESSION TABLES

TABLE E.1. IMPACT ON TEACHING PRACTICES

Dependent Variable	12-Month Impact (1)	Baseline Mean (2)
Part 1: 10 Essential Practices		
1. Teacher has textbook	-0.01	0.91
	(-0.23)	
2. Classroom walls have written materials	0.08	0.95
	(1.19)	
3. Classroom wall materials are grade-level appropriate	0.01	0.54
	(0.09)	
4. Reading comprehension pedagogy	-0.01	0.23
	(-0.09)	
5. Active, student-centered teaching	0.09	0.60
	(0.88)	
6. Teacher behavior supports student-centered classroom	-0.07	0.02
	(-1.14)	
7. Formative assessment	0.05	0.91
	(0.89)	
8. Teacher has lesson plan	0.01	0.86
	(0.15)	
9. Classroom has supplementary materials	0.06	0.70
	(0.73)	
10. Written notes on formative assessment	0.13	0.38
	(1.19)	
Number of 10 essential practices observed, Part 1	0.52*	6.10
	(1.76)	
Score for Part 2	4.91	136.43
	(0.53)	
0/1 for demonstrates teaching best practices: 7+ Part 1, 150+ Part 2	0.06	0.27
	(0.72)	
N	449	113

Notes: Estimations use difference-in-difference modeling. Robust t-statistics clustered at the school level are in parentheses. *=p<0.10

TABLE E.2. IMPACT ON PARENTAL ATTITUDES TO READING

	Treatment			Control			
	Mean	Standard Error	*n*	Mean	Standard Error	*n*	t-statistic
Parental Attitude to Reading in 2015, Standardized Z-Score	-0.05	0.10	293	0.08	0.10	281	0.95

Note: Robust t-statistics clustered at the school level.

TABLE E.3. IMPACT ON STUDENT READING AT HOME

	Treatment			Control			
	Mean	Standard Error	*n*	Mean	Standard Error	*n*	t-statistic
Proportion of Students Reading Nontextbooks at Home	0.92	0.02	981	0.88	0.02	970	1.93*

Notes: Robust t-statistics clustered at the school level. *=p<0.10

TABLE E.4. IMPACT ON STUDENT PARTICIPATION IN READING ACTIVITIES

	Treatment			Control			
	Mean	Standard Error	*n*	Mean	Standard Error	*n*	t-statistic
Proportion of Students Participating in Out-of-School Reading Activities	0.22	0.03	977	0.22	0.03	956	0.09

Note: Robust t-statistics clustered at the school level.

U.S. Agency for International Development
1300 Pennsylvania Avenue, NW
Washington, DC 20523
Tel: (202) 712-0000
Fax: (202) 216-3524
www.usaid.gov