



USAID
FROM THE AMERICAN PEOPLE

IMPACT EVALUATION OF THE EARLY GRADE READING PROGRAM (EGRP) IN NEPAL

MIDLINE EVALUATION REPORT

July 2018

This publication was produced at the request of the United States Agency for International Development. It was prepared independently by the National Opinion Research Center at the University of Chicago.

IMPACT EVALUATION OF THE EARLY GRADE READING PROGRAM (EGRP) IN NEPAL MIDLINE EVALUATION REPORT

July 2018

Prepared for the United States Agency for International Development

Under GS-10F-0033M/AID-OAA-M-13-00013

This publication was produced at the request of the United States Agency for International Development (USAID). It was prepared independently by Alicia Menendez and Gregory Haugan with the support of Varuni Dayaratna.

The views expressed in this publication do not necessarily reflect the views of the United States Agency for International Development or the United States Government.

TABLE OF CONTENTS

EXECUTIVE SUMMARY	VI
PROJECT BACKGROUND	VI
EVALUATION METHODOLOGY OVERVIEW	VI
ANSWERING THE EVALUATION QUESTIONS	VII
CONCLUSIONS	VIII
RECOMMENDATIONS	ix
I. INTRODUCTION	I
I.1. CONTEXT AND PROJECT BACKGROUND	I
2. EVALUATION QUESTIONS AND METHODOLOGY OVERVIEW	3
2.1 EVALUATION QUESTIONS	3
2.2 METHODOLOGY OVERVIEW	5
2.2.1. Treatment Assignment.....	5
2.2.2. Analytical Approach.....	6
2.2.3. Data Collection.....	6
2.2.4. Instrument Creation and Piloting.....	8
2.2.5. Assessor Training.....	9
2.2.6. Fielding the Survey and Quality Assurance	10
2.2.7. Data generation, cleaning, and finalization.....	11
3. FINDINGS	11
3.1 SOME LEARNER CHARACTERISTICS	11
3.2 EGRP IMPACT ON EGRA SCORES	12
3.2.1. Cohort 1.....	13
3.2.2. Cohort 2.....	22
3.3. MECHANISMS BEHIND THE IMPACT OF EGRP	24

3.3.1. Classroom Reading Materials	25
3.3.2. Teacher Training, Support, and Instructional Practices	27
3.3.3. Parental and Community engagement.....	34
4. CONCLUSIONS	38
5. LIMITATIONS	39
6. RECOMMENDATIONS	40
7. REFERENCES	41
8. ANNEXES	42
ANNEX I: EVALUATION STATEMENT OF WORK	43
ANNEX II: EVALUATION METHODS	53
ANNEX III: MATCHING PROCEDURE	55
ANNEX IV: ADDITIONAL ANALYSES	69
ANNEX V: SAMPLE	78

TABLE OF FIGURES

Figure 1: EGRP Theory of Change	2
Figure 2: Location of Sample Schools by Treatment Group	5
Figure 3: Age Distribution by Grade, Baseline and Midline.....	12
Figure 4: Percentage of zero score in Oral Reading Fluency, cohort I, by grade and learner language	16
Figure 5: Percentage of zero scores in Reading Comprehension, Cohort I, by grade and learner language	17
Figure 6: Percentage of learners reaching Minimum Fluency Threshold (45 CWPM), cohort I, by grade and learner language.....	18
Figure 7: Percentage of learners reaching Minimum Reading Comprehension Threshold (80%), cohort I, by grade and learner language	19
Figure 8: Mean Oral Reading Scores at Midline, cohort I Treatment, by Gender and Grade.....	20

Figure 9: EGRP Effect on Oral Reading Scores at Midline, cohort 1, by Gender and Grade.....	21
Figure 10: Percentage of learners reaching Minimum Fluency Threshold (45 CWPM), cohort 1, by grade and learner language.....	21
Figure 11: Percentage of learners reaching Minimum Reading Comprehension Threshold (80%), cohort 1, by grade and learner language.....	22
Figure 12: Nepali Textbooks available to All-Almost All Learners, Cohort 1	26
Figure 13: Nepali Textbooks available to All-Almost All Learners, Cohort 2.....	26
Figure 14: Supplementary Reading Materials available in the Classroom.....	27
Figure 15: Teacher Training on Language and Reading Instruction	28
Figure 16: Support and Supervision of Teachers	29
Figure 17: School has EGRP Materials at Midline, as reported by Teachers.....	30
Figure 18: Teacher Reading Instruction Practices Index, Cohort 1	31
Figure 19: Teacher Reading Instruction Practices Index, Cohort 2	32
Figure 20: Observer's Teaching Assessment Index, Cohort 1	33
Figure 21: Observer's Teaching Assessment Index, Cohort 2	33
Figure 22: Training of School Management Committee Members	34
Figure 23: Management Index, Cohort 1	35
Figure 24: Management Index, Cohort 2	35
Figure 25: Reading at home.....	37
Figure A2. 1: Figure A2.1. Difference in difference estimator.....	54
Figure A3. 1: Kernel Density of unmatched propensity score by treatment status – Cohort 1 and Control.....	61
Figure A3. 2: Kernel Density of propensity score by treatment status and matching algorithm – Cohort 1 and Control.....	62
Figure A4. 1: Oral Reading Fluency Distributions, by Grade and Learner Language. Cohort 1	77

TABLE OF TABLES

Table 1: Evaluation Question Matrix.....	4
Table 2: Early Literacy Skills, EGRA Subtasks	7
Table 3: Baseline and Midline Samples.....	8
Table 4: Home language of learners and teachers.....	12
Table 5: Effect of EGRP on EGRA Subtasks, cohort 1, Nepali (L1) Learners, by Grade	14
Table 6: Effect of EGRP on EGRA Subtasks, cohort 1, Non-Nepali (L2) Learners, by Grade.....	15
Table 7: Percentage able to read 45 cwpm and respond to 80 percent of reading comprehension questions correctly at baseline and midline, cohort 1, by grade and learner language	19
Table 8: Effect of EGRP on EGRA Subtasks, cohort 2, Nepali (L1) Learners, by Grade	22
Table 9: Effect of EGRP on EGRA Subtasks, cohort 2, Non-Nepali (L2) Learners, by Grade.....	24
Figure A2. 1: Figure A2.1. Difference in difference estimator.....	54
Table A3. 1: Mean, Standard Deviation, Min., Max, and observations.....	55
Table A3. 2: Logit on the probability of treatment – Cohort 1 and Control.....	56
Table A3. 3: Log-likelihood ratio test – Cohort 1 and Control.....	57
Table A3. 4: Hit-Miss Rate and Pseudo R2 tests – Cohort 1 and Control	58
Table A3. 5: Balance between treatment and control characteristics of the unmatched sample	59
Table A3. 6: Balance between treatment and comparison school characteristics at baseline – Cohort 1 and Matched Comparison Group.....	63
Table A3. 7: Balance between Treatment and Comparison at Baseline. Individual Characteristics. Cohorts 1 and 2 and matched comparison groups.....	64
Table A3. 8: Balance between Treatment and Comparison at Baseline. Individual Characteristics. L1 Learners, Cohorts 1 and 2 and matched comparison groups	66
Table A3. 9: Balance between Treatment and Comparison at Baseline. Individual Characteristics. L2 Learners, Cohorts 1 and 2 and matched comparison groups	67
Figure A4. 1: Oral Reading Fluency Distributions, by Grade and Learner Language. Cohort 1	77

ACRONYMS

ADS	Automated Directives System
ABE	Assistance to Basic Education
ACR	All Children Reading
CLA	Central Line Agency
DEC	Development Experience Clearinghouse
DiD	Difference-in-Difference
EGR	Early Grade Reading
EGRA	Early Grade Reading Assessment
EGRP	Early Grade Reading Program
FY	Fiscal Year
GoN	Government of Nepal
IE	Impact Evaluation
MoEST	Ministry of Education, Science, and Technology
MT	Mother Tongue
NORC	NORC at the University of Chicago
RM	Reading Motivator
SOW	Scope of Work
SMC	School Management Committee
TLMs	Teaching and Learning Materials
USAID	U.S. Agency for International Developments

EXECUTIVE SUMMARY

NORC at the University of Chicago, through the USAID Reading and Access Evaluation Contract, serves as the independent evaluator for the external impact evaluation (IE) of the Early Grade Reading Program (EGRP) in Nepal.

PROJECT BACKGROUND

The EGRP in Nepal is being implemented by RTI International. The program has two overarching goals: 1) Improve early grade reading performance of students in Grades 1-3; and 2) Build the Government of Nepal's (GON) capacity to deliver an EGRP that can be replicated nationwide.

Program activities are divided between 3 main components: 1) Improved early grade reading (EGR) instruction; 2) Improved national and district-level early grade reading service delivery; 3) Increased family and community support for early grade reading. Component 1 seeks to support teachers with coaching and professional development while providing classroom instructional materials. Component 2 works to improve the GON capacity for data collection and analysis, policymaking, and management. Component 3 engages local NGOs in working with school management and parent-teacher associations to conduct advocacy campaigns, trainings, reading materials development, and other activities.

Focusing on grades 1-3, the EGRP is being rolled out in 2 cohorts. Cohort 1 includes 6 districts (Saptari, Bhaktapur, Kanchanpur, Banke, Manang, Kaski), while Cohort 2 covers 10 districts (Dang, Bardiya, Dadeldhura, Parsa, Rupandehi, Dolpa, Dhanusa, Surkhet, Mustang, Kailali).

Under Cohort 1, all students (regardless of mother tongue) are currently receiving the EGRP package and will continue to do so. The package consists of the distribution of Nepali Teaching and Learning Materials (TLMs), a ten-day in-service teacher training on the use of TLMs in 2016, orientations for school management committees and head teachers, a teacher coaching and mentoring program, public service announcements through radio to promote early grade reading, and parent and community engagement activities. Cohort 2 will receive the same set of EGRP interventions received by cohort 1 schools. As of the midline data collection, cohort 2 was still in light-intensity implementation mode, having received only some components of the EGRP, namely, delivery of supplementary reading materials but not all TLMs, orientations for head teachers and school management committees on fundamentals and evidence-based practices to improve early grade reading skills, and public service announcements through radio to promote early grade reading. High-intensity roll-out of the EGRP in cohort 2 districts began in the 2018-19 school year, and there are ongoing discussions with the GoN about including an additional set of activities targeting non-native Nepali speakers (Nepali L2), which are yet to be defined but would begin in 2019.

EVALUATION METHODOLOGY OVERVIEW

The main questions the IE seeks to inform on the extent to which the EGRP improved the reading outcomes of native (L1) and non-native (L2) Nepali speakers, and the extent to which the Nepali L2 component of the program –if implemented– generate additional impacts on L2 learners. In addition,

the IE seeks to answer additional questions regarding the extent to which the EGRP results in changes in teachers' reading instruction practices in the classroom, and the extent to which it generates changes in school management support for EGR.

The EGRP was not implemented randomly. Thus, NORC used a quasi-experimental evaluation, which serves as a rigorous alternative to a randomized evaluation, and allows for the credible estimation of program impacts.

NORC first matched comparison and treatment schools in each cohort, selecting schools from the control districts that are most similar to schools in treatment districts in terms of language, baseline EGRA scores, and other observable characteristics. Next, NORC estimated the program impact via a Difference-in-Difference (DiD) approach. DiD is a widely used and simple methodology that compares the changes between baseline and midline in the treatment group with the changes between baseline and midline in a comparison group. The DiD approach assumes that, in the absence of treatment, the two groups of schools would evolve in the same way (parallel trends) over time.

ANSWERING THE EVALUATION QUESTIONS

Evaluation Question 1. To what extent did the EGRP Nepali L1 program improve the reading outcomes of pupils who speak Nepali as a first language (L1 learners) in cohorts 1 and 2?

Answer: At midline, the EGRP had large effects on all measured reading skills for both L1 and L2 learners in cohort 1. The largest effects were found for L1 students, particularly in grades 1 and 3. In contrast, in cohort 2, where most program components have yet to be implemented, there is little evidence that the program has had any impact on L1 learners up to midline.

The EGRP has benefited L1 students with both low and high performance. It reduced the number of zero scores among learners and increased the percentage of learners that reach the tentative benchmarks of 45 correct words per minute and 80 percent oral reading comprehension that the GoN is considering.

For cohort 1, we estimate the program led to an increase of 5.4 and 12.8 correct words per minute (cwpm) in oral reading fluency (ORF) in grades 1 and 3, respectively, and a decrease in the percentage of students with zero scores in ORF of approximately 20 percentage points in both grades 1 and 3.

Evaluation Question 2. To what extent did the EGRP Nepali L1 program improve the reading outcomes of pupils who speak Nepali as a second language (L2 learners) in cohorts 1 and 2?

Answer: In cohort 1, L2 learners generally benefited from the program and reduced the number of learners scoring zero in assessment subtasks. We estimate the EGRP led to increases of 1.2 and 4.5 cwpm in ORF for L2 learners in grades 1 and 2, respectively, and an estimated decrease of about 10 percentage points in the percentage of students in both grades 1 and 3 with zero scores in ORF.

However, the program tends to benefit L1 learners more than L2 learners. As noted at baseline, there is a very large gap between L1 and L2 learners' reading skills, equivalent to approximately one full year of schooling. At midline, the program appears to be increasing the gap. L2 learners not only lag behind

in terms of reading skills, it is also clear that there is a serious deficiency in overall oral Nepali language comprehension among L2 learners.

Evaluation Question 3. To what extent did the EGRP Nepali L2 intervention improve the reading outcomes of pupils who speak Nepali as a second language (L2 Learners) in cohort 2?

Answer: This question can only be answered at endline should the GoN decide to move forward on any Nepali L2 interventions.

Evaluation Question 4. To what extent has the EGRP Nepali L1 program changed teachers' reading instruction practices in the classroom?

Answer: For cohort 1, the EGRP has had impact on teachers' reading instruction practices at midline. While the teaching reading instruction practice index shows a positive impact of the EGRP, the more subjective teaching assessment done by the observer does not however, we have some reservations about the way in which quality of instructions was measured. Teaching practices did not change in Cohort 2. This was expected, given that teachers in these districts did not receive training.

Evaluation Question 5. To what extent has EGRP Nepali L1 program changed the school leadership and management index (as defined in monitoring index), demonstrating active support for EGR?

Answer: The EGRP Nepali L1 program has generated a modest improvement of almost one point (out of 14) in the management index for cohort 1. However, there has been no impact in the index among schools in cohort 2. Of note, only 35 and 25 percent of the school management committee members report receiving training in cohorts 1 and 2, respectively.

CONCLUSIONS

In cohort 1, EGRP benefited both L1 and L2 learners. This is highly desirable given the 45 words per minute and 80 percent oral reading comprehension benchmarks the GoN is considering.

Examining the channels through which the program has functioned, there is no evidence that the program has led to changes in parents' at-home support for their children's reading development. School and SMC support for reading activities shows a very modest improvement, but only around a third of the SMC members interviewed report receiving training.

There is mixed evidence that the program has had an effect on teachers' reading instruction, at least in the way captured by the classrooms observation exercise. While there is a positive effect on teachers' classroom practices, there is no effect on the quality of teaching as judged by observers.

The program has been quite successful at ensuring access to materials, including students' access to Nepali-language textbooks and workbooks, and additional children's reading materials, and teachers' access to teaching guidelines, materials, and curriculum. Almost all teachers report using these resources. Thus, **it is likely that the positive effects of the program have functioned via a combination of at least some improvement in teaching practices with broad access and use of learning and teaching materials.**

The lack of impact on parental engagement and mixed results on teaching quality suggests there is room for improving the community outreach and professional training components of EGRP. Additional gains to learners' test scores could be seen if these components show more evidence of having the expected effects.

RECOMMENDATIONS

A number of recommendations stem from our findings:

Special attention to L2 Learners: The disadvantage in early grade reading skills of L2 learners relative to L1 learners is evident, and the EGRP, while benefiting everyone, appears to widen the gap between the language groups. The situation could have long-lasting consequences in terms of economic development, growth, and social cohesion. Special attention should be devoted to better support non-Nepali speakers in the crucial early years of their schooling. At a minimum, teachers need basic training in providing effective reading instruction for non-Nepali learners in their classrooms.

Improve teacher support supervision: Cohort I teachers received frequent support supervision from one or more sources; however, the levels of satisfaction with this support are low. We recommend exploring why teachers do not find these interactions useful, and adapting support supervision plans accordingly.

Review implementation: A substantial portion of cohort I SMC members report not receiving any training. Some classrooms do not have textbooks, workbooks and/or supplementary reading materials. We recommend systematically confirming all cohort I schools received the full EGRP package.

I. INTRODUCTION

NORC at the University of Chicago, through the USAID Reading and Access Evaluation Contract, serves as the independent evaluator for the external impact evaluation (IE) of the Early Grade Reading Program (EGRP) in Nepal.

The EGRP-Nepal, implemented by RTI International, the Ministry of Education, Science, and Technology (MoEST) and its Central Line Agencies (CLAs) works to develop and test an early grade reading program that the government of Nepal can adopt and rollout to all districts in the country in a cost effective and sustainable manner.

The main purpose of this IE is to assess the causal impact of EGRP-Nepal on the reading outcomes of primary school children – Grades 1, 2, and 3 – who speak Nepali as their first language (L1 Learners) and children who speak Nepali as their second language (L2 Learners). The evaluation measures reading outcomes using subtasks of the Early Grade Reading Assessment (EGRA) tool, widely used for measuring various aspects of reading proficiency.

The evaluation's key audiences and stakeholders include the Government of Nepal (GoN), USAID, RTI International, the donor community and NGOs operating in working in the education sector in Nepal. The evaluation findings will be used to inform programmatic decisions and guide future allocation of resources as well as contribute to the evidence base on what works in improving early grade literacy in linguistically complex settings.

This report presents summary findings from the midterm evaluation, where all schools included in the baseline sample were re-visited and re-assessed.

I.1. CONTEXT AND PROJECT BACKGROUND

A USAID-supported, nationally representative EGRA conducted in Nepal in 2014 found that 34 percent of second graders and 19 percent of third graders could not read a single word of Nepali. Moreover, the assessment showed significant regional disparities, as well as larger deficiencies among students who spoke a language other than Nepali at home.

USAID's EGRP in Nepal is being implemented by RTI International and supports the MoEST and its CLAs—the Curriculum Development Center, Department of Education, Education Review Office, National Center for Educational Development, and Non-Formal Education Center—to develop and test an early grade reading program that is effective, replicable, cost-efficient, and sustainable.

The EGRP has two principal goals: 1) To improve early grade reading performance of students in Grades 1-3; and 2) To build the GoN's capacity to deliver an early grade reading program that can be replicated nationwide.

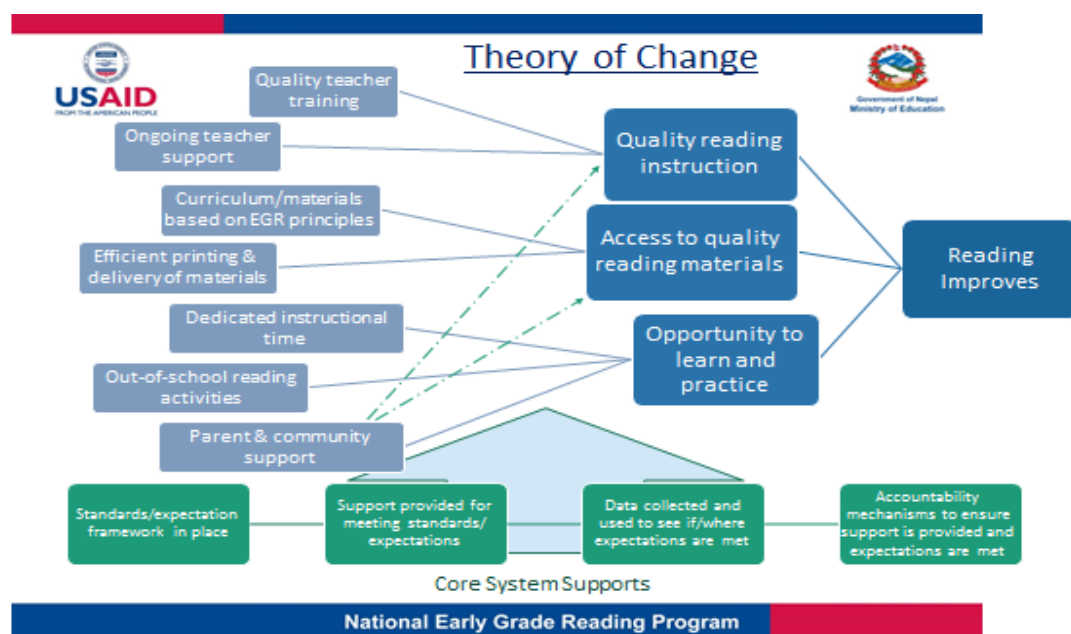
The program has 3 main intermediate results:

- 1) Improve Early Grade Reading Instruction by:

- a. designing, distributing, and using evidence-based early grade reading instructional materials
 - b. providing in-service professional development for teachers in public schools on reading instruction and the use of these materials
 - c. providing monitoring and coaching for teachers in early grade reading instruction
 - d. improving classroom-based and district-based early grade reading assessment processes
- 2) Improve National and District Early Grade Reading Service Delivery by:
- a. improving early grade reading data collection and analysis systems
 - b. institutionalizing policies, standards, and benchmarks that support improved early grade reading instruction
 - c. improving the planning and management of financial, material, and human resources devoted to early grade reading
 - d. facilitating adoption and geographical expansion of national standards for early grade reading improvement
- 3) Increase Family and Community Support for Early Grade Reading by:
- a. increasing family engagement to support reading
 - b. increasing parent–teacher association/school management committee ability to contribute to quality reading instruction
 - c. increasing parent and community capacity to monitor reading progress

As depicted in Figure 1 below, USAID envisions a theory of change where core inputs, including teacher training, on-going teacher support, early grade reading materials, dedicated instruction time, out-of-school-reading activities, and parent and community support, result in quality reading instruction and access to quality reading materials in school, and opportunities to learn and practice reading both in and out of school. Improvements in these three intermediate results lead to the final goal of improving reading outcomes. Specifically, USAID/Nepal hypothesizes that the EGRP will improve the reading skills of both L1 and L2 learners.

Figure 1: EGRP Theory of Change



The EGRP, which focuses on grades 1, 2, and 3, is being rolled out in 2 cohorts. Cohort 1 includes 6 districts -Banke, Bhaktapur, Saptari, Kanchanpur, Kaski, and Manang- while cohort 2 covers 10 additional districts -Dhankuta, Parsa, Rupandehi, Dang, Bardiya, Surkhet, Dolpa, Kailali, Dadeldhura, and Mustang.

In cohort 1, all schools in the six districts (regardless of mother tongue) received the full EGRP package of interventions: distribution of Nepali Teaching and Learning Materials (TLMs) and a ten-day in-service teacher training on the use of TLMs in 2016; continuing training during the following year, which included head teacher and school management committee (SMC) member orientation; a teacher coaching, mentoring, and support model currently implemented through reading motivators (RMs), who are teachers or resources persons within the GoN system; parent and community level engagement activities; and Public Service Announcements on the radio and newspapers to promote early grade reading in the community.

Cohort 2 will receive the same set of EGRP interventions received by cohort 1 schools. There are ongoing discussions with the GoN about including an additional set of activities targeting L2 learners, which are yet to be defined. As of the midline data collection, cohort 2 was still in light intensity implementation mode, having received only some components of the EGRP, namely, delivery of supplementary reading materials but not all TLMs, orientations for head teachers and school management committees on fundamentals and evidence-based practices to improve early grade reading skills, equipment, and Public Service Announcements on the radio and newspapers to promote early grade reading in the community. The EGRP will be rolled out in high intensity in cohort 2 districts in the 2018-19 school year, with potentially any additional EGRP interventions targeting Nepali L2 learners to be rolled out in 2019 (depending on the GoN's determination).

For the purpose of this report, to make a clear distinction between the EGRP interventions in cohorts 1 and 2, we refer to the cohort 1 EGRP interventions as the EGRP Nepali L1 program and the additional interventions that target L2 learners in cohort 2 as the EGRP Nepali L2 interventions.

2.EVALUATION QUESTIONS AND METHODOLOGY OVERVIEW

2.1 EVALUATION QUESTIONS

The questions for this evaluation were discussed among different stakeholders, including the USAID/Nepal Mission, USAID/E3/ED, EGRP, and NORC. In addition, USAID/Nepal Mission officers and EGRP representatives were in meetings with GoN (ERO, CDC, etc.) during the consultation period.

Following multiple meetings and conversations, all parties agreed on the following questions:

Q1. To what extent did EGRP Nepali L1 program improve the reading outcomes of pupils who speak Nepali as a first language (L1 learners) in cohorts 1 and 2?

Q2. To what extent did EGRP Nepali L1 program improve the reading outcomes of pupils who speak Nepali as a second language (L2 learners) in cohorts 1 and 2?

Note: Only to be answered should the GoN decide to move forward on any Nepali L2 interventions.

Q3. To what extent did the EGRP Nepali L2 intervention improve the reading outcomes of pupils who speak Nepali as a second language (L2 Learners) in cohort 2?

The IE also seeks to answer two additional questions about intermediate outcomes:

Q4. To what extent has the EGRP Nepali L1 program changed teachers' reading instruction practices in the classroom?

Q5. To what extent has EGRP Nepali L1 program changed the school leadership and management index (as defined in monitoring index), demonstrating active support for EGR?

The evaluation matrix below presents each evaluation question along with the data sources and analysis methods used to address it. The remaining methodological sections of this report will discuss the data sources, data collection and data analysis methods in more detail.

Table 1: Evaluation Question Matrix

Questions	Data Source	Data Analysis Method
Q1. To what extent did EGRP Nepali L1 program improve the reading outcomes of pupils who speak Nepali as a first language (L1 learners) in cohorts 1 and 2?	EGRA	Compare the average change in EGRA outcomes of the treatment group with the average change in outcomes among a statistically matched comparison group of schools.
Q2. To what extent did EGRP Nepali L1 program improve the reading outcomes of pupils who speak Nepali as a second language (L2 learners) in cohorts 1 and 2?		
<i>Note: Only to be answered should the GoN decide to move forward on any Nepali L2 interventions.</i> Q3. To what extent did the EGRP Nepali L2 intervention improve the reading outcomes of pupils who speak Nepali as a second language (L2 Learners) in cohort 2?	EGRA	<i>Note: Only to be answered should the GoN decide to move forward on any Nepali L2 interventions.</i> At endline, compare average change in EGRA outcomes of the cohort 2 L2 students with the average change in learner outcomes among a statistically matched comparison group of schools. The evaluation will measure the potential additional effect of the EGRP Nepali L2 interventions on L2 learners as compared to the effects of the EGRP Nepali L1 program effects.
Q4. To what extent has the EGRP Nepali L1 program changed teachers' reading instruction practices in the classroom?	Teacher survey and classroom observation	Compare the average change in outcomes of the treatment group and the average change in outcomes in a statistically matched comparison subgroup of schools.

Questions	Data Source	Data Analysis Method
Q5. To what extent has EGRP Nepali LI program changed school leadership and management index (as defined in monitoring index), demonstrating active support for EGR?	Head teacher and SMC member surveys and classroom inventory	Compare average change in management index in the treatment group with the average change in a statistically matched comparison group of schools.

2.2 METHODOLOGY OVERVIEW

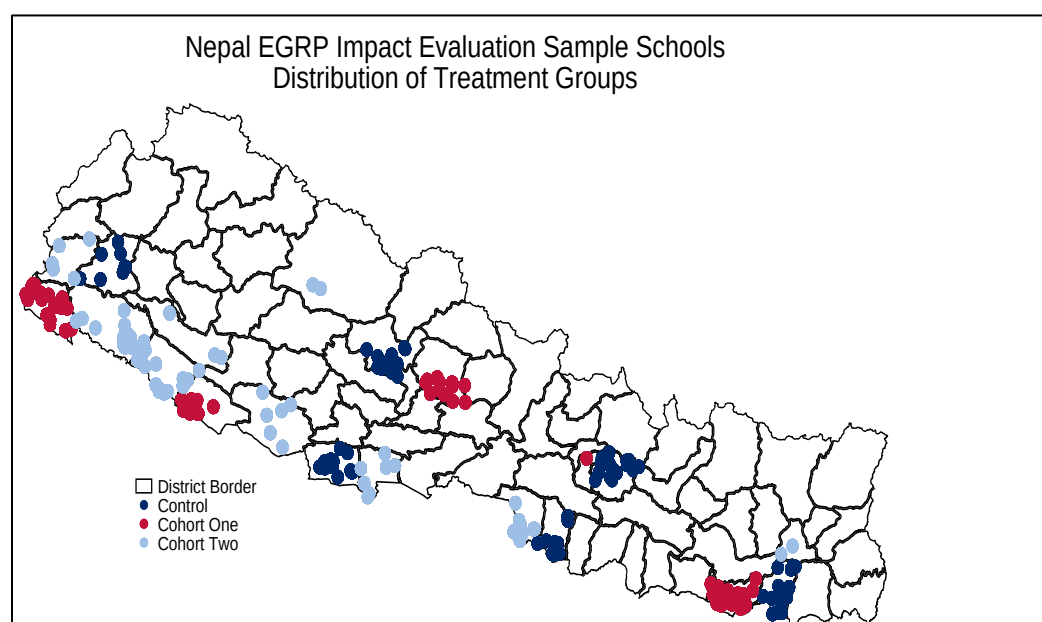
This evaluation uses a quasi-experimental design to measure impact. Using statistical techniques, NORC created a credible group of comparison schools against which to measure changes in schools that receive the EGRP. In this section, we explain the approach.

2.2.1. Treatment Assignment

The EGRP focuses on grades 1, 2, and 3 and is being rolled out in 2 cohorts of districts. Cohort 1 includes 6 districts while Cohort 2 covers 10 additional districts. The EGRP, the MoEST and USAID/Nepal decided that all schools within a treatment district would receive EGRP interventions and, therefore, comparison schools would necessarily need to be found in other districts. To this end, the EGRP selected a group of comparison districts to match the general characteristics of the treatment districts. The dimensions taken into account for the selection of comparison districts were landscape, climate, socio-cultural settings, and economic activity. The comparison districts selected to match treatment districts in both Cohort 1 and Cohort 2 are: Doti, Myagdi, Kapilvastu, Bara, Sunsari, and Kavre.

The geographical distribution of schools in each of the groups is shown in Figure 2.

Figure 2: Location of Sample Schools by Treatment Group



2.2.2. Analytical Approach

Given the selection and roll out of the intervention, a randomized controlled trial was not an option. Hence, our impact evaluation is based on quasi-experimental methods where a comparison group is formed by statistical methods, rather than by random assignment.

We first used matching techniques to form a comparison group of schools from the selected comparison districts that best match the schools receiving treatment. The objective is to make the two groups –treatment and comparison- as similar as possible. The details used to construct the comparison group of schools are described in detail in Annex III where we also show the matched sample balance for cohort 1 and its comparison group, and cohort 2 and its comparison group at baseline.

Once the comparison group of schools is formed we use a Difference-in-Difference (DiD) approach to measure impact. DiD is a widely used and simple methodology that compares the changes between baseline and midline in the treatment group with the changes between baseline and midline in a comparison group. Clearly, both groups of schools do not need to be identical at baseline, given that the comparison relies on the relative changes and not levels.

The DiD approach assumes that, in the absence of treatment, the two groups of schools would evolve in the same way (parallel trends) over time. While we cannot verify this assumption, using matching to ensure treatment and comparison groups are as alike as possible, increases the probability that the groups' trajectories over time are identical.

Finally, to further assure that the groups are as similar as possible and there is no bias, we take into account the basics characteristics of the learners in the analysis and produce adjusted DiD. To do so, we analyze L1 (Nepali) and L2 (Non-Nepali) learners separately and we take into account age and gender of the learners.

More details about the methodology can be found in Annex II.

2.2.3. Data Collection

All data collection and associated work related to this evaluation was and will be handled by RTI and its partners in Nepal. CAMRIS International, the Mission's Monitoring, Evaluation, and Learning (MEL) contractor provided quality assurance oversight of the data collection process. The data collection follows the schedule below:

Baseline: Data collection was originally planned for the last term of the school year 2015-16, in February-March 2016. However, it was interrupted due to earlier than normal exams and was completed in April-May of 2016, the first term of school year 2016-17.

Midline: Midline data collection took place in the last term of the 2017-18 academic year, in February-March 2018

Endline: Current plans indicate that data collection will take place at the end of the 2019-20 academic year, in February-March 2020.

Instruments. The evaluation measures reading outcomes using subtasks of the Early Grade Reading Assessment (EGRA), a widely used tool to measure various aspects of reading proficiency¹. The EGRA subtasks included in the assessment used are described in Table 2 below; all subtasks are administered in Nepali.

Table 2: Early Literacy Skills, EGRA Subtasks

Early Literacy Skill	Sub-test	Measurement
Phonetic Awareness	Letter sound knowledge	Number of letter sounds correctly identified out of 100 in 60 seconds
Matra Knowledge	Matra (or syllables) knowledge	Number of matra sounds correctly identified out of 100 in 60 seconds
Decoding	Nonword decoding	Number of nonwords correctly decoded out of 50 in 60 seconds
Fluency	Oral passage reading	Number of words in a reading passage of approximately 61 words read fluently (with accuracy)
Reading Comprehension	Oral recall	Number of questions (out of 6) about a reading passage (read by student) answered correctly
Listening Comprehension	Oral recall	Number of questions (out of 3) about an passage read aloud (by facilitator) answered correctly

In addition to the EGRA, the following data collection instruments were developed by EGRP/RTI and administered:

- Student Questionnaire: administered to each student selected for assessment
- Parent Questionnaire: administered at the school to one randomly selected parent of a student selected for assessment per school
- Head Teacher Questionnaire: administered to the head teacher in each school visited
- Teacher Questionnaire: administered to one teacher, preference for grade 2 teacher
- SMC Questionnaire: administered to the SMC member whose schools were selected for assessment
- School Inventory: administered at each school visited
- Classroom Inventory: administered in one of the sampled classes, preference for grade 2
- Classroom Observation: administered during reading and writing lessons in one selected classroom for each school visited, preference for grade 2.

All instruments can be found in Annex VII.

Samples. For the baseline and midline data collection, data was collected in grades 1, 2 and 3, creating a cross-section of learners in those grades.

¹ See RTI International. 2015. Early Grade Reading Assessment (EGRA) Toolkit, Second Edition. Washington, DC: United States Agency for International Development for details about this assessment.

At baseline, the sample comprised up to 12 randomly selected students per grade (when possible) at 86 cohort 1 schools, 86 cohort 2 schools, and 120 comparison schools. The sample of comparison schools was larger to maximize the probabilities of a good matching with cohort 1 and cohort 2 schools. While the theoretical sample planned was for 12 students per grade per school, there were numerous instances at baseline where schools had fewer than 12 students per grade, particularly among the comparison schools. The midline data collection re-visited the same schools as baseline and found similar enrollment issues. For the rest of the data collected – from teachers, head teachers, parents, etc. - there is only one observation per school.

Table 3: Baseline and Midline Samples

	Treatment Schools				Comparison Schools	
	Cohort 1		Cohort 2		Baseline	Midline
	Baseline	Midline	Baseline	Midline		
Schools visited	86	85	86	86	120	120
Parents	86	85	86	86	120	120
Teachers	86	85	86	86	120	120
Head teachers	86	85	86	86	120	120
SMC member	86	85	86	86	120	120
School Inventory	86	85	86	86	120	120
Classroom Inventory	86	85	86	86	120	120
Classroom Obs.	86	85	86	86	120	120
Grade 1 learners	839	782	870	821	1110	993
Grade 2 learners	870	822	842	811	1057	1014
Grade 3 learners	885	827	882	829	1132	984
Total learners	2594	2431	2594	2461	3299	2991

Table 3 shows the number of schools visited in each group at baseline and midline, the total number of learners assessed by grade, and the samples for parents, teachers, classroom observations and other data collected at the schools.

More details about the sample can be found in Annex V.

2.2.4. Instrument Creation and Piloting²

Nepal EGRP organized a five-day workshop from August 30 – September 4, 2017 to review the EGRA and EMES instruments prior to the midline assessment. The workshop brought together representatives from the Ministry of Education (MOE), Department of Education (DOE), Education Review Office (ERO), Curriculum Development Center (CDC), National Center for Educational Development (NCED), and educational experts from universities. The purpose of the workshop was to review all of the EGRA and EMES instruments prior to the midline data collection. During the workshop, participants provided recommendations to improve the instruments, and revisions to the instruments

² This sub-section, along with sub-sections 2.2.5-2.2.7, were provided by RTI and EGRP. NORC has provided some light editing to facilitate the inclusion of these contents into the body of the report.

were made as long as they did not impede comparability to the baseline data collection. The EGRA instruments remained the same, but some of the EMES instruments were revised. The revised tools were pre-tested in the schools in Bhaktapur. Students' performance were assessed by using the adapted EGRA tools. Besides the EGRA tool, the background and socio-economic information of the students were collected by using the Student Questionnaire Form. Based on the reflections and observations from the pre-test, the EMES tools were further revised and finalized prior to the pilot in December 2017.

Based on what was learned from the 2017 CB-EGRA, ERO agreed to pilot the EGRA instruments by adding a Silent Reading Passage as one of the sub-tasks in the EGRA tool. The tool was pilot-tested in 20 schools in four districts. An eight-day training was organized to orient the assessors. To ensure the consistency of the assessment, two rounds of Assessors' Accuracy Measures (AAM) were conducted during the training, and only the assessors who scored more than 90% in the AAM were deployed in the field to collect the data. A total of 240 students were assessed during the pilot testing. The analysis showed the positive relationship between the Comprehension sub-task and Silent Reading Passage sub-task. Thus, based on the result of the pilot study, ERO agreed to add the Silent Reading Passage sub-task in the midline instrument.

2.2.5. Assessor Training

The assessors were selected from an open competition of 385 applicants, and based on their qualifications, experience and competence, 150 candidates were selected in the first screening. The 150 persons were then individually interviewed to assess their physical fitness, particularly speaking and hearing ability. Finally, 127 assessors were selected and engaged.

A six-day training was organized to orient the assessors to administer the EGRA and the Student Questionnaire Form and the EMES. Out of 127 assessors, 77 assessors were trained as EGRA assessors and remaining 50 were trained as the Education Management Efficiency Survey (EMES) surveyors. The team supervisors were identified within the groups. During the training, two different schools were visited to provide hands-on experience for the assessors to practice administering the instruments. The assessors' training was facilitated jointly by ERO and EGRP technical staff.

The training was closely monitored by government counterparts, USAID staff, and USAID's Monitoring Evaluation Learning (MEL) project staff. The feedback received from different stakeholders during the training were immediately incorporated to ensure the quality of the training. During the training, two rounds of Assessors' Accuracy Measures (AAM) were conducted. The field assignment of the assessors were decided based on the final AAM score that they obtained during the training. Only the assessors who achieved more than 90% of score were deployed in the field as the EGRA assessors. The average inter-rater reliability score of the final AAM during the training was 98.3%. The figure was quite convincing to confirm the uniformity and reliability of the assessors. The AAM was also conducted for EMES supervisors on the Classroom Observation tool, and the highest rated assessors were selected for the EMES data collection.

2.2.6. Fielding the Survey and Quality Assurance

Data collection for the midline took place during February 8 - March 14, 2018. Besides Nepali, the instructions of EGRA tools were translated into four different local languages—Maithili, Awadhi, Bhojpuri, and Doteli—to enhance the communication between the assessors and the students. While deploying the assessors to different districts to collect the data, the local language competency and understanding of the local context of the assessors were taken into consideration. A team composed of five assessors with two EGRA assessors, two EMES surveyors and one team supervisor were deployed in each school. Assessors spent two days in each school to collect EGRA and EMES data. Twenty-five teams worked simultaneously in different districts. The data collection was electronic, so each of the assessors was equipped with a tablet with EGRA and EMES instruments, Tangerine Software and 3G SIM card for the immediate data feeding and transfer into the server. Each of the teams were provided at least two additional tablets as a contingency in the event there were unanticipated problems with the technology. Moreover, the teams were also equipped with paper instruments as a further back-up. In any event, all the tablets worked smoothly and none of data were collected using the pencil and paper tools.

Three different approaches—provision of Field Supervisors, school-based monitoring, and conducting daily Quick-Checks (field Inter-rater Reliability (IRR))—were arranged to ensure the quality of the assessment.

Onsite monitoring visits were conducted by USAID's MEL project staff, USAID staff, Central Line Agencies of the Ministry of Education including ERO, NCED, DOE and CDC as well as by central, regional and district EGRP office representatives. Furthermore, the District Education Offices from each district also monitored the process. The feedback and other comments provided during the monitoring visits were recorded and communicated to the EGRP central office, and immediate actions were taken to correct whatever issues or challenges arose.

Each assessment team was led by a supervisor who ensured quality data collection by providing appropriate support to the assessors whenever required. The supervisors also worked as a bridge for communication between EGRP, the sub-contractor, and the assessors. The sub-contractor (IIDS) also provisioned the assessment coordinators to coordinate with all the team supervisors and members. The support channel and communication system was very smooth and the assessors received the support they needed throughout the data collection process.

A further, important measure used to ensure the quality of the assessment was the provision of Quick-Checks. A Quick-Check is an assessment of the assessors before taking the EGRA where a pair of the assessors assess one non-sampled student in each school and identify the deviation of each other's assessment. It is a form of field-based inter-rater reliability (IIR). The analysis of the Quick-Check was impressive: At the end of the data collection the field IRR average for quick-check pairs was 99.8%. This high degree of agreement between the pairs ensured the credibility of the assessment.

The role of the EGRP Kathmandu central office was to provide technical support for the assessment process as required. The EGRP central office, regional offices, and the district teams were involved in monitoring the entire process, and provided prompt support to the assessors. In addition, EGRP regional and district teams also coordinated between the District Education Offices and the assessment

teams in order to ensure smooth data collection. The EGRP central team provided trouble shooting related to the use of tablets and other field issues.

The role of the Education Review Office (ERO) and other government agencies was to provide overall direction and oversight of the midline survey. The midline training was facilitated by ERO personnel along with EGRP staff, and ERO personnel also monitored the process independently. Besides ERO, other CLAs such as NCED and DOE also closely monitored the training and data collection process and provided valuable feedback to ensure the quality of the entire process. District Education Offices from all 20 districts also played a vital role during the data collection process. All District Education Offices made at least two monitoring visits in their respective districts. Overall, a high level of government ownership was evident throughout the midline survey process.

2.2.7. Data generation, cleaning, and finalization

As noted above, the EGRA midline data were collected electronically using the Tangerine survey data collection application on tablet devices and uploaded to the Tangerine server by the assessors using a wireless internet connection. Once the data were uploaded to the Tangerine servers, it was accessed by RTI statisticians in the .csv format, with one file per instrument.

These files were then imported into Stata where they were cleaned and checked. Practice observations, incomplete observations (false starts or abrupt stops), and duplicates were identified, documented, and deleted. During the data collection period, 14 data quality monitoring reports were provided to the EGRP team. The reports provided information on the count of assessments completed per team per day and were shared with the field supervisors for cross-checking, and if any discrepancies were found, they were rectified. All student assessment data were scored and any extreme values were investigated for assessor error and either deleted or corrected. School-level and student-level weights were then applied to the data to ensure that the dataset was representative to the cohort level of the Early Grade Reading Program. In each step of the process, the work was checked for quality and accuracy by a senior statistician.

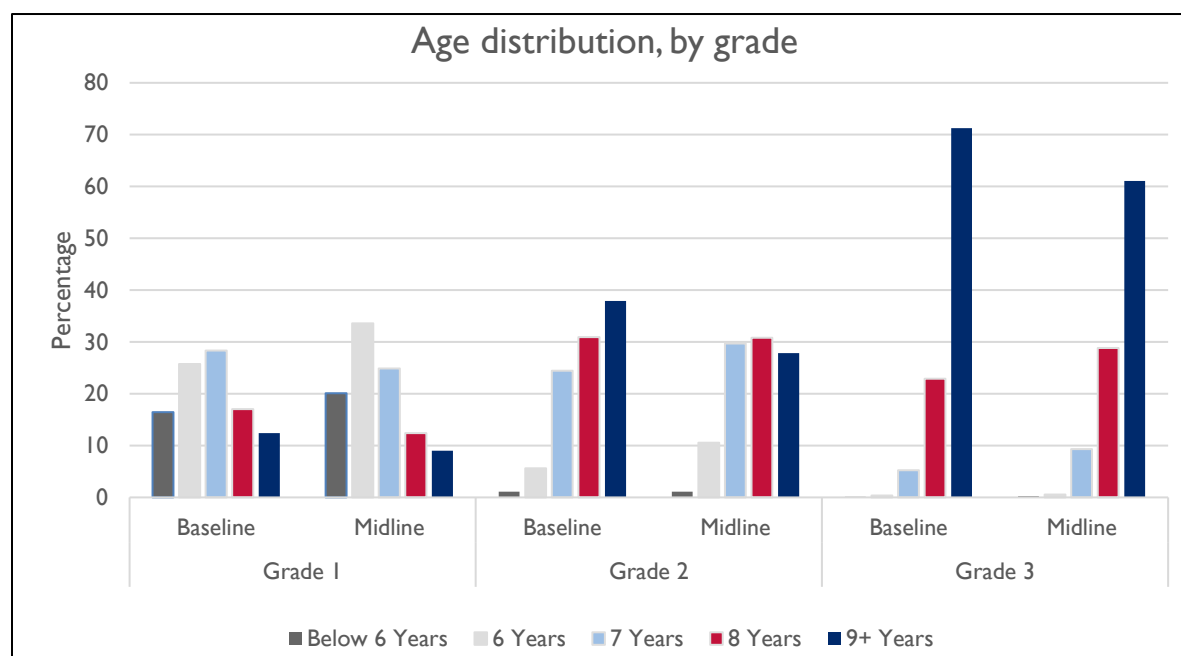
The final dataset was delivered to the external evaluator via a secure server. Outcomes and EGRP impact were then assessed by the external evaluator using complex survey weighted analysis.

3. FINDINGS

3.1 SOME LEARNER CHARACTERISTICS

More than half the learners assessed are girls (55.2%) in both baseline and midline samples. Figure 3 presents learners' age distribution in 5 categories: below 6 years of age, 6 years, 7 years, 8 years, and 9 or more. Assuming those "below 6" are 5 years old, as the minimum age for admission at grade one is five (UNESCO, 2015), and the group classified as "9 or more" has an average age of 10, the average age of the learners in the sample is slightly higher at baseline than at midline, 8.1 and 7.8, respectively.

Figure 3: Age Distribution by Grade, Baseline and Midline



Note: Sample weights applied to recover population representativeness.

Nepali is the national official language and the medium of teaching and learning in Nepal. However, there are 123 languages spoken as mother tongue in the country. Table 4 shows the distribution of home languages for learners and teachers in our sample. Most students and teachers report a language other than Nepali as their mother tongue, but this proportion is substantially higher among students than among teachers.

Table 4: Home language of learners and teachers

	Baseline		Midline	
	Learners	Teachers	Learners	Teachers
Nepali (L1)	32.6%	49.0%	33.5%	47.0%
Non-Nepali (L2)	67.4%	51.0%	66.5%	53.0%
Maithali	19.5%	16.5%	21.4%	16.1%
Bhojpuri	26.9%	15.7%	27.8%	17.5%
Tharu	11.7%	14.9%	15.1%	14.2%
Tamang	2.3%	2.6%	1.6%	2.9%
Awadhi	12.5%	0.0%	13.1%	5.9%
Others	27.0%	50.2%	21.0	43.4%
Observations	8487	292	7883	291

Note: Sample weights applied to recover population representativeness

3.2 EGRP IMPACT ON EGRA SCORES

In this section, we address the first two evaluation questions:

EQ1: To what extent did EGRP (Nepali L1 program) improve the reading outcomes of pupils who speak Nepali as a first language (L1 learners) in cohorts 1 and 2?

EQ2: To what extent did EGRP (Nepali L1 program) improve the reading outcomes of pupils who speak Nepali as a second language (L2 learners) in cohorts 1 and 2?

At midline the EGRP had large effects on all measured reading skills for both L1 and L2 learners in cohort 1. The largest effects were found for L1 students.

In cohort 2, where most components of the program have yet to be implemented, there is little evidence that the program has had any impact on L1 or L2 learners up to the midline.

Below, we present and discuss the evidence that allows us to answer evaluation questions 1 and 2. We start with the analysis for cohort 1. As mentioned, cohort 1 has received the full set of EGRP interventions and, therefore, at midline already had the preconditions to show a potential impact. We continue with the analysis for cohort 2, which had only received part of the EGRP interventions at midline, mostly limited to the delivery of supplementary reading materials, orientations for head teachers and school management committees, and Public Service Announcements on the radio and newspapers to promote early grade reading in the community.

3.2.1. Cohort 1

We show the effect of the EGRP on each skill measured by the EGRA. Table 5 and 6 show results by grade for L1 and L2 learners respectively. The tables show the mean score for each EGRA subtask at baseline and midline, for treatment and comparison schools, along with their difference. It can be seen that at baseline the two groups are quite similar, as intended. To estimate the impact of EGRP we compare the change between baseline and midline in the treatment group to the change between baseline and midline for the comparison group, this is the DiD, and is shown in column 7. In column 8 we included the DiD but added controls for students' characteristics. The final column shows the effect size in units of standard deviation (std. dev.).³

For example, focusing on the Oral Reading Fluency (ORF) subtask for grade 3, we see that the average number of correct words at baseline is 18.4 for the comparison group, and 18.2 for the treatment group, with a small difference of 0.2 words between the groups. At midline, the averages are 14.8 and 29.2 correct words for comparison and treatment groups respectively, amounting to a difference between the groups of 14.4 words. The simple DiD between the groups is, therefore, 14.6 words (14.4 - (-0.2)). When we adjust to take into account learners' basic characteristics, the adjusted DiD is 12.8 words, showing a clear advantage of the cohort 1 treatment group over their comparison counterparts. The effect is statistically significant and its size is equivalent to 0.56 of a std. dev. This effect is clearly large; an improvement of 12.8 words due to the EGRP is more than the difference between average ORF in Grade 2 (ORF=22.2 cwpm) and in Grade 3 (ORF=29.2 cwpm) for the treatment group at midline, for example.

³ Effect size refers to the difference between treatment and comparison groups as a proportion of the standard deviation of the distribution. In our case we use the pooled standard deviation of the groups at endline.

Table 5: Effect of EGRP on EGRA Subtasks, cohort I, Nepali (L1) Learners, by Grade

	Baseline			Midline			DiD (7=6-3)	Adjusted DiD (8)	Adj. Effect Size (9)
	Comp p (1)	Treat t (2)	Diff (3=2-1)	Comp (4)	Treat (5)	Diff (6=5-4)			
Correct Letter Sounds Per Minute [Max=100]									
Grade 1	11.5	16.9	5.4	10.7	23.3	12.6	7.2***	5.9**	0.37
Grade 2	22.7	25.2	2.5	23.2	33.0	9.8	7.3	5.4	0.27
Grade 3	33.5	33.0	-0.5	28.0	39.2	11.2	11.7***	10.1***	0.50
Correct Matra Per Minute [Max=100]									
Grade 1	3.7	5.8	2.1	4.3	13.8	9.5	7.4**	6.7**	0.45
Grade 2	12.7	13.4	0.7	17.0	23.9	6.9	6.2	5.2	0.24
Grade 3	21.4	21.9	0.5	20.2	29.5	9.3	8.8**	7.2*	0.33
Correct Invented Words Per Minute [Max= 50]									
Grade 1	0.9	1.2	0.3	1.0	3.6	2.6	2.3***	2.1**	0.43
Grade 2	3.5	4.1	0.6	4.3	7.3	3.0	2.4	2.2	0.29
Grade 3	6.1	6.8	0.7	4.9	8.8	3.9	3.2**	2.6**	0.31
Oral Reading Fluency									
Grade 1	2.4	3.4	1.0	2.4	9.3	6.9	5.9***	5.4**	0.48
Grade 2	10.2	10.7	0.5	12.0	22.2	10.2	9.7*	8.7*	0.44
Grade 3	18.4	18.2	-0.2	14.8	29.2	14.4	14.6***	12.8***	0.56
Reading Comprehension, Percentage Correct									
Grade 1	3.9	5.9	2.0	4.5	16.7	12.2	10.2***	9.6**	0.47
Grade 2	17.7	17.6	-0.1	22.6	34.0	11.4	11.5	8.8	0.29
Grade 3	32.5	28.5	-4.0	28.1	44.4	16.3	20.3***	19.1***	0.59
Listening Comprehension, Percentage Correct									
Grade 1	15.1	15.8	0.7	10.9	20.9	10.0	9.3*	9.4**	0.34
Grade 2	27.1	26.0	-1.1	24.6	33.9	9.3	10.4*	4.7	0.15
Grade 3	35.4	36.4	1.0	38.2	43.5	5.3	4.3	4.5	0.13

Note: Propensity score matching weights applied. Adjusted DiD includes, student gender and age. Effect size refers to the difference between treatment and comparison groups as a proportion of the standard deviation of the distribution. In our case we use the pooled standard deviation of the groups at endline. *** p<0.01, ** p<0.05, * p<0.1

Table 5 shows that for all EGRA subtasks and across grades the effect of EGRP at midline is positive and, in most cases, statistically significant for L1 learners. There are some subtasks for which the effect for grade 2, although positive, is not statistically significant at conventional levels. In general, those effects are smaller than those for grades 1 and 3, and the lack of statistical significance seems to be the result of a sample underpowered to detect effects of that size⁴.

Table 6 replicates the information presented in Table 5 but now focusing on L2 learners. There are several findings worth noting. First, the average scores for L2 learners are much lower than those displayed in Table 5 for L1 learners. L2 learners' scores are approximately one full grade behind those of L1 learners. This is true for all subtasks in which learners were assessed. For example, in the letter

⁴ The impact is always statistically significant when analyzing all grades together.

sound subtask, column (4) shows the comparison group at midline where first grade L1 learners scored 10.7 correct letter sounds per minute (clspm) compared to L2 learners who scored only 5.5 words. Similarly, L1 learners scored 23.2 clspm in grade 2, while L2 learners only reach that level in grade 3.

The difference is also very large in the Listening Comprehension subtask. On average, the grade 1 L1 learners in the comparison group at midline (Table 5, column 4) answered 10.9 percent of the listening comprehension questions correctly, while the Grade 1 L2 learners (Table 6, column 4) answered only 2.8 percent. For grade 2, these figures are 24.6 percent for L1 learners and 10.9 for L2 learners, and for grade 3, L1 learners responded 38.2 percent correctly and while L2 learners only answered 24.2 percent of the questions correctly. This suggests a serious deficiency in overall oral Nepali language comprehension among L2 learners rather than problems in reading skills exclusively.

Second, the effect of the EGRP is positive for L2 learners for all subtasks and grades. While the program shows a positive impact in all subtasks, the absolute levels of reading competence remain very low, on average, for this group. For example, in grade 3, L2 learners that received EGRP read on average only 13.8 words per minute and only answered 37 percent of the listening comprehension questions correctly.

Finally, comparing Tables 7 and 8, it is clear that the impact of the program is lower for L2 learners than for L1 learners, further increasing the gap between the two groups. One exception is the listening comprehension subtask, where the effect size is larger for L2 than for L1 learners. Although the comprehension levels remain lower than desired, the program seems to have an important impact in the listening comprehension of L2 students.

Table 6: Effect of EGRP on EGRA Subtasks, cohort 1, Non-Nepali (L2) Learners, by Grade

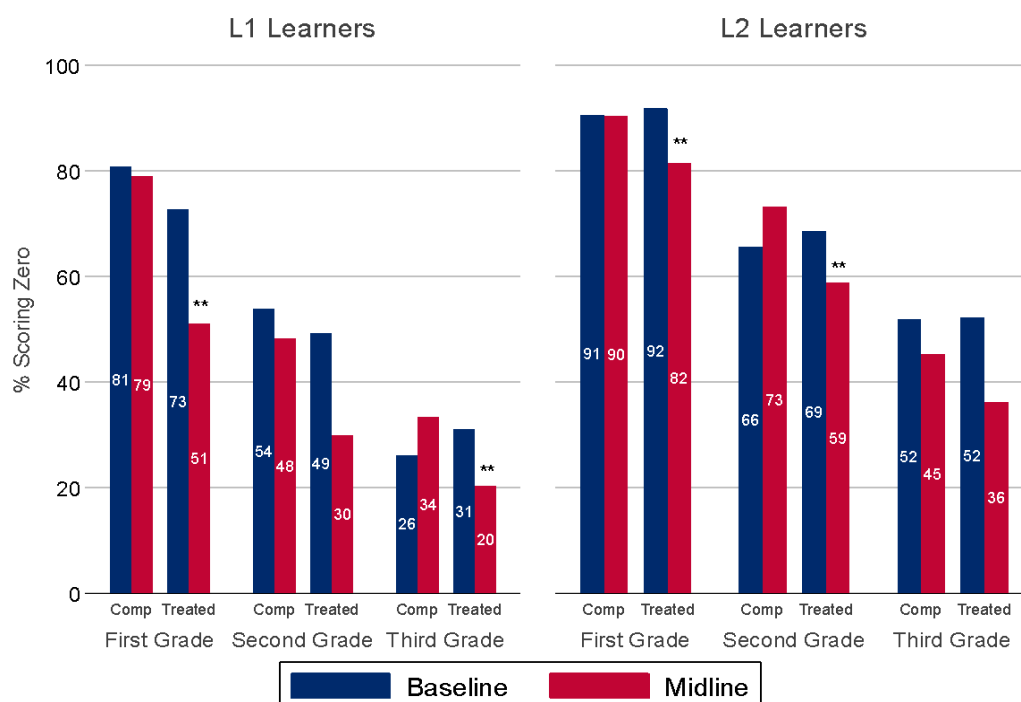
	Baseline	Midline		Baseline	Midline			Adjusted	Adj.
	Comp	Treat	Diff	Comp	Treat	Diff	DiD	DiD	Effect
	(1)	(2)	(3=2-1)	(4)	(5)	(6=5-4)	(7=6-3)	(8)	Size
									(9)
Correct Sound of Letters Per Minute [Max=100]									
Grade 1	8.1	6.5	-1.6	5.5	8.1	2.6	4.2***	5.6***	0.57
Grade 2	17.3	16.4	-0.9	14.2	16.1	1.9	2.8	3.7	0.25
Grade 3	23.2	22.6	-0.6	23.1	27.4	4.3	4.9*	4.9**	0.26
Correct Matra Per Minute [Max=100]									
Grade 1	1.7	1.5	-0.2	1.8	2.8	1.0	1.2	2.0**	0.32
Grade 2	6.3	7.6	1.3	6.6	8.3	1.7	0.4	0.8	0.06
Grade 3	12.8	12.8	0.0	16.5	19.0	2.5	2.5	2.2	0.11
Correct Invented Words Per Minute [Max=50]									
Grade 1	0.4	0.4	0.0	0.2	0.7	0.5	0.5**	0.7**	0.35
Grade 2	1.6	2.2	0.6	1.8	2.7	0.9	0.3	0.5	0.10
Grade 3	3.7	4.4	0.7	4.2	6.7	2.5	1.8	1.5	0.19
Oral Reading Fluency [Max=100]									
Grade 1	0.7	0.6	-0.1	0.6	1.3	0.7	0.8*	1.2***	0.35
Grade 2	4.0	4.5	0.5	3.4	5.1	1.7	1.2	1.4	0.15

	Baseline	Midline		Baseline	Midline		DiD	Adjusted	Adj.
	Comp	Treat	Diff	Comp	Treat	Diff	(7=6-3)	DiD	Effect
	(1)	(2)	(3=2-1)	(4)	(5)	(6=5-4)		(8)	Size
									(9)
Grade 3	9.5	9.6	0.1	9.3	13.8	4.5	4.4**	4.5**	0.29
Reading Comprehension, Percentage Correct									
Grade 1	1.0	0.7	-0.3	0.7	2.7	2.0	2.3**	2.8***	0.41
Grade 2	6.3	6.5	0.2	5.5	9.1	3.6	3.4	4.0	0.25
Grade 3	12.6	13.4	0.8	15.9	22.1	6.2	5.4*	5.1	0.20
Listening Comprehension, Percentage Correct									
Grade 1	4.1	5.3	1.2	2.8	16.0	13.2	12.0***	12.1***	0.54
Grade 2	12.1	13.5	1.4	11.8	26.7	14.9	13.5***	14.2***	0.47
Grade 3	15.8	17.4	1.6	24.2	37.2	13.0	11.4**	10.8*	0.32

Note: Propensity score matching weights applied. Adjusted DiD includes, student gender and age. Effect size refers to the difference between treatment and comparison groups as a proportion of the standard deviation of the distribution. In our case we use the pooled standard deviation of the groups at endline. *** p<0.01, ** p<0.05, * p<0.1

Percentage of Students with Zero Scores. The implementation of the EGRP in Nepal was motivated in part by the high percentage of young learners unable to read in Nepali. As mentioned previously, a 2014 assessment supported by USAID found that 34 percent of second graders and 19 percent of third graders were unable to read a single word in Nepali. Thus, beyond raising average reading assessment scores and words read per minute, part of the goal of the EGRP is to target the weakest learners in order to reduce the prevalence of illiteracy.

Figure 4: Percentage of zero score in Oral Reading Fluency, cohort I, by grade and learner language

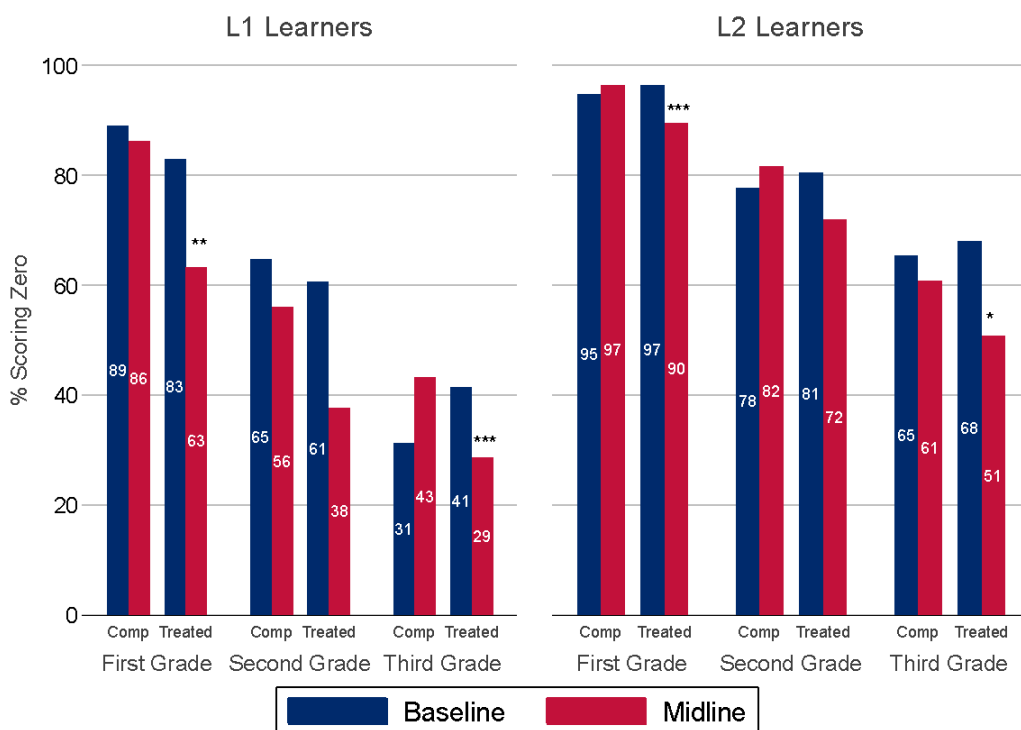


Note: Propensity score matching weights applied. DID *** p<0.01, ** p<0.05, * p<0.1

The percentage of students scoring zero on the oral reading and reading comprehension sections are of obvious interest. As shown in Figure 4, the percentage of students with zero scores in the oral reading section were comparable at baseline between treatment and comparison schools across all grade levels. These baseline rates also stand out for being quite high. Among L1 learners, 80 percent of first graders, approximately 50 percent of second graders, and around 30 percent of third graders were unable to read a single word during the assessment at baseline, figures that are substantially higher than in the aforementioned 2014 assessment, and reflecting the weakness of students in the targeted schools. Among the L2 learners the percentages are substantially higher. Even in third grade more than half of the learners assessed were not able to read one word.

Between baseline and midline, these figures remained essentially unchanged in the comparison group, while in the intervention schools each grade level saw a reduction in zero scores. The impact of the program for the L1 learners is around 20 percentage points in grade 1 and 3 (effects size 0.36 std. dev.), and 10 percentage points for grade 2 (effect size 0.18 std. dev.). The same is true for L2 learners that received the program. The reduction of zero scores was around 10 percentage points for grade 1 and 3 (effect size 0.4 and 0.16)⁵ and 17 per grade 2 (effect size 0.4). The stars in the figure indicate when the EGRP effect is statistically significant.

Figure 5: Percentage of zero scores in Reading Comprehension, Cohort I, by grade and learner language



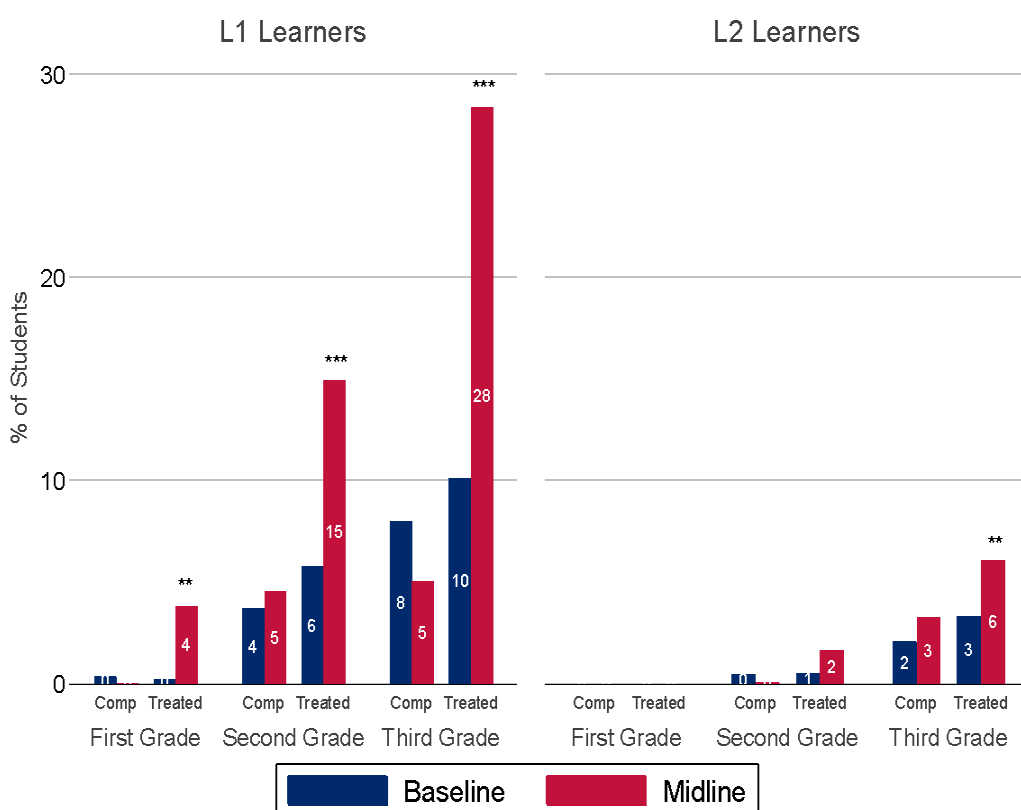
Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

⁵ Note that comparing effect sizes in terms of units of standard deviations when the underlying distributions are very different, can be misleading, as the measurement artificially inflates the effectiveness of interventions done on more homogeneous groups, all else equal. In our case, the standard deviation of oral reading, for example, among L2 learners is much smaller than that of L1 learners.

These reductions in zero scores are also reflected in the reduction of zero scores in the reading comprehension subtask. Figure 5 shows zero scores for Cohort 1 and comparison groups by grade and learner language. The proportion of zero scores is high, as expected given the frequency of zero scores in the reading subtask, but on average, L1 and L2 learners that received the program have reduced the proportion of zero scores in reading comprehension. For L1 students, the program had an impact of 17 percentage points (effect size is 0.37 of a std. dev.) in first grade, 14 percentage points (0.18 std. dev.) in second grade and 24 percentage points (0.48 std. dev.) in third grade. The stars in the figure indicate where the EGRP effect is statistically significant. The effect size of the program is also sizeable for L2 learners, 9 percentage points for grade 1 and 13 for grade 2 and 3 (effect sizes of 0.43, 0.35 and 0.24 std. dev. respectively). It is important to note that the proportion of L2 learners with zero comprehension scores is still very high at midline.

Tentative Benchmarks. In Figures 6 and 7, we show the percentage of learners in treatment and comparison groups able to read 45 correctly identified words per minute and to answer 80% of the reading comprehension questions correctly, respectively. These standards are the achievement level that the Government of Nepal's reading benchmarks.

Figure 6: Percentage of learners reaching Minimum Fluency Threshold (45 CWPM), cohort 1, by grade and learner language

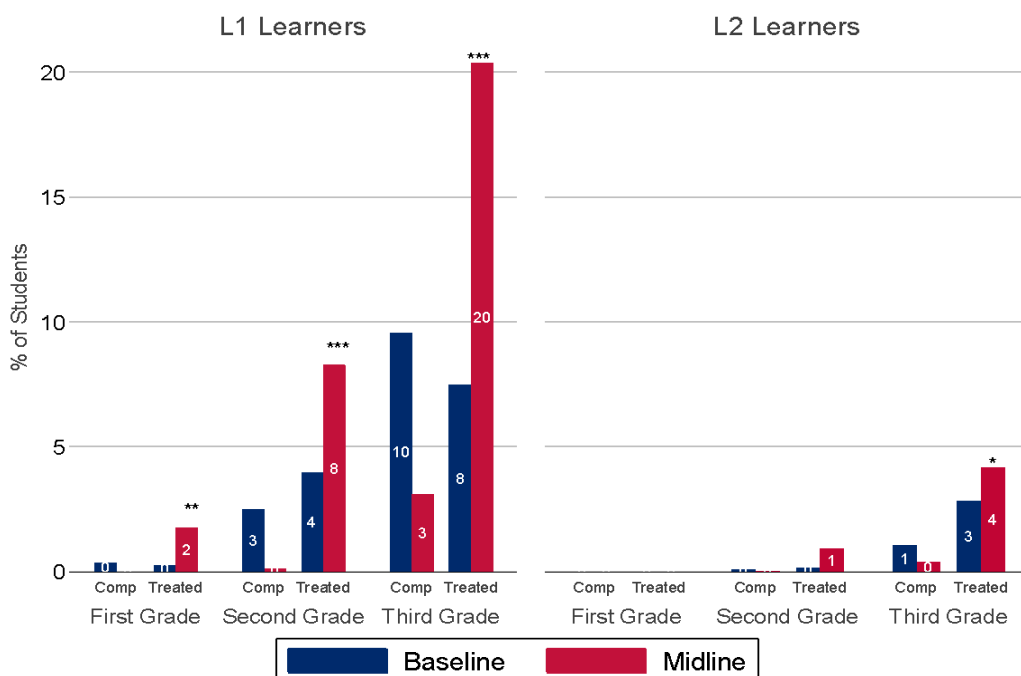


Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

L1 learners reach the thresholds in much higher proportions than learners that have a different mother tongue than Nepali. Not surprisingly, grade 3 learners do better than those in Grade 2. However, L2

learners in grade 3 perform worse than L1 learners attending grade 2, reflecting more than one school year gap between the two groups. In general, very few learners were able to reach the benchmark at baseline. At midline, the positive impact of the EGRP, particularly among L1 students, is evident. The change in the proportion of learners reaching the thresholds is larger for treated than for comparison learners. Although there is a positive effect among L2 learners, the data once more suggests that additional and focused attention is required for them.

Figure 7: Percentage of learners reaching Minimum Reading Comprehension Threshold (80%), cohort I, by grade and learner language



Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Finally, we calculated the percentage of learners able to reach both 45 cwpm and 80 percent of the reading comprehension questions at midline. Table 7 shows percentage estimates for grades 2 and 3 by the language of the learners at baseline and midline.

Table 7: Percentage able to read 45 cwpm and respond to 80 percent of reading comprehension questions correctly at baseline and midline, cohort I, by grade and learner language

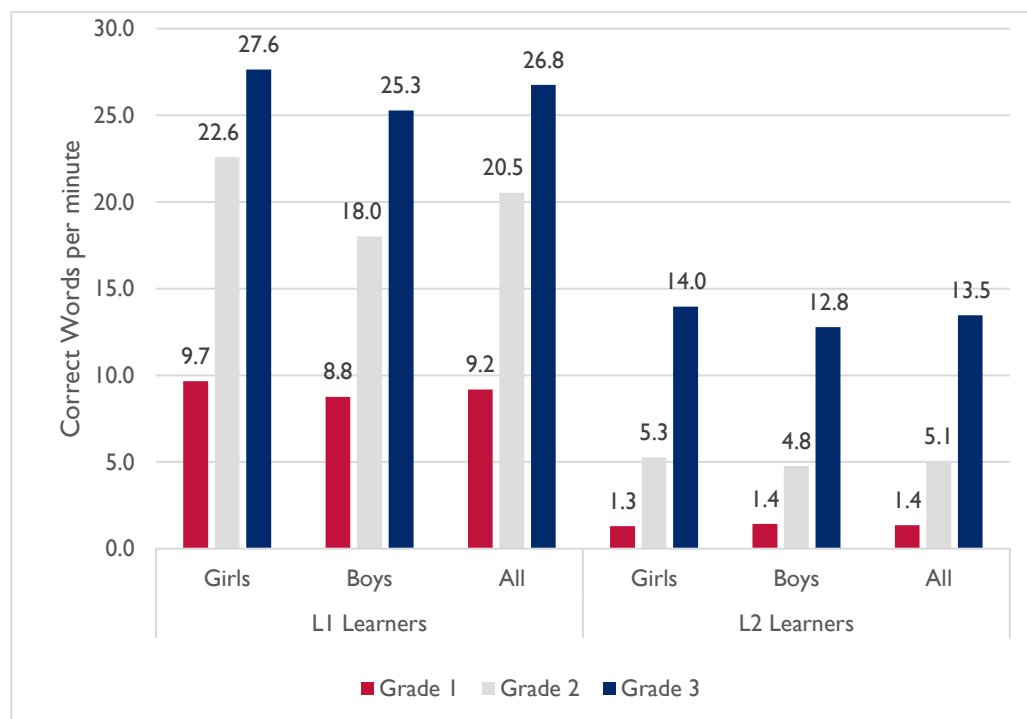
	L1 Learners		L2 Learners	
	Baseline	Midline	Baseline	Midline
Grade 2 (n)	5.1% (18)	10.1% (29)	0.0% (0)	0.6% (4)
Total sample	358	288	512	534
Grade 3 (n)	11.1% (45)	26.3% (80)	2.0% (10)	2.9% (15)
Total sample	401	302	484	525

Note: Sample weights applied to recover population representativeness

The figures show substantial progress between midline and baseline, particularly among L1 learners. At midline more than a quarter of the L1 learners in grade 3 reach the threshold. In contrast L2 learners are far from those levels and less than 3 percent reach the benchmark.

Differential Impact by Gender: In Figure 8 we show the average correct words per minute for the cohort I treatment group for girls and boys, separately for L1 and L2 learners. Girls seem to perform slightly better than boys but none of the differences in means are statistically significant.

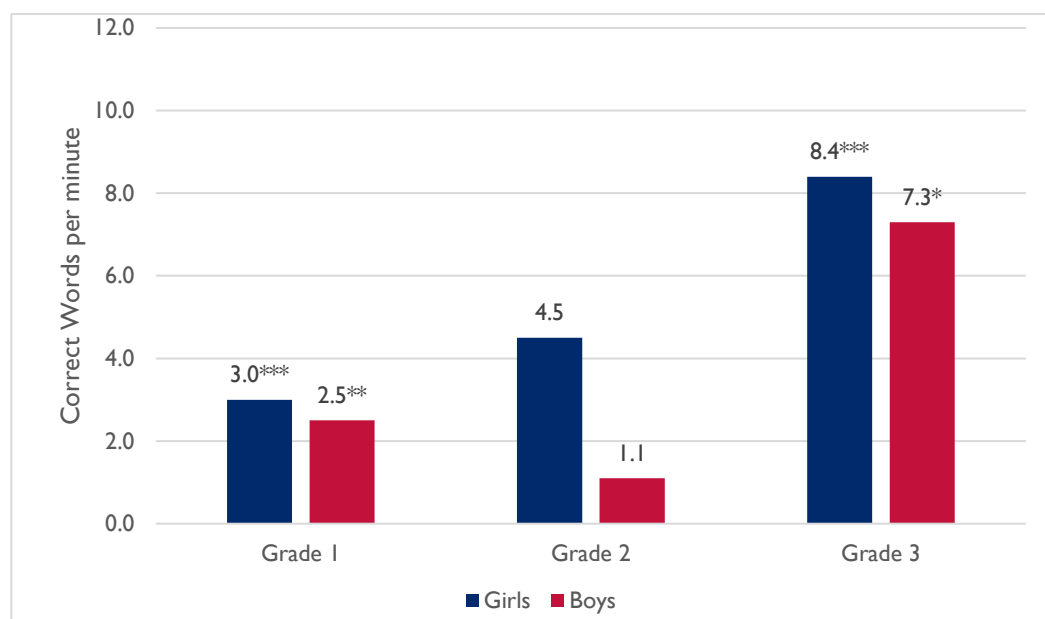
Figure 8: Mean Oral Reading Scores at Midline, cohort I Treatment, by Gender and Grade



EGRP does not show differential effects for girls and boys. Figure 9 shows the EGRP effect on oral reading by gender. All the estimated effects of the EGRP on boys and girls are positive and significant (significance levels shown by stars) for grades 1 and 3. However, when comparing between boys and girls, while there are some differences in favor of girls, these differential effects are not significant.

In Annex IV, we show more details. We present the EGRP effects by gender for all the EGRA subtasks and the estimated differences in effects between boys and girls, which are not significant.

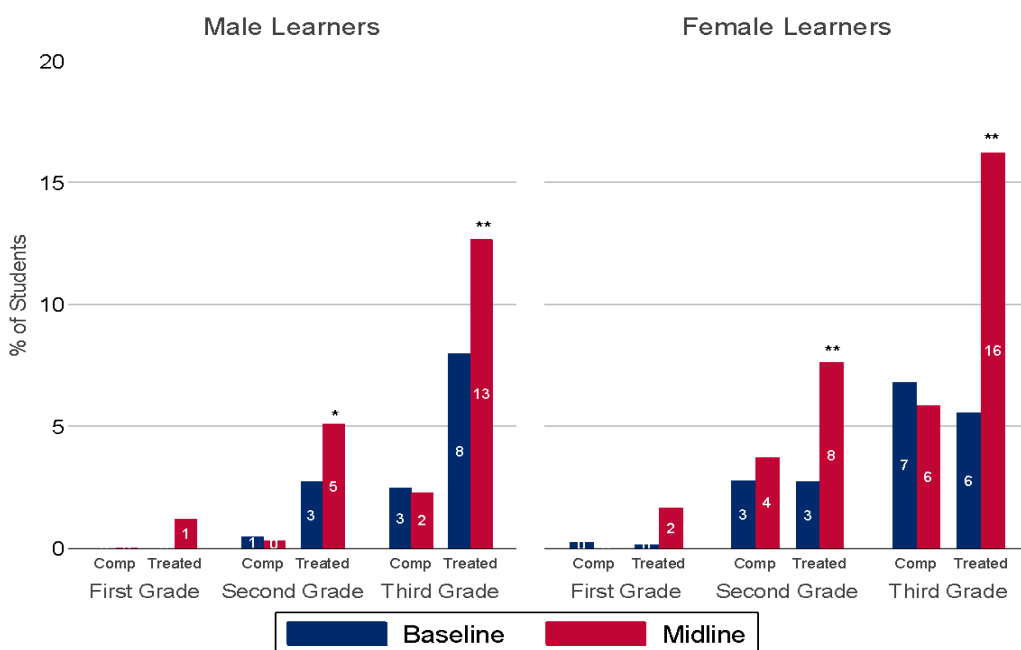
Figure 9: EGRP Effect on Oral Reading Scores at Midline, cohort I, by Gender and Grade



Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Finally, we show the percentage of girls and boys reaching a benchmark of 45 cwpm and the percentage reaching 80 percent in oral reading comprehension, for each grade, in Figures 10 and 11.

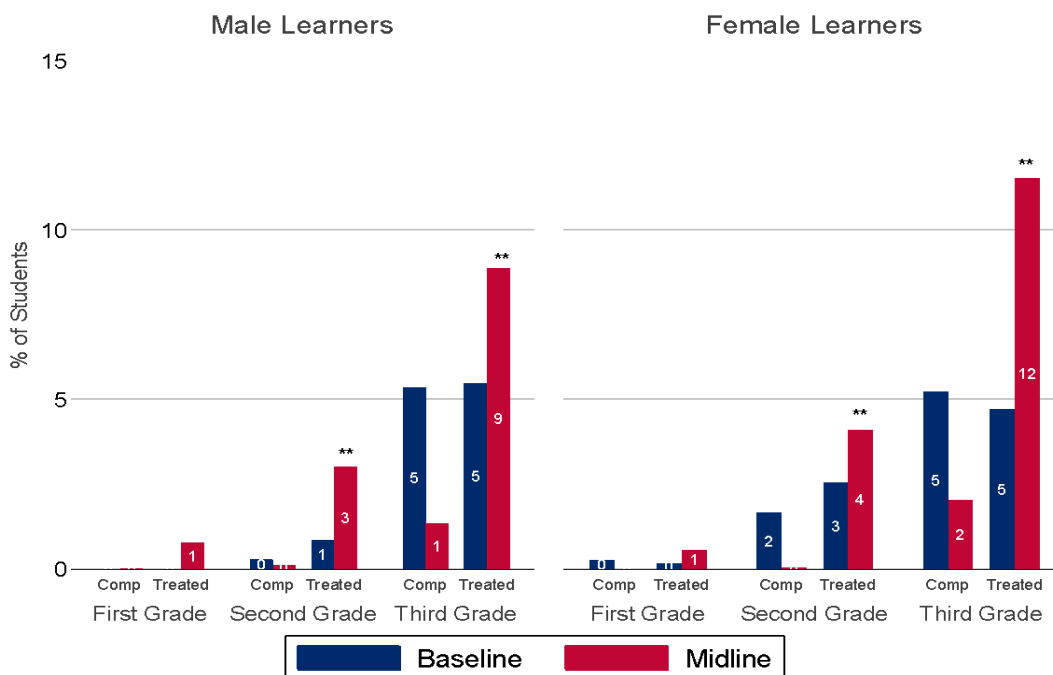
Figure 10: Percentage of learners reaching Minimum Fluency Threshold (45 CWPM), cohort I, by grade and learner language



Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

We find that both girls and boys have benefited from the EGRP at midline. The fraction of learners – boys and girls - reaching the 45 cwpm benchmark and the 80% comprehension threshold increased in the treatment group relative to the comparison group. Again, the data suggest a slight and non-significant difference between boys and girls in favor of girls.

Figure 11: Percentage of learners reaching Minimum Reading Comprehension Threshold (80%), cohort 1, by grade and learner language



Note: Propensity score matching weights applied. *** p<0.01, ** p<0.05, * p<0.1

3.2.2. Cohort 2

In cohort 2 districts, where most aspects of the program are still to be implemented, there is little evidence that the EGRP has had any effect on learners' reading outcomes.

The only effects seen for LI learners in cohort 2 occurred in Grade 3, and were markedly smaller than effects found in cohort 1. For third graders in cohort 2, the EGRP is associated with increases in 5.8 words per minute in oral reading fluency, and an increase in 10.5 percentage points in reading comprehension questions answered correctly. When analyzing mechanisms in Section 3.3, we speculate about possible reasons behind these findings.

Table 8: Effect of EGRP on EGRA Subtasks, cohort 2, Nepali (LI) Learners, by Grade

	Baseline			Midline			DiD (7=6-3)	Adjusted DiD (8)	Adj. Effect Size (9)
	Comp (1)	Treat (2)	Diff (3=2-1)	Comp (4)	Treat (5)	Diff (6=5-4)			
Correct Letter Sounds Per Minute [Max=100]									
Grade 1	13.8	15.6	1.8	10.9	12.4	1.5	-0.3	-0.5	-0.04

	Baseline			Midline			DiD (7=6-3)	Adjusted DiD (8)	Adj. Effect Size (9)
	Comp (1)	Treat (2)	Diff (3=2-1)	Comp (4)	Treat (5)	Diff (6=5-4)			
Grade 2	27.2	25.3	-1.9	23.1	21.9	-1.2	0.7	1.4	0.08
Grade 3	37.7	38.2	0.5	32.7	36	3.3	2.8	3.4	0.17
Correct Matra Per Minute [Max=100]									
Grade 1	4.7	5.5	0.8	3.5	4.3	0.8	0.0	-0.1	-0.01
Grade 2	15.9	15	-0.9	15.5	12.9	-2.6	-1.7	-0.4	-0.02
Grade 3	25.9	27.9	2.0	24.6	28.4	3.8	1.8	2.6	0.12
Correct Invented Words Per Minute [Max= 50]									
Grade 1	1.1	1.2	0.1	0.7	0.9	0.2	0.1	0.0	0.00
Grade 2	4.4	4.3	-0.1	4.7	3.5	-1.2	-1.1	-0.5	-0.08
Grade 3	7.8	9.3	1.5	6.8	8.3	1.5	-0.0	0.3	0.04
Oral Reading Fluency									
Grade 1	2.3	2.4	0.1	1.5	2	0.5	0.4	0.3	0.05
Grade 2	12.5	10.7	-1.8	11.4	9.7	-1.7	0.1	0.8	0.05
Grade 3	24	23.6	-0.4	20	24.6	4.6	5.0*	5.8**	0.28
Reading Comprehension, Percentage Correct									
Grade 1	4.1	3.8	-0.3	3.2	4.1	0.9	1.2	1.1	0.10
Grade 2	21.7	18.2	-3.5	22.8	16.9	-5.9	-2.4	-1.4	-0.05
Grade 3	43.4	38.3	-5.1	36	40.4	4.4	9.5**	10.5**	0.34
Listening Comprehension, Percentage Correct									
Grade 1	16	14	-2	10.7	12.9	2.2	4.2	4.8	0.21
Grade 2	29.7	22	-7.7	30	26.7	-3.3	4.4	3.7	0.12
Grade 3	36.6	37.3	0.7	42.8	42.3	-0.5	-1.2	-1.3	-0.04

Note: Propensity score matching weights applied. *** p<0.01, ** p<0.05, * p<0.1. Adjusted DiD includes, student gender and age. Effect size refers to the difference between treatment and comparison groups as a proportion of the standard deviation of the distribution. In our case we use the pooled standard deviation of the groups at endline.

Among L2 learners, we do not find any significant effect of the EGRP in cohort 2, as expected.

As the implementation of the cohort 2 intervention package begins in earnest in the 2018 school year, we expect the effects will begin to approximate the estimated effects found for cohort 1. Of course, if the cohort 2 treatment package eventually contains additional inventions specifically targeting Nepali L2 learners, we would expect to detect outsized impacts on non-native Nepali-speaking students in cohort 2. However, this portion of the program has yet to be defined and implemented; hence, its effects cannot be measured until the next round of data collection.

Table 9: Effect of EGRP on EGRA Subtasks, cohort 2, Non-Nepali (L2) Learners, by Grade

	Baseline			Midline			DiD (7=6-3)	Adjusted DiD (8)	Adj. Effect Size (9)
	Comp (1)	Treat (2)	Diff (3=2-1)	Comp (4)	Treat (5)	Diff (6=5-4)			
Correct Letter Sounds Per Minute [Max=100]									
Grade 1	8.4	8.1	-0.3	5	6.8	1.8	2.1	2.2	0.25
Grade 2	19.9	17.5	-2.4	15	15.5	0.5	2.9	3.4	0.22
Grade 3	28.5	25.9	-2.6	24.7	23.5	-1.2	1.4	0.6	0.03
Correct Matra Per Minute [Max=100]									
Grade 1	2	1.8	-0.2	1.5	1.9	0.4	0.6	0.8	0.15
Grade 2	8.1	7.9	-0.2	7.3	7.8	0.5	0.7	1.1	0.09
Grade 3	18.4	17.4	-1	18	15.1	-2.9	-1.9	-2.7	-0.14
Correct Invented Words Per Minute [Max= 50]									
Grade 1	0.4	0.4	0	0.1	0.5	0.4	0.4	0.3	0.18
Grade 2	1.8	2	0.2	1.7	1.8	0.1	-0.1	-0.1	-0.02
Grade 3	5.2	5.5	0.3	4.6	4	-0.6	-0.9	-1.2	-0.18
Oral Reading Fluency [Max=100]									
Grade 1	0.9	0.9	0	0.5	0.7	0.2	0.2	0.2	0.08
Grade 2	5	5	0	3.9	4.1	0.2	0.2	0.5	0.06
Grade 3	14.1	12.9	-1.2	11.1	9.9	-1.2	-0.0	-0.6	-0.04
Reading Comprehension, Percentage Correct									
Grade 1	1.1	1.3	0.2	0.7	1.2	0.5	0.3	0.5	0.1
Grade 2	8.4	7.7	-0.7	6.3	7.9	1.6	2.3	2.5	0.16
Grade 3	20	20	0	18.8	17.2	-1.6	-1.6	-2.5	-0.1
Listening Comprehension, Percentage Correct									
Grade 1	5.1	5.4	0.3	5	5.5	0.5	0.2	-0.6	-0.04
Grade 2	13.7	12.3	-1.4	12.7	15.8	3.1	4.5	5.1	0.21
Grade 3	19.7	20	0.3	26.3	24.3	-2	-2.3	-2.5	-0.08

Note: Propensity score matching weights applied. *** p<0.01, ** p<0.05, * p<0.1. Adjusted DiD includes, student gender and age. Effect size refers to the difference between treatment and comparison groups as a proportion of the standard deviation of the distribution. In our case we use the pooled standard deviation of the groups at endline.

In the next section we analyze the mechanisms behind the positive impact of EGRP among cohort 1 learners, and the lack of any effect for cohort 2.

3.3. MECHANISMS BEHIND THE IMPACT OF EGRP

In this section, we study the channels through which the program may have produced an effect on learners' reading skills.

The EGRP theory of change states that learners' reading skills will improve if they are exposed to high-quality reading instruction, have access to high-quality reading materials in school, and have reading opportunities both in and out of school.

Teacher training and continuous teacher support is expected to increase the quality and quantity of reading instruction, as well as the exposure to reading opportunities. EGR materials development and delivery gives learners access to high-quality reading resources. Parental and community engagement and support for early reading increases reading opportunities and contributes to quality reading instruction through school support.

We use data from teachers, head teachers, SMC members, classroom inventory and observations, parents, and schools to address the extent to which the program has been implemented in schools from cohort 1 and 2 districts, and compare them with comparison schools. It is important to keep in mind we only have 291 observations in total in each of these categories (1 observation per school; 85 observations in cohort 1 treatment schools, 86 in cohort 2 treatment schools, and 120 observations in comparison schools).

Using the available data, we start by exploring the availability of reading materials in classrooms. Then we focus on teacher training and support, teacher use of EGRP materials and how this translates into changes in reading instruction practices in the classroom. Finally, we study changes in school and school management support for reading activities and parental engagement

3.3.1. Classroom Reading Materials

Through classroom inventories, enumerators recorded whether learners had Nepali reading materials, such as textbooks. The fraction of learners that had textbooks and workbooks were recorded in five categories: none, very few, less than half, half or just over half, all or almost all. Enumerators also checked whether supplementary reading materials were available in the classrooms.

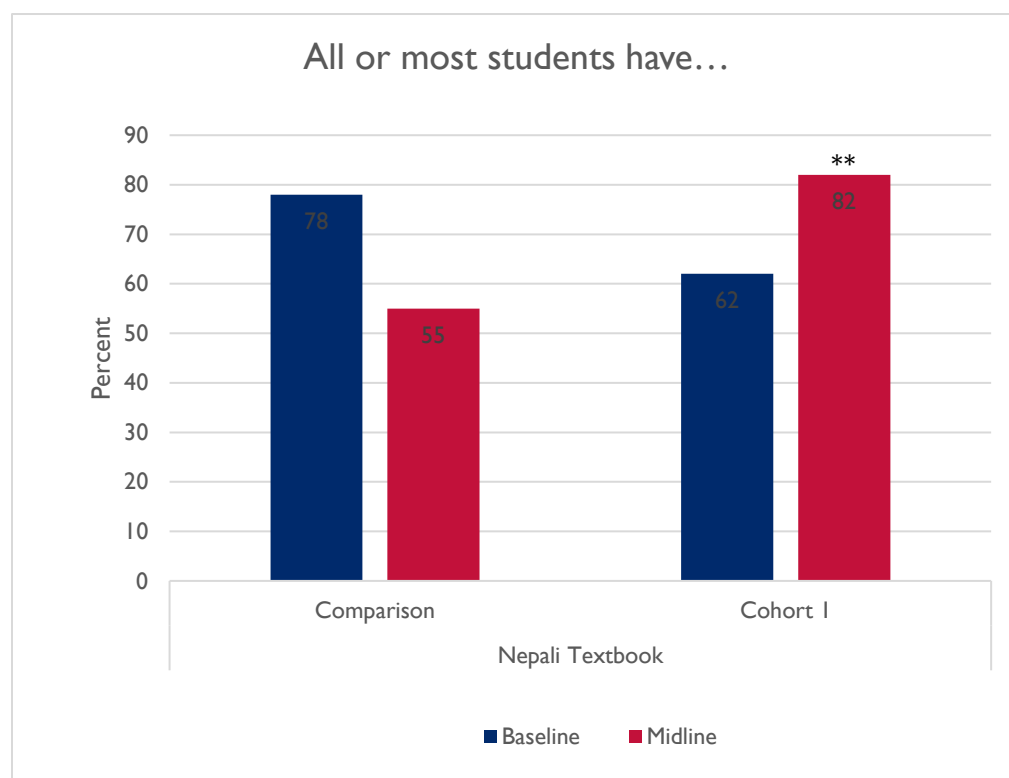
Figures 12 and 13 show the fraction of classrooms visited where all or almost all learners have a Nepali textbook for cohorts 1 and 2, respectively. **We find access to textbooks has increased for cohort 1 and the change is significant. The access, however, is still not universal.** For cohort 2, changes in access to these materials are not significant, which is expected as the program only provided supplemental readings in those districts.

In cohort 1, learners also received Nepali workbooks which were not available before. At midline 91 percent of the classrooms had workbooks for all or almost all the learners.

Although not every child has a Nepali textbook and workbook, the availability of these materials is much higher than the availability of local languages resources. In no group –cohort 1, 2 or their respective comparison groups- are there more than a few schools where most learners have non-Nepali textbooks or workbooks.

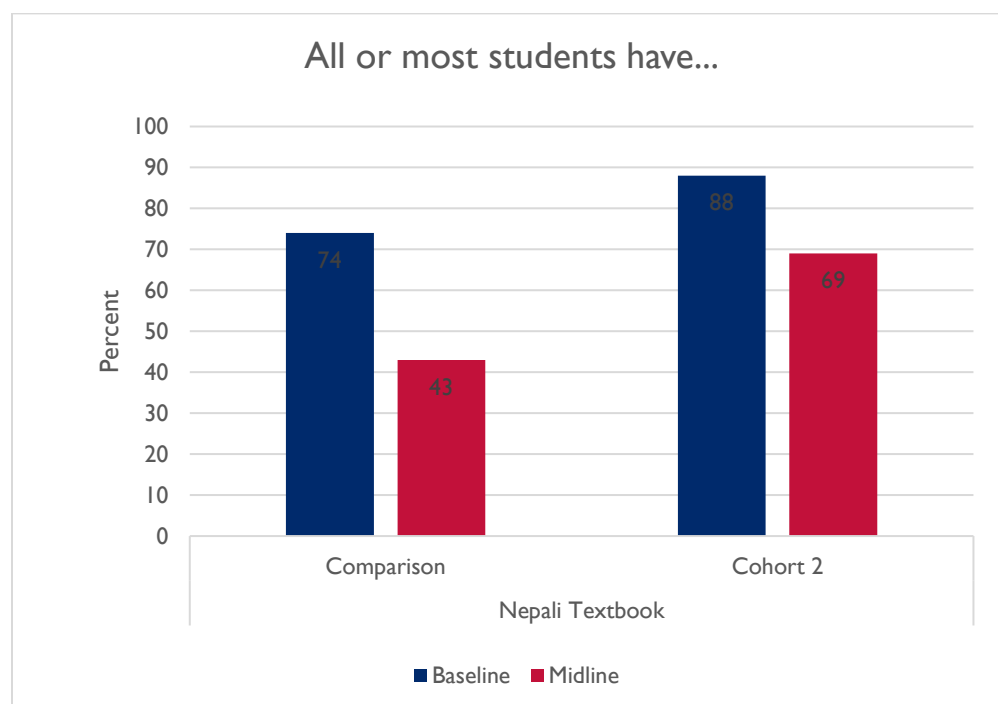
Figure 14 shows the proportion of classrooms where **supplemental reading materials** are available. In this case we see that all groups have increase the percentage of classrooms with additional reading resources; however, **cohort 1 and cohort 2 have seen a substantial and significantly higher increase.** Still, around 10 percent of the cohort 1 classes visited and 20 percent of cohort 2, are lacking supplemental reading materials.

Figure 12: Nepali Textbooks available to All-Almost All Learners, Cohort 1



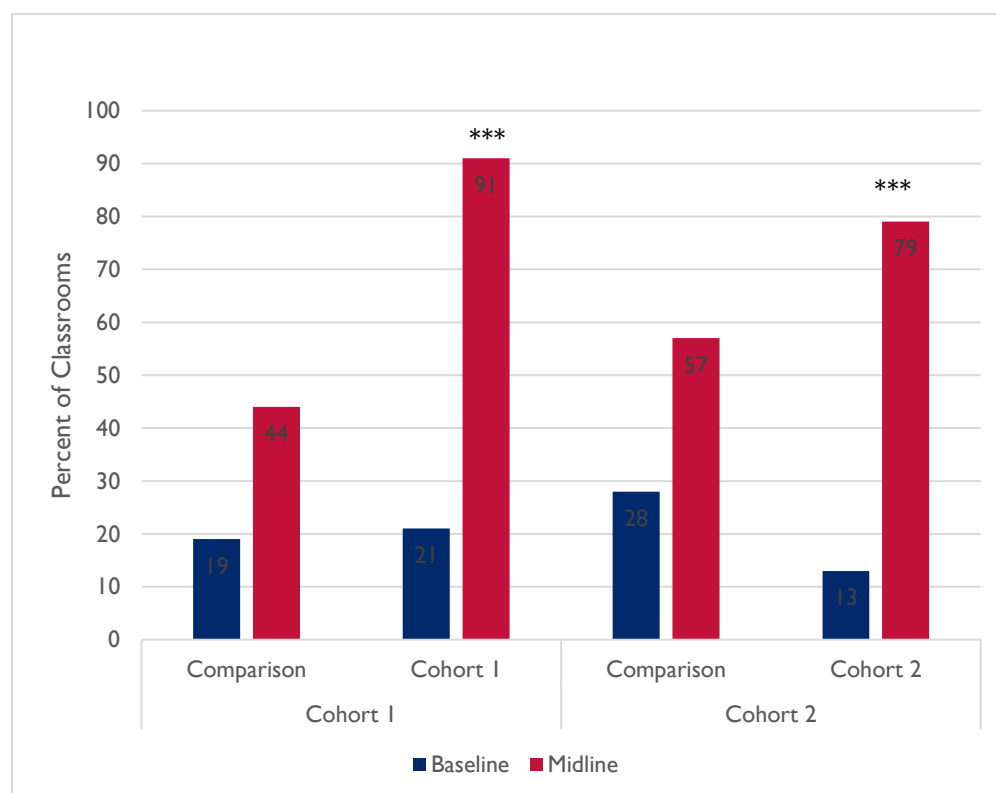
Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Figure 13: Nepali Textbooks available to All-Almost All Learners, Cohort 2



Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Figure 14: Supplementary Reading Materials available in the Classroom

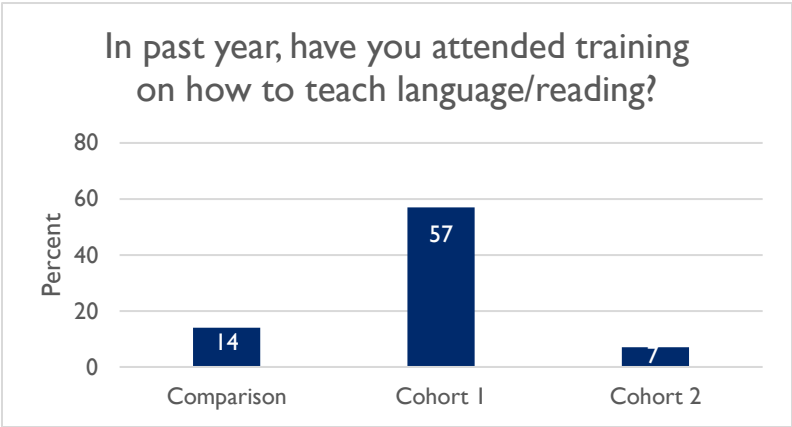


Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

3.3.2. Teacher Training, Support, and Instructional Practices

Teachers were asked whether at any time in previous year, they had attended a seven-day teacher professional development training on how to teach language and reading. Figure 15 shows that while the majority of teachers -57 percent - in cohort 1 answered affirmatively, only 7 percent of cohort 2 teachers, and 14 percent of teachers in the comparison group did. This is consistent with what we would expect, given that training was conducted with cohort 1 teachers. However, the percentage for cohort 1 is lower than expected. While it is possible that not all the teachers attended the training, it is likely that some cohort 1 teachers had received EGRP training earlier. The question refers only to the previous year, but EGRP training started in 2016 for cohort 1.

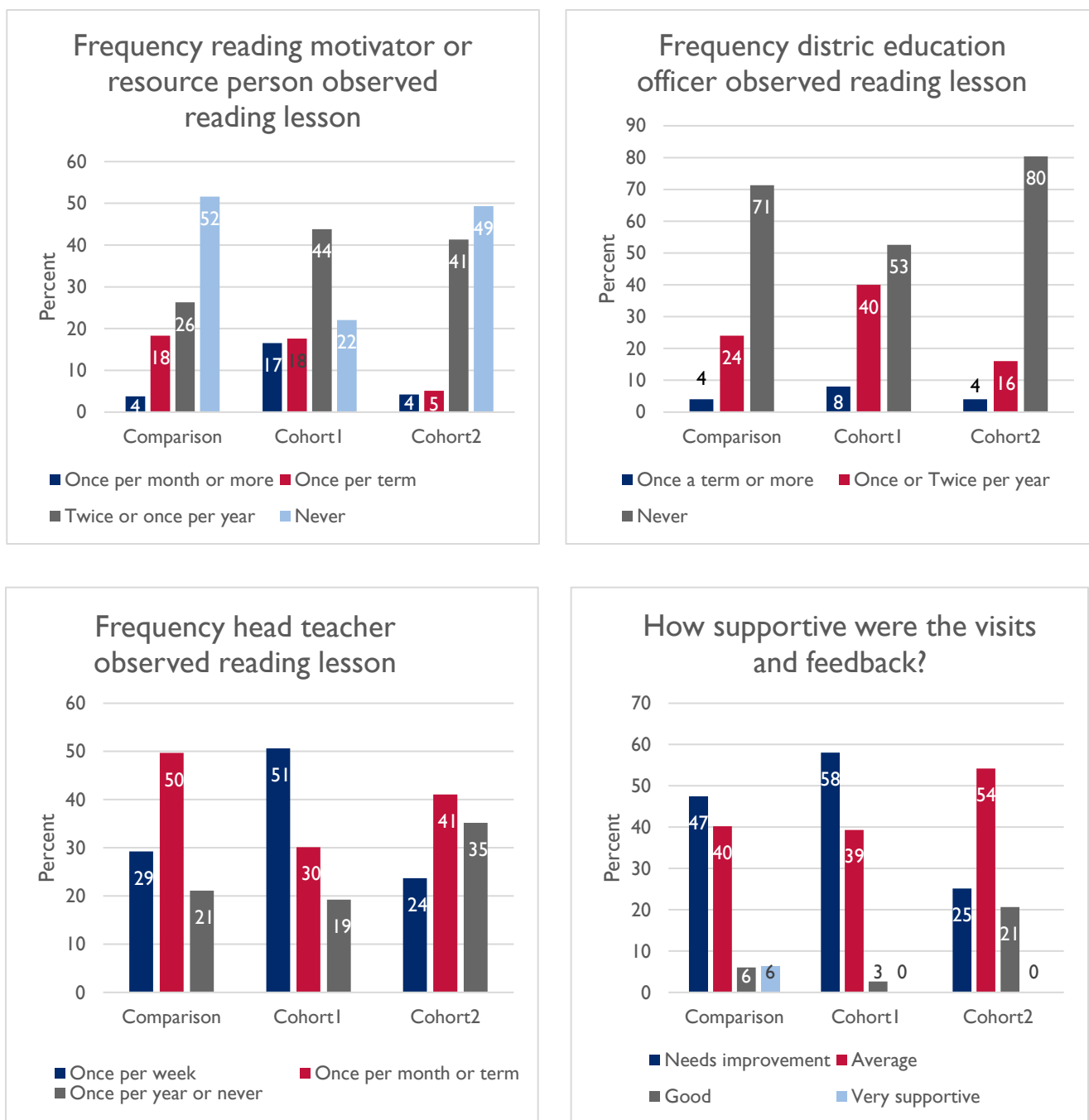
Figure 15: Teacher Training on Language and Reading Instruction



We also found **teachers in cohort 1 were more likely to receive support and supervision from different sources**. In Figure 16, we show that 35 percent of the teachers in cohort 1 report that a motivator or resource person observed their reading class at least once per term. Only 22 percent of them have not received this supervision. Among teachers in cohort 2 and comparison groups, around 50 percent of the teachers report no supervision of this type and the percentage that received it at least once per term goes down to 9 and 22 percent, respectively. Cohort 1 teachers also report more supervision from education officers than their peers from cohort 2 and the comparison groups, although more than half of them indicated that the district education officer never observed their reading classes. It is also much more common among teachers in cohort 1 to have the head teacher observing their reading classes every week.

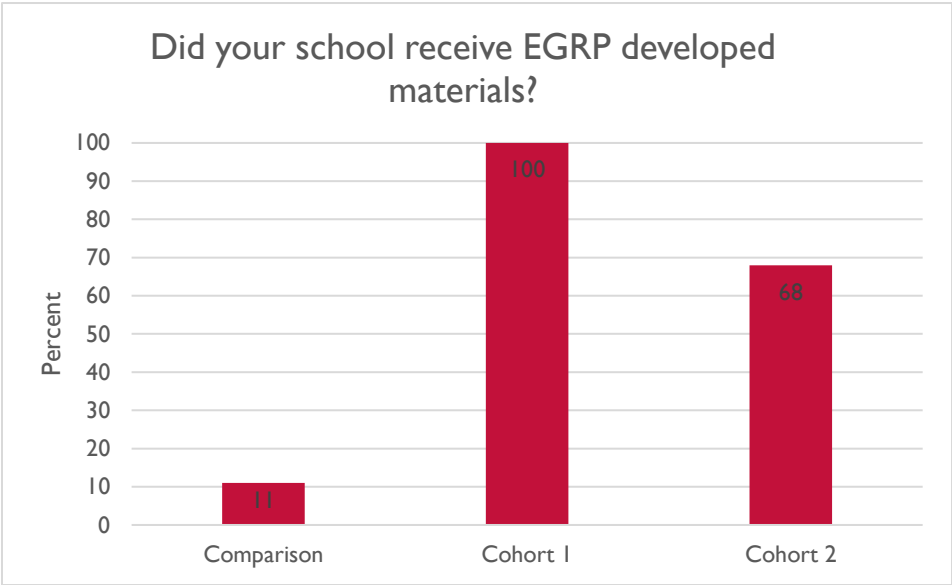
While the frequency of the visits was higher in **cohort 1**, the group **does not seem satisfied with the support and feedback**. More than half of the teachers said it needs improvement, and very few report it being good or very supportive. The figures do not appear too different from those in the comparison group.

Figure 16: Support and Supervision of Teachers



Teachers were also asked if their schools have EGRP developed materials. **Every teacher in cohort 1 confirmed receiving the material, while 68 percent of the teachers in cohort 2 did so.** Only 11 percent of the teachers in the comparison group responded affirmatively. Given that the comparison group is located in completely different districts, this is most likely explained by a few cases of mistakes in the data collection or incorrect teachers' reports.

Figure 17: School has EGRP Materials at Midline, as reported by Teachers



Summarizing, we find high levels of implementation of the program in cohort 1 schools in terms of access to EGRP materials, teacher training and teacher support and supervision, although teachers do not seem to find such a support particularly useful.

We turn now to explore whether these components of the program have translated into improved reading instruction practices. The aim of this analysis is to answer evaluation question 4:

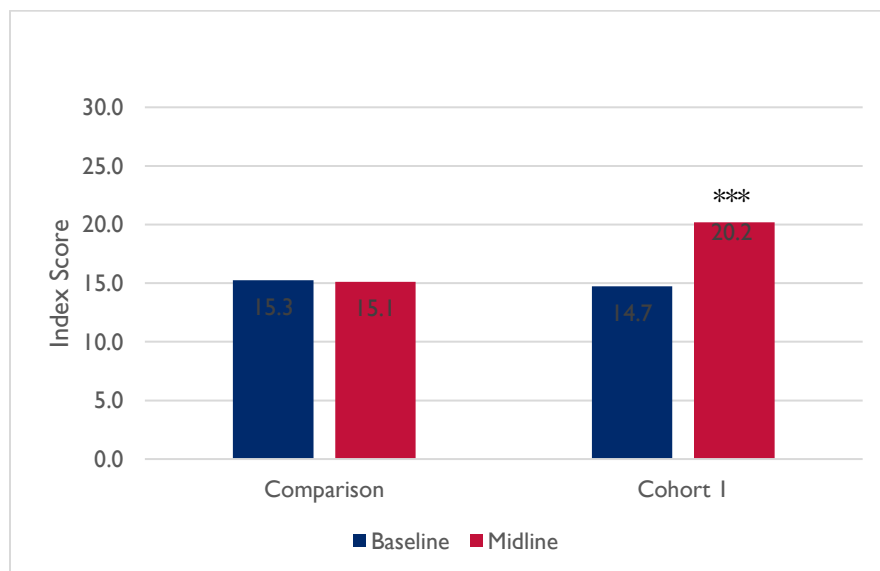
EQ4: To what extent has the EGRP Nepali LI program changed teachers' reading instruction practices in the classroom?	For cohort 1, the EGRP has had some impact on teachers' reading instruction practices at midline. The teaching reading instruction practice index shows a positive impact of the EGRP, although the more subjective teaching assessment done by the observer does not. Cohort 2 does not show any impact as expected given that teachers in these districts did not receive training.
---	--

We created two indexes to measure teachers' reading instruction practices in the classroom. The first index, which we call Teacher Reading Instruction Practices Index, includes 30 items describing desirable actions during an early grade reading lesson. We score each of them with one point if they were observed during the reading lesson; therefore, the index minimum is zero and its maximum is 30. The items included are the following:

- Did the teacher show how to pronounce sounds/letters/words/syllables correctly?
- Did students pronounce sounds/letters/words correctly?
- Did students practice reading/pronouncing sounds/letters/words separating?
- Did students practice reading/pronouncing sounds/letters/words put together?
- Did the teacher read text w/ proper sound/pattern/rhythm for students to listen?

- Did students have an opportunity to read alone/in pairs w/ proper sound/pattern/rhythm?
- Did the teacher introduce new vocabulary words or discuss meaning of vocabulary words?
- Did the teacher ask students to use vocabulary words in sentence/activity oral/write?
- Did the teacher have students answer question before/while reading/listening to text?
- Did the teacher ask students questions about read/listening text after text finished?
- Did comprehension questions include at least 1 question where answer not explicitly stated?
- Did the teacher make students read the text?
- Were students able to answer the questions asked based on the reading text?
- Did students have an opportunity to practice writing accuracy?
- Did students have an opportunity to do any original writing?
- Overall, did the teacher call on all students in the classroom?
- Overall, did the teacher call on, and respond to, boys and girls equally?
- Did the teacher use at least two different kinds of grouping?
- During the lesson, were most of the students primarily doing what the teacher asked?
- During the lesson, did more than half of the children volunteer to answer questions?
- If children were reading, the majority of children's eyes on the text as they read?
- If students responded correctly, did the teacher give them positive feedback?
- If students responded incorrectly, did the teacher give constructive feedback?
- Did the teacher use the instructional materials adequately?
- Were the materials used appropriately?
- During the lesson, did the teacher move around to monitor students work individually or in groups?
- Did the teacher use the teach model, guide & students practice (I do, we do, you do)?
- Did the teacher help students having difficulty w/ an activity individually/groups?
- During lesson, did the teacher do in/formal check of students' understanding/performance?
- Did the teacher provide an opportunity for students to ask questions/discuss ideas?

Figure 18: Teacher Reading Instruction Practices Index, Cohort I

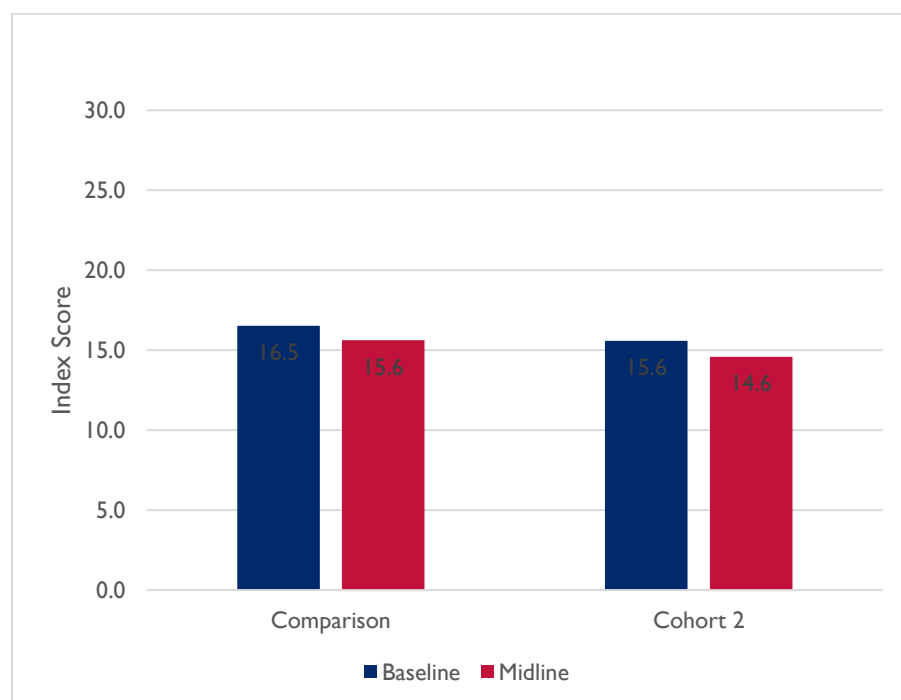


Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Figure 18 shows a **positive effect of the EGRP on the Teacher Reading Instruction Practices Index for cohort 1**. The net change in the index between baseline and midline is positive for cohort 1. The DiD estimate shows a large and statistically significant effect of 5.45 points in the index for cohort 1 teachers.

In contrast, **we do not find an effect of the EGRP on the Teacher Reading Instruction Practices Index for cohort 2**. This is not surprising given that the teachers in cohort 2 schools did not receive training. Figure 19 shows the index for cohort 2 and comparison teachers at baseline and midline. The scores are quite similar across groups and time.

Figure 19: Teacher Reading Instruction Practices Index, Cohort 2



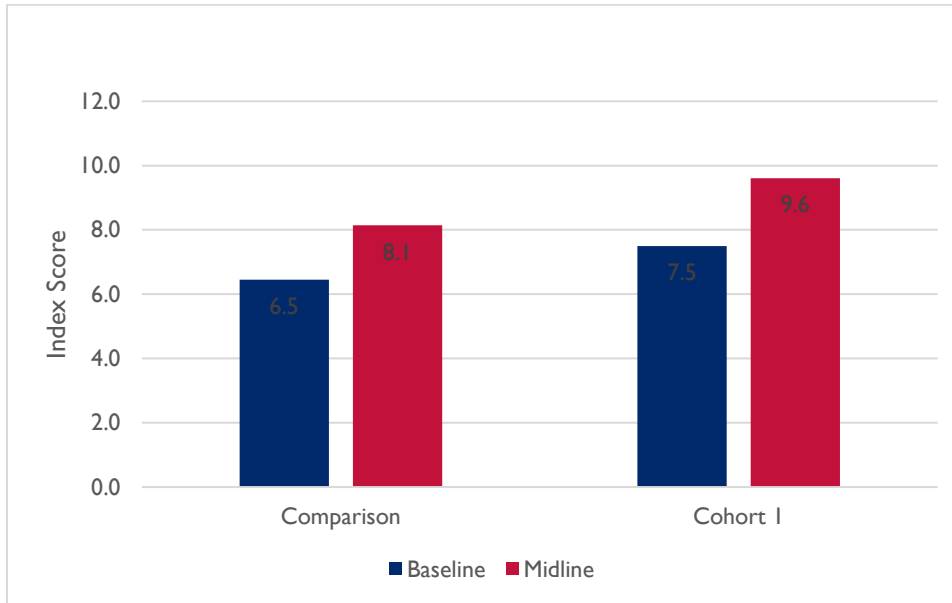
Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Lesson observers also assessed teachers' readiness and preparation, students' motivation to the lesson, students' participation and the use of child-centered pedagogical methods at the end of the class they observed. During the lesson these four areas were classified as good, so-so, and needing improvement by the observer. We assigned 3 points for a good score, 2 points for so-so qualification and 1 when improvement was needed; therefore, this index can oscillate from 4 to 12 points. We denominate this index Observer's Teaching Assessment. It is important to mention that we recommend (see Section 6) a different and more rigorous approach to assess the quality of teaching.

We show the Observer's Teaching Assessment Index for cohort 1 classroom observations and their comparison group in Figure 20. Classroom observers judged the quality of teaching in cohort 1 and their comparison group similarly. Both groups had some increase in the index over time. **No impact due to the EGRP program can be detected in the Observer's Teaching Assessment Index of cohort 1.**

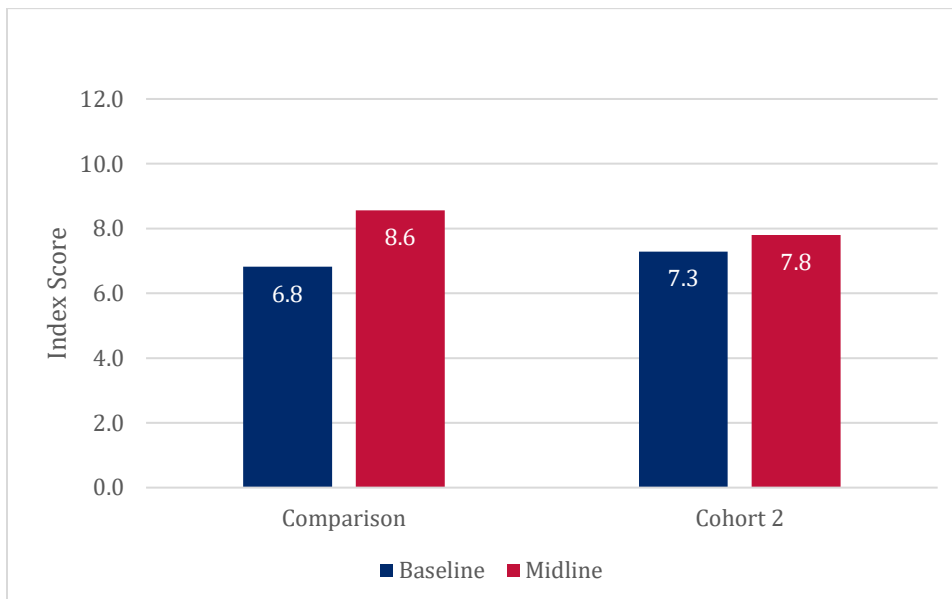
Similarly, Figure 21 shows the index for cohort 2 and its comparison group at baseline and midline. Not surprisingly given their lack of training, classroom observers assessed the quality of teaching similarly between cohort 2 and their comparison group. **We do not find that EGRP has improved the Observer's Teaching Assessment Index for cohort 2.** The net change for cohort 2 is in fact negative.

Figure 20: Observer's Teaching Assessment Index, Cohort 1



Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Figure 21: Observer's Teaching Assessment Index, Cohort 2



Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

In addition, teachers self-reported their approach to supporting students with reading difficulties, their attitudes about how students learn to read and about how to teach early grade reading, and if they assign reading homework to learners. In general, we do not find differences in the support provided, attitudes, or the assignment of reading homework between teachers in cohort 1 and teachers in their comparison group. While some items show the desired trends, others do the opposite and mostly nothing appears statistically significant. The cohort 2 findings are similar, and again, the result is not surprising given that teachers in that cohort had no training. The details are included in Annex IV.

3.3.3. Parental and Community engagement

Effect on School Leadership and Management

One of the components of the EGRP that aims to engage parents and community in the acquisition of early grade reading skills, consists of increasing the ability of the parent–teacher association and the school management committee (SMC) to contribute to quality reading instruction.

Figure 22: Training of School Management Committee Members

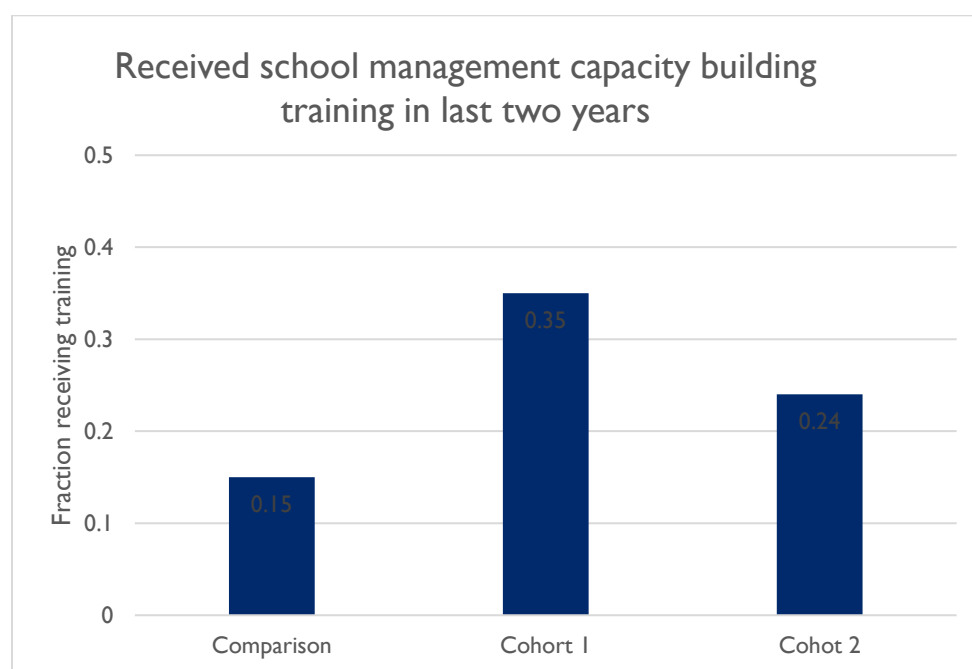


Figure 22 shows the percentage of SMC members that received training. In each school, the question was administered to the SMC members whose schools were selected for assessment. The proportion is higher in cohort 1 schools than in cohort 2 and comparison groups. Around a quarter of the SMC members in cohort 2 reported receiving training, while the proportion is over a third for cohort 1. None of these fractions is particularly high.

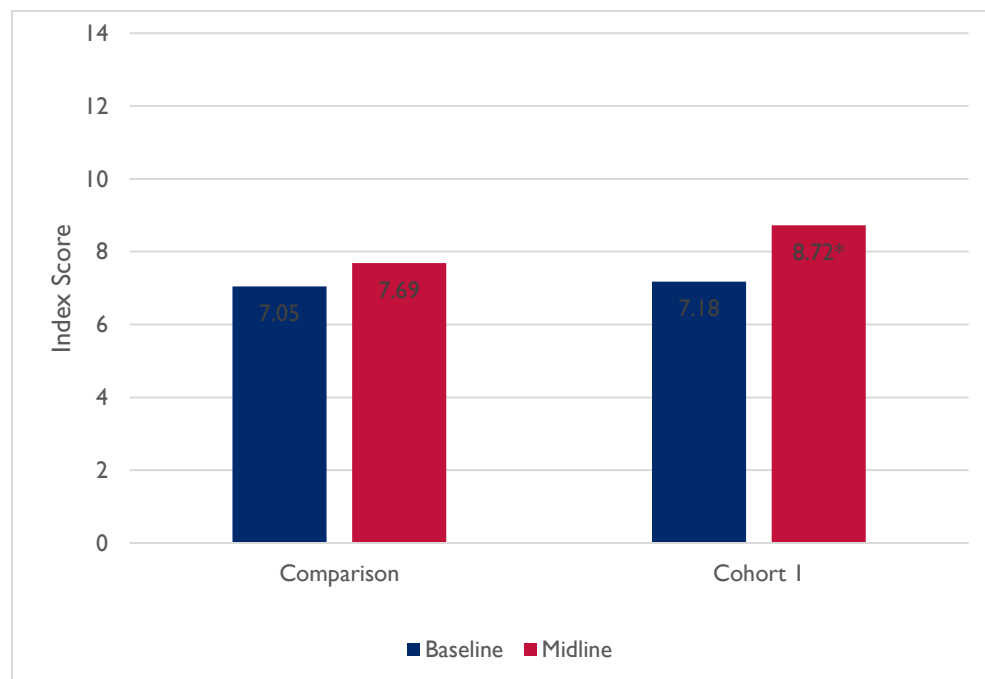
USAID/Nepal, the EGRP team and local stakeholders defined a School Leadership and Management Index. The index includes 14 items as follows:

- Number one mission of the school is to ensure quality education
- Number one purpose of Grade 2 learning is to achieve basic language/numeracy skills
- School provides guidance to parents to help their children become readers
- School has an active parent-teacher association (PTA)
- School prioritizes early grade reading
- School offers reading activities to promote initiatives or programs (from Head Teacher report)
- Reading or literacy are mentioned in the SIP
- School tracks number of students who are meeting reading/literacy standards
- School provide student report cards to parents
- Is there a book corner or classroom library?
- School offers initiatives designed to promote reading (from SMC member report)
- SMC meets frequently
- The head teacher shares with SMC information on student learning
- SMC member conducts supervisory visits

This information is collected through interviews with head teachers, SMC members and classroom observations, and each item weights equally, resulting in an index that goes from 0 to 14.

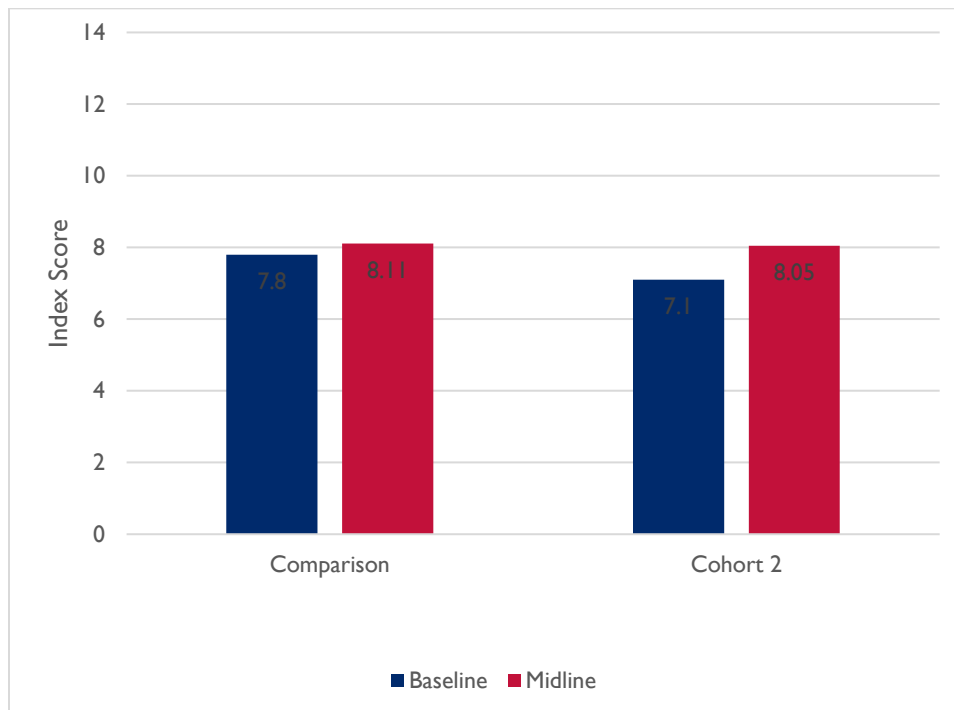
We show in Figures 23 and 24 the management index at baseline and midline for cohort 1 and its comparison group, and for cohort 2 and its comparison group, respectively. In cohort 1, the index was positively impacted by the program. The impact is 0.9 of a point or 6.4 percent, and it is statistically significant. In contrast, we find no impact of the EGRP on the management index among cohort 2 schools.

Figure 23: Management Index, Cohort 1



Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Figure 24: Management Index, Cohort 2



Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

This evidence allows us to answer the last evaluation question:

EQ5: To what extent has the EGRP Nepali LI program changed the school leadership and management index (as defined in monitoring index), demonstrating active support for EGR?	The EGRP Nepali LI program has generated a modest improvement of almost one point (out of 14) in the management index for cohort 1. However, there is no impact in the index among schools in cohort 2. Of note, only 35 and 25 percent of the school management committee members report receiving training.
---	---

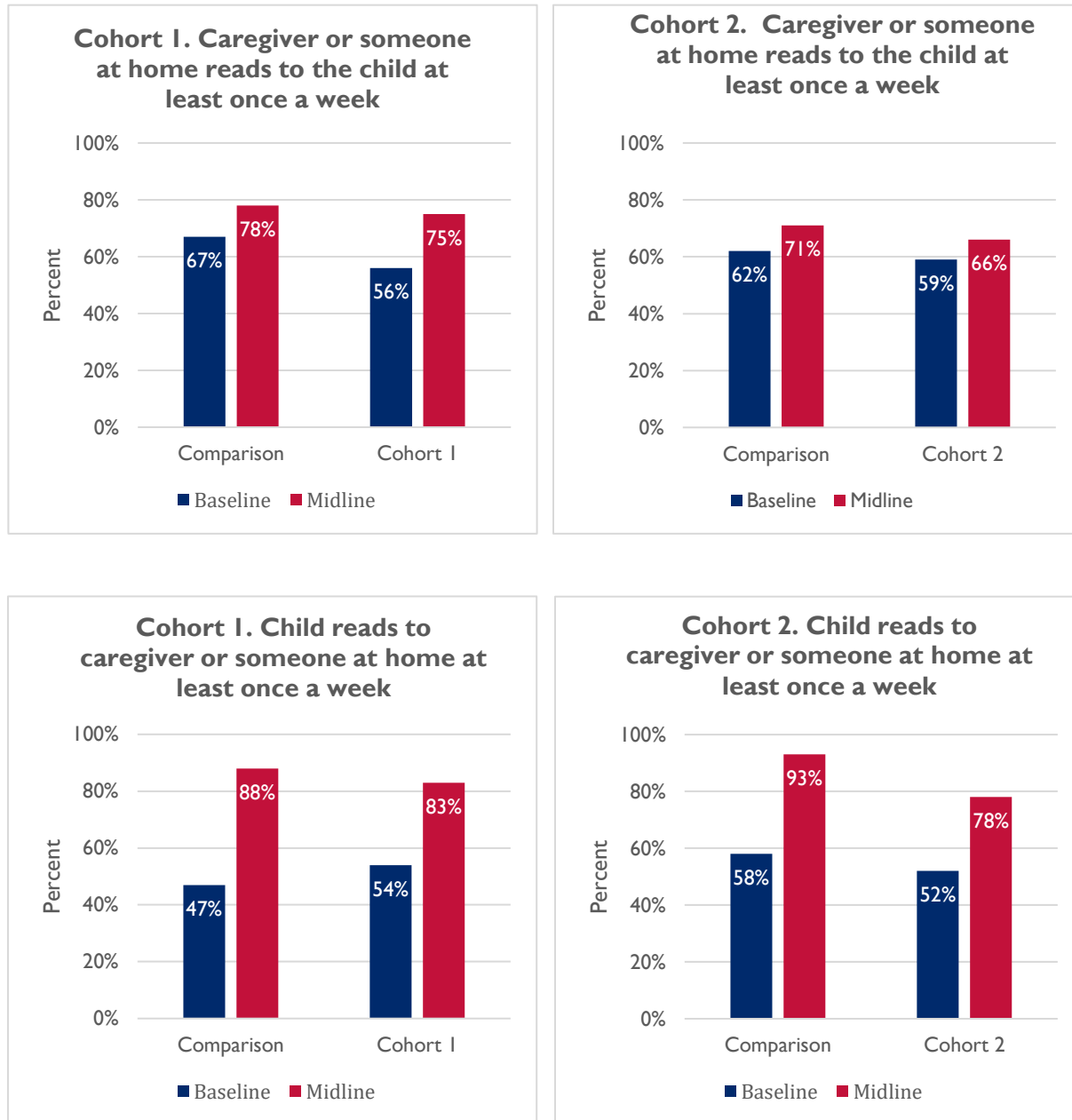
Effect on Parental Engagement and Awareness

The EGRP includes numerous activities devoted to parental and community engagement, such as reading camps and festivals, simple and low-cost reading materials development, support of print-rich school and classroom environment, parent conferences, etc.

We explore the effects of the EGRP on parental behavior related to reading with children at home. Parents report if they or someone else reads with their children at home and whether their child reads to them or to someone else at home. Figure 25 shows the averages at baseline and midline for cohort 1 and 2 and their corresponding comparison groups. For all groups, there is some increment between baseline and midline in the percentages of reading to child or listening the child at home at

least once a week. The EGRP, however, does not show an additional improvement over the comparison group on these indicators.

Figure 25: Reading at home



Note: Propensity score matching weights applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

4. CONCLUSIONS

The EGRP had positive and large effects at midline among learners in cohort 1, where the program was fully rolled-out. In contrast, we do not find an effect on learners in cohort 2 schools, where there was only a partial implementation of the program and important components, such as teacher training, were still pending at midline. We focus therefore mostly on cohort 1.

In cohort 1, both L1 and L2 learners benefited from the program. This is highly desirable given that the performance of both groups of students is far from the levels that the GoN considers to be the minimum reading standards. **However, the program tends to benefit L1 learners more than L2 learners.** As noted at baseline, there is a very large gap between L1 and L2 learners' reading skills. The gap is approximately the equivalent of one full year of schooling – for example, on average, L2 grade 3 learners perform at the level of L1 grade 2 learners. At midline, the program appears to be increasing the gap by having a larger impact on L1 than L2 learners. L2 learners not only lag behind L1 learners in terms of their reading skills, it is also clear that there is a serious deficiency in overall oral Nepali language comprehension among L2 learners.

The EGRP has benefited students with both low and high performance. It reduced the number of zero scores among learners and also increased the percentage of learners that reach the tentative benchmarks of 45 correct words per minute and 80 percent oral reading comprehension that the GoN is considering.

Examining the channels through which the program has functioned, there is no evidence that the program has led to changes in parents' at-home support for their children's reading development. School and SMC support for reading activities shows a very modest improvement but only around a third of the SMC members interviewed report receiving training.

There is mixed evidence that the program has had an effect on teachers' reading instruction, at least in the way captured by the classrooms observation exercise. While there is a positive effect on teachers' classroom practices, there is no effect on the teaching assessment as judged by observers. It is important to mention that we recommend in Section 6 a different and more rigorous approach to assess the quality of teaching.

The program has been quite successful at ensuring access to materials, including students' access to Nepali-language textbooks and workbooks, and additional children's reading materials, and teachers' access to teaching guidelines, materials, and curriculum. Almost all teachers report using these resources. Thus, it is **likely that the positive effects of the program have functioned via a combination of at least some improvement in teaching practices with broad access and use of learning and teaching materials.**

The lack of impact on parental engagement and teaching quality suggests there is still room for improving the community outreach and professional training portions of the program, and additional gains to learners' test score outcomes could still be seen if these program components begin to show more evidence of having the expected effects.

5. LIMITATIONS

Representativeness of the Sample. The sample is only representative of the districts where EGRP is being implemented. Findings and results are not generalizable at the national level or other geographical areas, within or outside Nepal, or to other languages.

Methodology. The DiD methodology assumes the treatment and comparison groups, in the absence of the program, would display the same trends or, in other words, would move in parallel. This is an assumption that we cannot verify directly. However, using matching to ensure that treatment and comparison groups are as alike as possible increases the probability that the groups' trajectories over time are identical. To further assure that the groups are as similar as possible and there is no bias, we also take into account the basic characteristics of the learners in the analysis and produce adjusted DiD. Finally, the lack of significant differences in the difference-in-differences estimates shown in Table 8 of the main text for Cohort 2, where most aspects of the treatment program had still yet to be implemented at midline, strongly supports the parallel trends assumption that, in absence of the EGRP, the treatment and control groups follow similar paths.

Sample size. Samples of parents, teachers, head teachers, SMC members, classroom inventory and observation, and school inventory are small. At midline there are a total of only 85 observations for cohort 1 and 86 for cohort 2 in each of those categories. This limits the type of analyses that can be done and the precision of estimates. In addition, it is not possible to link most of these data to particular students; for example, we cannot link a particular parent to a learner.

Data collection schedule. Baseline data collection started at the end of school year 2015-16 but only ended at the beginning of the following academic year, 2016-17, after a school break. In contrast, midline data collection took place at the end of the school year 2017-18. It is unlikely that this would make too much of a difference, but it should be kept in mind when comparing means between baseline and midline. There is no effect in terms of the evaluation as all groups –cohort 1, 2 and comparison- had the same data collection schedule.

6. RECOMMENDATIONS

A number of recommendations stem from our findings:

Special attention to L2 Learners: The disadvantage in early grade reading skills of L2 learners relative to L1 learners is evident, and the EGRP, while benefiting everyone, appears to widen the gap between the language groups. The situation not only negatively affects the L2 population, but might also have long-lasting consequences in terms of economic development and growth and social cohesion. Special attention should be devoted to better support non-Nepali speakers in the crucial early years of their schooling. At a minimum, teachers need basic training to acquire the skills needed to provide effective reading instruction for Nepali language learners in their classrooms.

Improve teacher support supervision: Cohort I teachers received frequent support supervision from one or more sources; however, the levels of satisfaction with this support are not high. We recommend exploring why teachers do not find these interactions useful and, in the light of the findings, adapt and redesign a comprehensive and coordinated systematic support supervision plan.

Review implementation. A substantial percentage of SMC members in cohort I report not receiving any kind of training. Some classrooms do not have textbooks, workbooks and/or supplementary reading materials. We recommend systematically confirming all schools in cohort I received the full EGRP package.

Finally, we have a recommendation for future data collection work that may enhance our ability to study the effect of the EGRP on teacher practices.

Data Collection: we recommend a revision of the classroom observations approach. The uptake of program principles by the teachers is a fundamental aspect to make it effective. The main goal of the classroom observations is to evaluate teachers' reading instruction practices. Classroom observations were conducted by enumerators, one per classroom. The first set of teaching practices recorded by enumerators consists largely of low inference (presence / absence) items that tend to be objective and ensure comparability of data across classrooms, although cannot provide the depth that rich narratives of the lessons can offer. On the other hand, the second set of teacher observations relates to teaching quality and requires the judgment of the observer.

We recommend that – at least in some subsample of schools – enumerators work in pairs during the classroom observation, each capturing a detailed narrative of the lesson, and that they complete the second set of items together after the observation, so that any judgements required could be verified at the point of data collection. These judgements could also be confirmed by reference to the lesson narratives. Finally, enumerator training to conduct classrooms observations needs to be intensive and conducted by experts.

7. REFERENCES

RTI International. 2015. Early Grade Reading Assessment (EGRA) Toolkit, Second Edition. Washington, DC: United States Agency for International Development

UNESCO, 2015, Education for All, National Review Report, Kathmandu, NEPAL July 2015

8. ANNEXES

ANNEX I: EVALUATION STATEMENT OF WORK

STATEMENT OF WORK

Impact Evaluation of Early Grade Reading Project (EGRP)

PURPOSE OF THE EVALUATION

The main purpose of the IE will be to assess the causal impact of EGRP-Nepal on reading outcomes of primary school children who speak Nepali as a first language (L1 learners) and children who do not speak Nepali as a first language (L2 learners). The evaluation will measure reading outcomes using subtasks of the Early Grade Reading Assessment (EGRA), a widely used tool to measure various aspects of reading proficiency. Furthermore, the evaluation will examine intermediate outcomes related to teacher and school management knowledge, attitudes and behaviors, as measured by the Education Management Efficiency Study (EMES). The evaluation's key audiences and stakeholders include USAID, government of Nepal, RTI, the donor community and NGOs operating in Nepal. The evaluation findings will be used to inform programmatic decisions and funding allocations, among other purposes. In addition, findings from this evaluation will contribute to the knowledge base on what works in improving early grade literacy in linguistically complex settings.

SUMMARY INFORMATION

Strategy/Project/Activity Name	<i>Early Grade Reading Project (EGRP)</i>
Implementer	<i>RTI International</i>
Cooperative Agreement/Contract #	<i>AID-367-TO-15-00002</i>
Total Estimated Ceiling of the Evaluated Project/Activity(TEC)	<i>\$53,870,553</i>
Life of Strategy, Project, or Activity	<i>March 2, 2015 – March 1, 2020</i>
Active Geographic Regions	<i>Dang, Bardiya, Dadeldhura, Parsa, Rupandehi, Dolpa, Dhanusa, Surkhet, Mustang, Kailali Saptari, Manang, Banke, Kanchanpur, Kaski and Bhaktapur</i>
Development Objective(s) (DOs)	<i>DO 3 – Increased Human Capital</i>
USAID Office	<i>Education Office</i>

NORC at the University of Chicago, through the USAID Reading and Access Evaluation Contract, has been charged with conducting the external impact evaluation (IE) of the Early Grade Reading Program (EGRP) in Nepal.

BACKGROUND

Description of the Problem, Development Hypothesis, and Theory of Change

In 2014, USAID supported a nationally representative Early Grade Reading Assessment, which provided concrete data on the foundational reading skills of Nepali children. The assessment found that 34 percent of second graders and 19 percent of third graders could not read a single word of Nepali.

Students in the Terai had both the lowest mean oral reading fluency score and the highest zero scores compared to other regions of Nepal and were, on average, reading 12 correct words per minute fewer than students in the Kathmandu Valley. Moreover, students who reported speaking Nepali at home performed better than students speaking another first language. USAID's Early Grade Reading Program (EGRP) in Nepal was designed to improve the reading skills of Nepali students.

The **goals** to be achieved by the conclusion of this five-year task order are:

Reading skills improved: Public primary school students in grades 1-3 in the 16 target districts with improved reading skills

GON service strengthened: The Contractor will have supported the Government of Nepal through Phase One of the NEGRP and completed the design and demonstration of a national model that the Government of Nepal can then implement nationwide within its budget.

To achieve these goals, the Contractor must implement activities aligned with the following **intermediate results (IR)**:

Improved Early Grade Reading Instruction (IR 1)

Improved National and District Early Grade Reading Service Delivery (IR2)

Increased Family and Community Support for Early Grade Reading (IR3)

EGRP GOAL: Public primary students in the early grades (1-3) in the 16 target districts with improved reading skills.

IR 1: Improved Early Grade Reading Instruction

Sub IR 1.1 Evidence-based early grade reading instructional materials designed, distributed, and in use

Sub IR 1.2 In-service professional development for teachers in public schools on reading instruction and the use of these materials provided

Sub IR 1.3 Monitoring and coaching for teachers in early grade reading instruction provided

Sub IR 1.4 Classroom- based and district-based early grade reading assessment processes improved

IR 2: Improved National and District Early Grade Reading Service Delivery

Sub IR 2.1 Early grade reading (EGR) data collection and analysis systems improved.

Sub IR 2.2 Policies, standards, and benchmarks that support improved early grade reading instruction institutionalized.

Sub IR 2.3 Planning and management of financial, material, and human resources devoted to early grade reading improved.

Sub IR 2.4 National standards for early grade reading improvement adopted and implemented in EGRP districts.

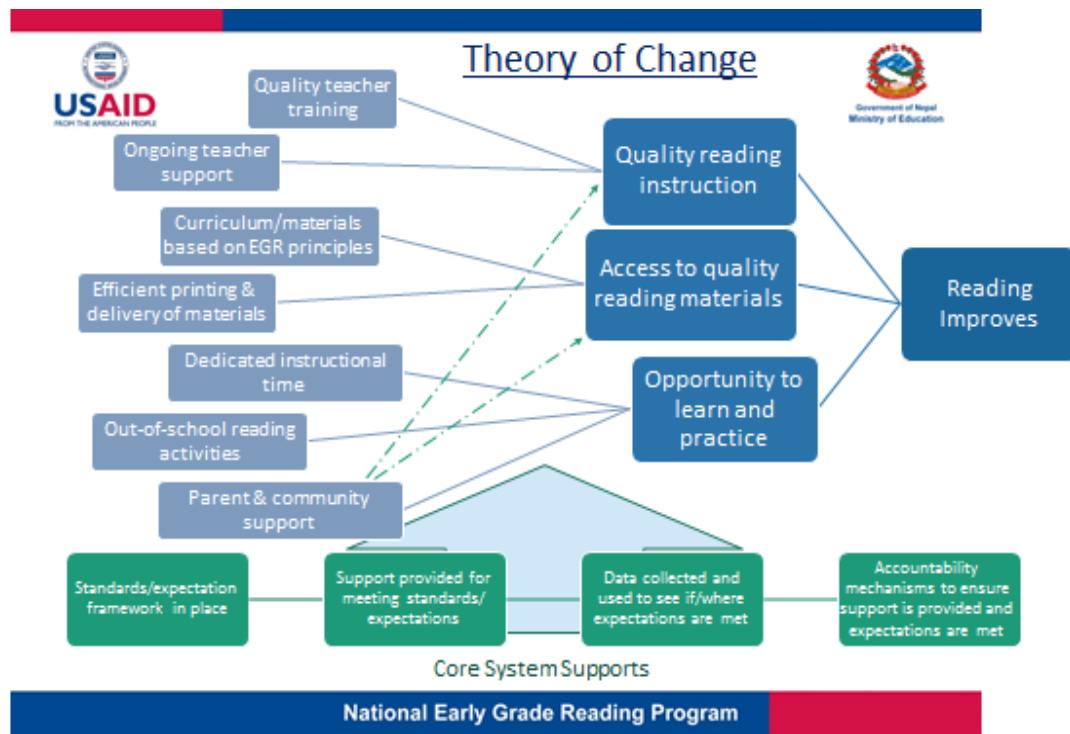
IR 3: Increased Family and Community Support for Early Grade Reading

Sub IR 3.1 Community awareness of the importance of reading and language for reading instruction as appropriate increased.

Sub IR 3.2 Family engagement to support reading increased

Sub IR 3.3 Parent Teacher Association/School Management Committee ability to contribute to quality reading instruction increased

Sub IR 3.4 Parent and community capacity to monitor reading progress increased



The Early Grade Reading Program aims to achieve the following objectives:

Increase the proportion of grade 1–3 public primary students who can read and understand grade-level text.

Improve national and district early grade service delivery by completing the design and demonstration of an evidence-based reading model which the Ministry of Education can feasibly replicate and scale up nationally.

Increase family and community support for early grade reading.

Summary Strategy/Project/Activity/Intervention to be evaluated

USAID/Nepal hypothesizes that implementing EGRP (Nepali L1) will improve the reading skills of L1 and L2 learners. However, implementing EGRP with accommodations for second language learners (Nepali L2 with and without mother tongue (MT) reading instruction), will improve the reading skills of L2 learners even more than under EGRP (Nepali L1) program. The EGRP-Nepal focuses on grades 1, 2, and 3 and will be rolled out in 2 cohorts of districts. cohort 1 includes 6 districts (Saptari, Bhaktapur, Kanchanpur, Banke, Manang, Kaski) while cohort 2 covers 10 districts (Dang, Bardiya, Dadeldhura, Parsa, Rupandehi, Dolpa, Dhanusa, Surkhet, Mustang, Kailali). Under cohort 1, all students (regardless of language) are currently receiving the EGRP (Nepali L1) package and will continue to do so in the second year. The teacher coaching, mentoring and support model is currently implemented through reading motivators (RMs) who are teachers or resource persons within the GON system). The treatment of cohort 2 will start in April 2018. There are ongoing discussions with the GON to determine if L2 learners could receive additional EGRP interventions (Nepali L2 with of without MT reading instruction) in 2019. Furthermore, the teacher coaching model for cohort 2 will change from the current RM

modality, where the head teachers and/or primary in charge would provide regular mentoring and coaching and regular teacher cluster meetings would be held.

Summary of the Project/Activity Monitoring, Evaluation, and Learning (MEL) Plan

USAID can share the EGRP Monitoring, Evaluation and Learning (MEL) Plan, which includes performance monitoring indicators and indicator reference sheets, as well as the EGRA and EMES conducted in 2014. The EGRP M&E team will also share its monitoring database –or the relevant parts of it- with NORC (at a later time).

EVALUATION QUESTIONS

The main purpose of the IE will be to assess the causal impact of EGRP-Nepal on reading outcomes of primary school children. Specifically, the IE will answer:

A. To what extent did EGRP (Nepali L1 program) improve the reading outcomes of pupils who speak Nepali as a first language (L1 learners) in cohorts 1 and 2?

B. To what extent did EGRP (Nepali L1 program) improve the reading outcomes of pupils who speak Nepali as a second language (L2 learners) in cohorts 1 and 2?

Note: Only to be answered should the GoN decide to move forward on any Nepali L2 interventions.

C. To what extent did the EGRP (Nepali L2 intervention) improve the reading outcomes of pupils who speak Nepali as a second language (L2 Learners) in cohort 2?

Additional Questions about Intermediate Outcomes:

To what extent has the EGRP Nepali L1 program changed teachers' reading instruction practices in the classroom?

To what extent has EGRP Nepali L1 program changed the school leadership and management index (as defined in monitoring index), demonstrating active support for EGR?

Questions	Suggested Data Sources (*)	Suggested Data Collection Methods	Data Analysis Methods
A. To what extent did EGRP (Nepali L1 program) improve the reading outcomes of pupils who speak Nepali as a first language (L1 learners) in cohorts 1 and 2?	EGRA	Assessment	The impact of the program is estimated by comparing the average outcomes of the treatment group and the average outcome among a statistically matched control subgroup of

Questions	Suggested Data Sources (*)	Suggested Data Collection Methods	Data Analysis Methods
B. To what extent did EGRP (Nepali L1 program) improve the reading outcomes of pupils who speak Nepali as a second language (L2 learners) in cohorts 1 and 2?	EGRA	Assessment	<i>schools. The exact econometric approach to the comparison will be decided once we have the data and can then assess the different possibilities.</i>
<i>Note: Only to be answered should the GoN decide to move forward on any Nepali L2 interventions.</i> C. To what extent did the EGRP (Nepali L2 intervention) improve the reading outcomes of pupils who speak Nepali as a second language (L2 Learners) in cohort 2?	EGRA	Assessment	<i>The evaluation can try to measure the potential additional effect that the EGRP Nepali L2 intervention might have on L2 learners over the EGRP Nepali L1 program effects.</i>

It was decided by EGRP-Nepal, the MOE and USAID/Nepal that all schools within a treatment district would receive EGRP interventions and, therefore, control schools would necessarily need to be found in other districts. A group of control districts was selected by RTI to match the characteristics of the treatment districts in general. The dimensions that were taken into account for the selection were landscape/climate, socio-cultural settings, and economic activity. The selected control districts to match cohort 1 treatment districts are: Doti, Myagdi, Kapilvastu, Bara, Sunsari, and Kavrepalanchowk.

Impact Evaluation Plans

The evaluation will use a quasi-experimental approach to evaluate EGRP-NEPAL. The implementer will collect the data to be used in the IE. The data collection schedule is as follows:

Baseline: It was originally planned to complete all data collection in February/March of the school year 2015-16. However, collection was interrupted due to exams and it was finalized in April/May school year 2016-17

Midline: End of school year 2017-18

Endline: End of school year 2019-20

Midline data collection:

EGRP will conduct a workshop with GoN to gain their support regarding midline data collection, tools, approach, etc.

Midline will include all the same schools visited during baseline in cohort 1, cohort 2 and Control Districts. A first check of schools will be done before going to the field to see if additional schools need to be included.

CAMRIS will conduct QA, participating in instruments pre-tests/adaptation, enumerator training, piloting, data collection fieldwork, and data cleaning. (SOW to be reviewed by USAID/Nepal and NORC).

Instruments. The instruments to be used are EGRA, student survey, teacher survey, and head-teacher survey. The evaluator will review data collection instruments and make recommendations for modifications, if needed.

Cohort 1 of EGRP includes six districts of the country: Saptari, Manang, Banke, Kanchanpur, Kaski and Bhaktapur. The evaluator will select control districts to match cohort 1 treatment districts.

Cohort 2 includes the following districts: Dang, Bardiya, Dadeldhura, Parsa, Rupandehi, Surkhet, Dolpa, Mustang, Dhankuta, and Kailali.

Sample Size

The implementer calculated the sample size. The sample was selected such that the impact of EGRP-Nepal will be measured at the cohort level and not at the district level. The original calculation used the following assumptions:

Grade 2 mean: 15wpm, SD = 28wpm (based on previous studies)

Grade 3 mean: 28wpm, SD = 24wpm (based on previous studies)

The ICC for the school clusters = 0.25

Power = 80%

MDE=6 wpm per year

Based on those parameters the sample size was estimated as 86 treatment schools in cohort 1 and cohort 2, with 10 students per grade, from grades 1-3 in each school (amounting to 30 students per school and 2,580 students in total); and 90 control schools, with 10 students each from grades 1-3 per school (for a total of 2,700 students in total).

DELIVERABLES AND REPORTING REQUIREMENTS

Evaluation Work plan: Within 4 weeks of the agreed-upon evaluation scope of work, a draft work plan for the evaluation shall be completed by the lead evaluator and presented to the Contracting Officer's Representative (AOR/COR). The work plan will include: (1) the anticipated schedule and logistical arrangements; and (2) a list of the members of the evaluation team, delineated by roles and responsibilities.

Evaluation Design: Within 2 weeks of the agreed-upon evaluation scope of work, the evaluation team must submit to the Agreement Officer's Representative/Contracting Officer's Representative

(AOR/COR) an evaluation design (which will become an annex to the Evaluation report). The evaluation design will include: (1) a detailed evaluation design matrix that links the Evaluation Questions in the SOW to data sources, data collection methods (i.e. test/survey administration procedures), and the data cleaning and analysis plan; (2) draft questionnaires and other data collection instruments or their main features; (3) the list of potential interviewees and sites to be visited and proposed selection criteria and/or sampling plan (must include calculations and a justification of sample size, plans as to how the sampling frame will be developed, and the sampling methodology); (4) known limitations to the evaluation design; and (5) a dissemination plan.

USAID offices and relevant stakeholders will take up to *10 business days* to review and consolidate comments through the AOR/COR. Once the evaluation team receives the consolidated comments on the initial evaluation design and work plan, they are expected to return with a revised evaluation design and work plan within *10 business days*.

Mid-term Briefing and Interim Meetings: The Mission and/or USAID/Washington may request that the evaluation team hold a mid-term briefing with Mission and USAID/Washington staff on the status of the evaluation, including potential challenges and emerging opportunities. The team will also provide the evaluation COR/manager with periodic briefings and feedback on the team's findings, as agreed upon during the in-briefing. Weekly briefings by phone may be conducted.

Final Presentation: The Mission and/or USAID/Washington may request that the evaluation team hold a final presentation in person or by virtual conferencing software to discuss the summary of findings and recommendations to USAID. This presentation will be scheduled as agreed upon during the in-briefing.

Draft Baseline, Midline and Endline Evaluation Report: The draft baseline, midline and endline evaluation reports should be consistent with the guidance provided by the contract COR. The report will address each of the questions identified in the SOW and any other issues the team considers having a bearing on the objectives of the evaluation. Any such issues can be included in the report only after consultation with USAID. The submission date for the draft evaluation report will be determined in the evaluation work plan. The draft evaluation report should be submitted within 60 days of transferring collected data to NORC. Once the initial draft evaluation report is submitted, USAID/Nepal and USAID/Washington will have 10 business days in which to review and comment on the initial draft, after which point the AOR/COR will submit the consolidated comments to the evaluation team. The evaluation team will then be asked to submit a revised final draft report 10 business days hence, and again the USAID/Nepal and USAID/Washington will review and send comments on this final draft report within 5 business days of its submission.

Final Evaluation Report: The evaluation team will be asked to take no more than 5 business days to respond/incorporate the final comments from the USAID/Nepal and USAID/Washington. The evaluation team leader will then submit the final report to the AOR/COR. All project data and records will be submitted in full and should be in electronic form in easily readable format, organized and documented for use by those not fully familiar with the intervention or evaluation, and owned by USAID.

EVALUATION TEAM COMPOSITION

All team members will be required to provide a signed statement attesting to a lack of conflict of interest or describing any existing conflict of interest.

The evaluation team shall demonstrate familiarity with USAID's evaluation policies and guidance included in the USAID Automated Directive System (ADS) in Chapter 200.

The expected roles/responsibilities of IE team vis-a-vis IP are as follows:

The IP (RTI International) will be responsible for collecting primary data following agreed-upon plans and procedures and using agreed-upon tools from an agreed-upon sample of beneficiaries. Data will be collected within timeframe specified in the activity work plan. The IP will also be responsible for data processing and cleaning. Cleaned data will be transferred to NORC for analysis. Additional data from the IP's M&E system will also be transferred. NORC will perform analyses of all provided data, prepare final evaluation report and present findings in person or via conference. Additional dissemination activities might be agreed upon, as well.

EVALUATION SCHEDULE

Baseline: It was originally planned to complete all data collection in February/March of the school year 2015-16. However, collection was interrupted due to exams and it was finalized in April/May school year 2016-17

Midline: End of school year 2017-18

Endline: End of school year 2019-20

Schedule

Timing (Anticipated Months or Duration)	Proposed Activities	Important Considerations / Constraints
Sept-Oct 2017	Preparation of the work plan and evaluation design	
Nov 2017	USAID review of the work plan and evaluation design	
Midline: End of school year 2017-18; Endline: End of school year 2019-20	Data Collection	
Midline: June 2018; Endline: June 2020	Data Analysis	
Midline: July 2018; Endline: July 2020	Report writing	
Midline: August 2018; Endline: August 2020	USAID review of Draft Report	
Midline: September 2018; Endline: September 2020	Incorporate USAID comments and prepare Final Report	

FINAL REPORT FORMAT

The evaluation final report should follow the template provided by Reading and Access Evaluation contract and be aligned with **ADS 201 mah, USAID Evaluation Report Requirements**.

The executive summary should be 2–5 pages in length and summarize the purpose, background of the project being evaluated, main evaluation questions, methods, findings, conclusions, and recommendations and lessons learned (if applicable).

The evaluation methodology shall be explained in detail in an Annex, with the summary in the main report. Limitations to the evaluation shall be disclosed in the report, with particular attention to the limitations associated with the evaluation methodology (e.g., selection bias, recall bias, unobservable differences between comparator groups, etc.)

The annexes to the report shall include:

Any details of data analyses that were not included in the main report.

The Evaluation SOW;

All data collection and analysis tools used in conducting the evaluation, such as questionnaires, checklists, and discussion guides;

All sources of information, properly identified and listed; and

Signed disclosure of conflict of interest forms for all evaluation team members, either attesting to a lack of conflicts of interest or describing existing conflicts of.

Any “statements of difference” regarding significant unresolved differences of opinion by funders, implementers, and/or members of the evaluation team.

Summary information about evaluation team members, including qualifications, experience, and role on the team.

In accordance with ADS 201, the contractor will make the final evaluation reports publicly available through the Development Experience Clearinghouse within three months of the evaluation’s conclusion.

CRITERIA TO ENSURE THE QUALITY OF THE EVALUATION REPORT

Per **ADS 201mah, Criteria to Ensure the Quality of the Evaluation Report**, draft and final evaluation reports will be evaluated against the following criteria to ensure the quality of the evaluation report.⁶

Evaluation reports should represent a thoughtful, well-researched, and well-organized effort to objectively evaluate the strategy, project, or activity.

Evaluation reports should be readily understood and should identify key points clearly, distinctly, and succinctly.

The Executive Summary of an evaluation report should present a concise and accurate statement of the most critical elements of the report.

Evaluation reports should adequately address all evaluation questions included in the SOW, or the evaluation questions subsequently revised and documented in consultation and agreement with USAID.

⁶ See ADS 201mah, USAID Evaluation Report Requirements and the Evaluation Report Review Checklist from the Evaluation Toolkit for additional guidance.

Evaluation methodology should be explained in detail (in an Annex, with a summary in the main body of the report) and sources of information properly identified.

Limitations to the evaluation should be adequately disclosed in the report, with particular attention to the limitations associated with the evaluation methodology (selection bias, recall bias, unobservable differences between comparator groups, etc.).

Evaluation findings should be presented as analyzed facts, evidence, and data and not based on anecdotes, hearsay, or simply the compilation of opinions.

Findings and conclusions should be specific, concise, and supported by strong quantitative or qualitative evidence.

If evaluation findings assess person-level outcomes or impact, they should also be separately assessed for both males and females.

If recommendations are included, they should be supported by a specific set of findings and should be action-oriented, practical, and specific.

OTHER REQUIREMENTS

In addition to the midline and endline reports, the evaluator will prepare dissemination materials, such as study briefs, presentations of midline and endline findings, and other products for communicating evaluation findings to study stakeholders. Dissemination materials should be written in a lay person language and be visually engaging.

All quantitative data collected by the evaluation team must be provided in machine-readable, non-proprietary formats as required by USAID's Open Data policy (see ADS 579). The data should be organized and fully documented for use by those not fully familiar with the project or the evaluation. USAID will retain ownership of the data collection tools and all datasets developed.

All modifications to the required elements of the SOW of the contract/agreement, whether in technical requirements, evaluation questions, evaluation team composition, methodology, or timeline, need to be agreed upon in writing by the COR. Any revisions should be updated in the SOW that is included as an annex to the Evaluation Report.

LIST OF ANNEXES

EGRP PMP
EGRA 2014
EMES 2014
EGRP – EGRA and EMES 2016 working papers
EGRP Baseline data collection report

ANNEX II: EVALUATION METHODS

Conditions. Most decisions about the roll-out of the program were made before NORC was invited to design the evaluation methodology, which therefore limited the range of methodological approaches that could be used for the IE.

EGRP-Nepal, the MOE and USAID/Nepal decided that all public schools in cohort 1 and cohort 2 districts –which we call treatment districts- would receive the EGRP interventions. Therefore, the IE takes a quasi-experimental approach. Cohort 1 includes 6 districts (Banke, Bhaktapur, Saptari, Kanchanpur, Kaski, and Manang) while cohort 2 covers 10 districts (Dhankuta, Parsa, Rupandehi, Dang, Bardiya, Surkhet, Dolpa, Kailali, Dadeldhura, and Mustang).

A group of comparison districts was selected by EGRP to match the characteristics of the treatment districts in general. The dimensions that were taken into account for the selection were landscape and climate, socio-cultural settings, and economic activity. The selected control districts to match treatment districts are: Doti, Myagdi, Kapilvastu, Bara, Sunsari, and Kavre.

Approach. NORC uses a quasi-experimental approach to evaluate EGRP-NEPAL, combining Difference-in-Difference (DiD) analysis and matching methods.

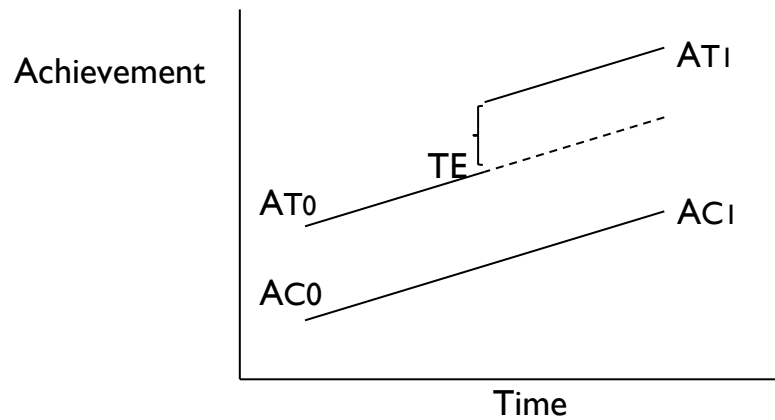
Identifying a credible comparison group is a critical aspect of an impact evaluation and there are several approaches to do so. Our impact evaluation is based on quasi-experimental methods where a comparison group is formed by statistical methods, rather than by random assignment.

First, NORC uses techniques to match comparison schools and treatment schools in each cohort. The goal is to select the schools from the control districts that best match in terms of characteristics the schools in the treatment districts. The matching is done taking into account language spoken by learners, learners' performance at baseline, school characteristics, etc. We include the details of the matching approach in Annex III.

The impact of the program is then estimated by comparing the average outcomes of the treatment group and the average outcome among a statistically matched control subgroup of schools. The NORC evaluation team conducts a Difference-in-Difference (DiD) analysis. This method involves comparing the changes between baseline and midline or endline test scores in treatment schools to changes between baseline and midline or endline test scores in comparison schools.

A graphical representation of the methodology is depicted by Figure 1 below.

Figure A2.1. Difference in difference estimator



Where:

A_{T0} is the average test score for a given grade at baseline in the treatment group

A_{C0} is the average test score for a given grade at baseline in the comparison group

A_{T1} is the average test score for a given grade at mid/endline in the treatment group

A_{C1} is the average test score for a given grade at mid/endline in the comparison group

and TE is the treatment effect

The idea behind the DiD method is to eliminate the differences that the treatment and comparison groups may have and that are constant overtime. As it is clear from the figure, the baseline levels do not need to be the same.

The DiD approach assumes that, in absence of the treatment, the two groups of schools would evolve in the same way; this is they follow parallel trends as shown in the figure in terms of the figure a (parallel trends). This is an assumption that we cannot verify. Using matching to ensure that treatment and comparison groups are as alike as possible, increases the probability that the groups' trajectories over time are identical.

Finally, to further assure that the groups are as similar as possible and that there is no bias, we take into account the basics characteristics of the learners in the analysis and produce adjusted DiD. To do so, we produced the analysis for L1 (Nepali) and L2 (Non-Nepali) learners separately and we included gender and age of the students.

ANNEX III: MATCHING PROCEDURE

This annex details the methodology and steps used to select the covariates and the matching algorithm to match treated and control schools using the Nepal EGRA baseline database. The steps described below are performed separately for both cohort 1 and comparison schools, and for cohort 2 and comparison schools. We illustrate the process by focusing on the matching results between cohort 1 and comparison schools. We show final baseline balance for cohort 1 and its comparison group at the end of the Annex.

Table A3.1 presents the mean, standard error, minimum and maximum of selected school characteristics available to do the matching process. As the table shows, the variables available for the matching come from the school administrative data, classroom observation, head teacher interview, and student assessments. Because only one teacher and one parent was interviewed per school, characteristics from the teacher and parent surveys were not used for the matching.

Table A3. 1: Mean, Standard Deviation, Min., Max, and observations.

Variable	Treatment – Cohort 1					Control				
	Mean	S.E	Min	Max	N	Mean	S.E	Min	Max	N
Total Enrollment	367.7	468.2	0	3243	86	226.2	236.8	0	1684	120
Enrollment Grade 1	31.3	23.2	0	123	86	29.1	29.3	0	195	120
Enrollment Grade 2	26.0	17.8	0	91	86	21.6	18.9	0	113	120
Enrollment Grade 3	27.5	19.0	0	106	86	22.1	19.0	0	134	120
Teachers Grade 2	4.1	2.3	0	16	86	4.4	1.9	0	14	120
Classrooms in Grade 1 (1+)	0.6	0.5	0	1	86	0.8	0.4	0	1	120
Classrooms in Grade 2 (1+)	0.6	0.5	0	1	86	0.8	0.4	0	1	120
Classrooms in Grade 3 (1+)	0.6	0.5	0	1	86	0.8	0.4	0	1	120
Nepali speakers % range	3.0	1.6	1	5	86	2.9	1.8	1	5	120
Classroom Grade observed	2.1	0.3	2	3	86	2.0	0.2	2	3	120
Number of girls present in classroom	7.9	5.2	0	28	85	7.6	10.0	0	92	120
Grade 2 is mono-grade classroom	0.5	0.5	0	1	86	0.4	0.5	0	1	120
Teacher Assistant literacy instruction	0.2	0.4	0	1	85	0.2	0.4	0	1	120
Guidance to parents to help children become readers	0.7	0.4	0	1	85	0.8	0.4	0	1	120
Ask parents to help with homework	0.9	0.3	0	1	85	1.0	0.2	0	1	119
Active parent-teacher association	0.7	0.5	0	1	85	0.7	0.5	0	1	120
School has improvement plan	0.8	0.4	0	1	85	0.8	0.4	0	1	120
Annual program and budget	0.6	0.5	0	1	85	0.6	0.5	0	1	120
School has library facility	0.4	0.5	0	1	85	0.5	0.5	0	1	120
School provides report cards to parents	0.6	0.5	0	1	85	0.4	0.5	0	1	120
School has annual report and social audit	1.0	0.7	0	3	84	1.2	0.7	0	3	120
Number of working computers in school	4.7	9.8	0	65	85	2.5	5.6	0	32	120
School have electricity	0.6	0.5	0	1	84	0.5	0.5	0	1	119
Source of water: Tap	0.3	0.4	0	1	84	0.6	0.5	0	1	119
Average Matra score grade 1	3.6	5.2	0	27.5	86	4.3	6.1	0	23.3	120
Average Matra score grade 2	9.8	9.2	0	35.1	86	12.9	11.5	0	45.7	120
Average Matra score grade 3	17.3	12.0	0	51.3	86	21.9	15.6	0	67.0	120
Average oral reading score grade 1	1.8	3.1	0	15.5	86	2.4	4.2	0	18.8	120
Average oral reading score grade 2	7.3	7.8	0	29.7	86	9.4	9.5	0	36.6	120
Average oral reading score grade 3	13.7	10.9	0	52.5	86	17.7	14.1	0	52.9	120

Variable	Treatment – Cohort I					Control				
	Mean	S.E	Min	Max	N	Mean	S.E	Min	Max	N
Average number of assets at home	4.9	1.2	2.8	7.6	86	5.1	1.1	2.7	7.7	120

The estimation was done using school-level variables without weighting. There is no consensus on whether to use sample weights when doing PSM, although the recommendation in the Stata documentation of the psmatch2 program is not to use the sampling weights when selecting a matching algorithm. All estimations presented here are unweighted, unless otherwise noted.

There are several ways to select covariates. Depending on the particular case, one could use variables identified as important in the relevant literature. Alternatively, one can run a stepwise logit to select the covariates to include. This is the method we pursue here.

The selection is done by dropping those covariates that had a p-value over 0.5 in the logit estimation. This cutoff point means that the t-statistic is under 1, which usually suggests the variable does not add additional information. Performing this exercise, the variables from Table A3.1 with p-values over 0.5 include seven outcome variables and two enrollment variables, which we have strong reasons for wanting to include. Thus, the only variables dropped based on this condition are: classrooms in grade 1 and classroom grade observed. The logit is then re-estimated with the remaining variables. The results are shown in Table A3.2.

Table A3. 2: Logit on the probability of treatment – Cohort I and Control

Variable	Odds ratio [S.E]
Total Enrollment	0.0044** [0.0016]
Enrollment Grade 1	-0.0076 [0.0115]
Enrollment Grade 2	0.0118 [0.0378]
Enrollment Grade 3	-0.0090 [0.0331]
Teachers Grade 2	-0.2212 [0.1479]
Classrooms in Grade 2 (1+)	-5.0135* [2.0423]
Classrooms in Grade 3 (1+)	2.8930 [2.0575]
Nepali speakers % range	0.4986* [0.2224]
Number of girls present in classroom	-0.1285 [0.0689]
Grade 2 is mono-grade classroom	0.4757 [0.5043]
Teacher Assistant literacy instruction	0.6252 [0.6196]
Guidance to parents to help children become readers	-0.9176 [0.6427]
Ask parents to help with homework	-1.7316 [1.0757]
Active parent-teacher association	-0.2808 [0.5892]

Variable	Odds ratio [S.E]
School has improvement plan	0.8179 [0.6837]
Annual program and budget	-0.6503 [0.5564]
School has library facility	-1.5051* [0.6449]
School provides report cards to parents	1.3731* [0.6024]
School has annual report and social audit	-0.2349 [0.3432]
Number of working computers in school	0.0738 [0.0695]
School have electricity	0.6334 [0.6138]
Source of water: Tap	-3.4406*** [0.8689]
Average Matra score grade 1	0.1693 [0.1656]
Average matra score grade 2	-0.1317 [0.1118]
Average matra score grade 3	-0.0461 [0.0929]
Average oral reading score grade 1	0.3842 [0.4045]
Average oral reading score grade 2	0.1323 [0.1621]
Average oral reading score grade 3	0.1845 [0.1384]
Average number of assets at home	-0.4810 [0.2711]
Constant	5.2281** [1.9524]
N	200

The next step is to perform a test suggested by Imbens (2010). The idea is to perform a log likelihood ratio test to different covariates in comparison to the full specification in order to determine the explanatory capacity of each particular covariate over the model. Imbens (2010) suggests that for linear models, the log likelihood ratio should be under a parameter of 1. The results for this test are presented in Table A3.3. The Log-Likelihood ratio goes below one when including the average reading comprehension scores for grade 3. This suggests that the matching can be done using only linear terms.

Table A3. 3: Log-likelihood ratio test – Cohort 1 and Control

Variable	Log Likelihood Ratio	Prob>chi 2	DF
Total Enrollment	122.77	2.43E-10	40
Enrollment Grade 1	122.36	1.55E-10	39
Enrollment Grade 2	121.82	1.03E-10	38
Enrollment Grade 3	121.80	5.62E-11	37
Teachers Grade 2	118.04	1.17E-10	36
Classrooms in Grade 2 (1+)	97.35	8.74E-08	35
Classrooms in Grade 3 (1+)	97.29	5.16E-08	34
Nepali speakers % range	96.42	3.97E-08	33

Variable	Log Likelihood Ratio	Prob>chi 2	DF
Number of girls present in classroom	94.85	3.85E-08	32
Grade 2 is mono-grade classroom	94.85	2.16E-08	31
Teacher Assistant literacy instruction	92.51	2.72E-08	30
Guidance to parents to help children become readers	89.93	3.73E-08	29
Ask parents to help with homework	83.90	1.73E-07	28
Active parent-teacher association	83.72	1.02E-07	27
School has improvement plan	83.05	7.05E-08	26
Annual program and budget	82.40	4.77E-08	25
School has library facility	78.40	1.09E-07	24
School provides report cards to parents	69.42	1.49E-06	23
School has annual report and social audit	65.97	2.81E-06	22
Number of working computers in school	65.92	1.56E-06	21
School have electricity	60.92	5.12E-06	20
Source of water: Tap	29.92	0.052839	19
Average number of assets at home	28.85	0.050227	18
Average Matra score grade 1	28.43	0.040172	17
Average Matra score grade 2	22.37	0.131505	16
Average Matra score grade 3	21.65	0.117389	15
Average Letter Sound score grade 1	15.39	0.35191	14
Average Letter Sound score grade 2	15.05	0.303953	13
Average Letter Sound score grade 3	14.76	0.254975	12
Average Invented Word score grade 1	9.15	0.608071	11
Average Invented Word score grade 2	9.00	0.531775	10
Average Invented Word score grade 3	7.64	0.570466	9
Average Oral Reading score grade 1	4.83	0.775709	8
Average Oral Reading score grade 2	3.17	0.868624	7
Average Oral Reading score grade 3	2.83	0.829738	6
Average Reading Comprehension score grade 1	2.45	0.784056	5
Average Reading Comprehension score grade 2	2.05	0.727223	4
Average Reading Comprehension score grade 3	0.89	0.828037	3
Average Listening Comprehension score grade 1	0.36	0.833586	2
Average Listening Comprehension score grade 2	0.36	0.546375	1

There are a couple of additional tests suggested in the literature. The first one is the “hit or miss” test by Heckman et al. (1998) and Heckman and Smith (1999). In this test, an observation is classified as ‘1’ if the propensity score is greater than the sample proportion of treated. Our covariates were grouped into 4 categories: School Enrollment (e.g., total enrollment, enrollment per grade level, teachers in grade 2, classrooms per grade level), School Characteristics (e.g., grade 2 classroom type, teacher assistant for literacy instruction, guidance to parents to help children become readers), School Inventory (e.g., Report cards to parents, school has library facility, annual program and budget), and Average Scores (e.g., Average Matra scores for grades 1, 2, and 3). The results of this test are presented in Table A3.4. Both Average Scores and School Inventory reach over 50% while the remaining categories are under 40%. This would suggest a moderate within sample prediction rate.

Table A3. 4: Hit-Miss Rate and Pseudo R2 tests – Cohort 1 and Control

Group	Hit-Miss Rate	Pseudo R- squared
School Characteristics	0.3495	0.0381
School Inventory	0.5146	0.1657
School Enrollment	0.3981	0.1205

Group	Hit-Miss Rate	Pseudo R-squared
Average Scores	0.5340	0.0868
School Characteristics + School Inventory	0.4660	0.2240
School Characteristics + School Enrollment	0.4126	0.1852
School Characteristics + Average Scores	0.4903	0.1354
School Inventory + School Enrollment	0.4563	0.2711
School Inventory + Average Scores	0.4806	0.2503
School Enrollment + Average Scores	0.4175	0.2221
School Characteristics + School Inventory + School Enrollment	0.4515	0.3442
School Characteristics + School Inventory + Average Scores	0.4951	0.3162
School Characteristics + School Enrollment + Average Scores	0.4466	0.3153
School Inventory + School Enrollment + Average Scores	0.4757	0.3615
All variables	0.4320	0.4507

An additional test is to look into the pseudo-R² when an additional set of covariates is added. The results are also presented in Table A3.4. Using all covariates provides the largest pseudo R-squared. Therefore, the recommendation is to use the specification presented in Table A3.2.

The next step is to compare how these characteristics balance between treatment and control, and try different matching algorithms. Table A3.5 presents the balance between treatment and control characteristics of the unmatched sample.

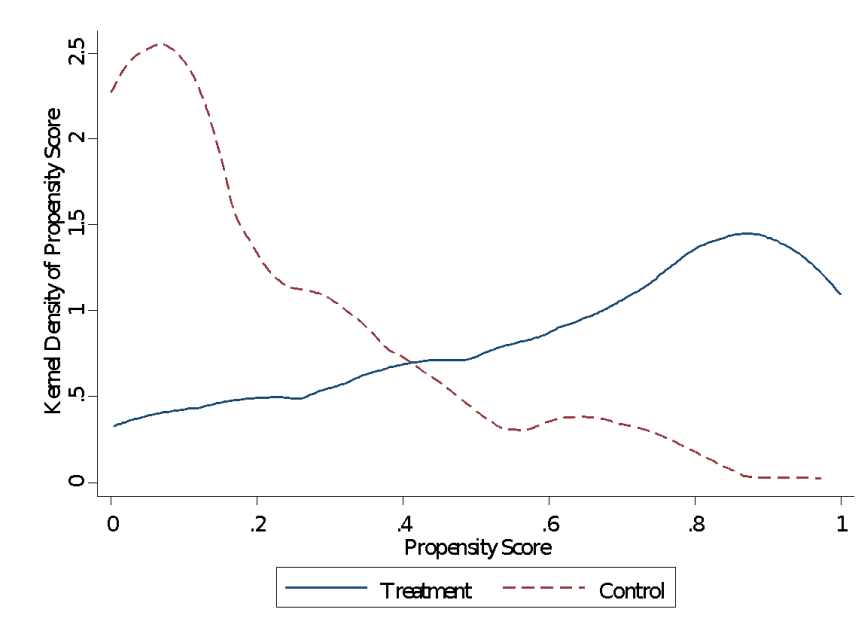
Table A3. 5: Balance between treatment and control characteristics of the unmatched sample

Variable	Unmatched			
	Cohort I	Control	T	P-value
Total Enrollment	404.04	183.15	1.36	0.18
Enrollment Grade 1	26.71	24.16	0.63	0.53
Enrollment Grade 2	23.84	19.67	1.08	0.28
Enrollment Grade 3	26.22	18.85	1.50	0.13
Teachers Grade 2	4.50	3.88	0.86	0.39
Classrooms in Grade 2 (1+)	0.67	0.83	** -2.06	0.04
Classrooms in Grade 3 (1+)	0.68	0.85	** -2.37	0.02
Nepali speakers % range	2.52	2.04	** 2.24	0.03
Number of girls present in classroom	6.05	7.87	-0.87	0.39
Grade 2 is mono-grade classroom	0.46	0.41	0.50	0.62
Teacher Assistant literacy instruction	0.22	0.23	-0.13	0.90
Guidance to parents to help children become readers	0.72	0.83	-1.29	0.20
Ask parents to help with homework	0.88	0.99	*** -2.55	0.01
Active parent-teacher association	0.62	0.71	-0.92	0.36
School has improvement plan	0.74	0.78	-0.48	0.63
Annual program and budget	0.50	0.62	-1.18	0.24

Variable	Unmatched			
	Cohort I	Contr ol	T	P-value
School has library facility	0.40	0.45	-0.49	0.62
School provides report cards to parents	0.62	0.35	***	2.83
School has annual report and social audit	1.22	1.18		0.79
Number of working computers in school	5.01	1.13	*	1.85
School have electricity	0.54	0.42		1.27
Source of water: Tap	0.37	0.52		-1.59
Average number of assets at home	5.01	4.88		0.50
Average Matra score grade 1	3.72	3.37		0.36
Average Matra score grade 2	10.02	10.76		-0.34
Average Matra score grade 3	17.37	21.20		-1.26
Average Letter Sound score grade 1	12.21	10.03		1.17
Average Letter Sound score grade 2	20.20	20.58		-0.13
Average Letter Sound score grade 3	27.68	29.89		-0.69
Average Invented Word score grade 1	0.64	0.76		-0.53
Average Invented Word score grade 2	2.91	3.42		-0.65
Average Invented Word score grade 3	5.55	6.98		-1.19
Average Oral Reading score grade 1	1.73	1.66		0.14
Average Oral Reading score grade 2	7.65	7.72		-0.04
Average Oral Reading score grade 3	14.70	16.13		-0.49
Average Reading Comprehension score grade 1	0.17	0.15		0.34
Average Reading Comprehension score grade 2	0.76	0.76		0.01
Average Reading Comprehension score grade 3	1.44	1.51		-0.26
Average Listening Comprehension score grade 1	0.35	0.26		1.01
Average Listening Comprehension score grade 2	0.56	0.53		0.40

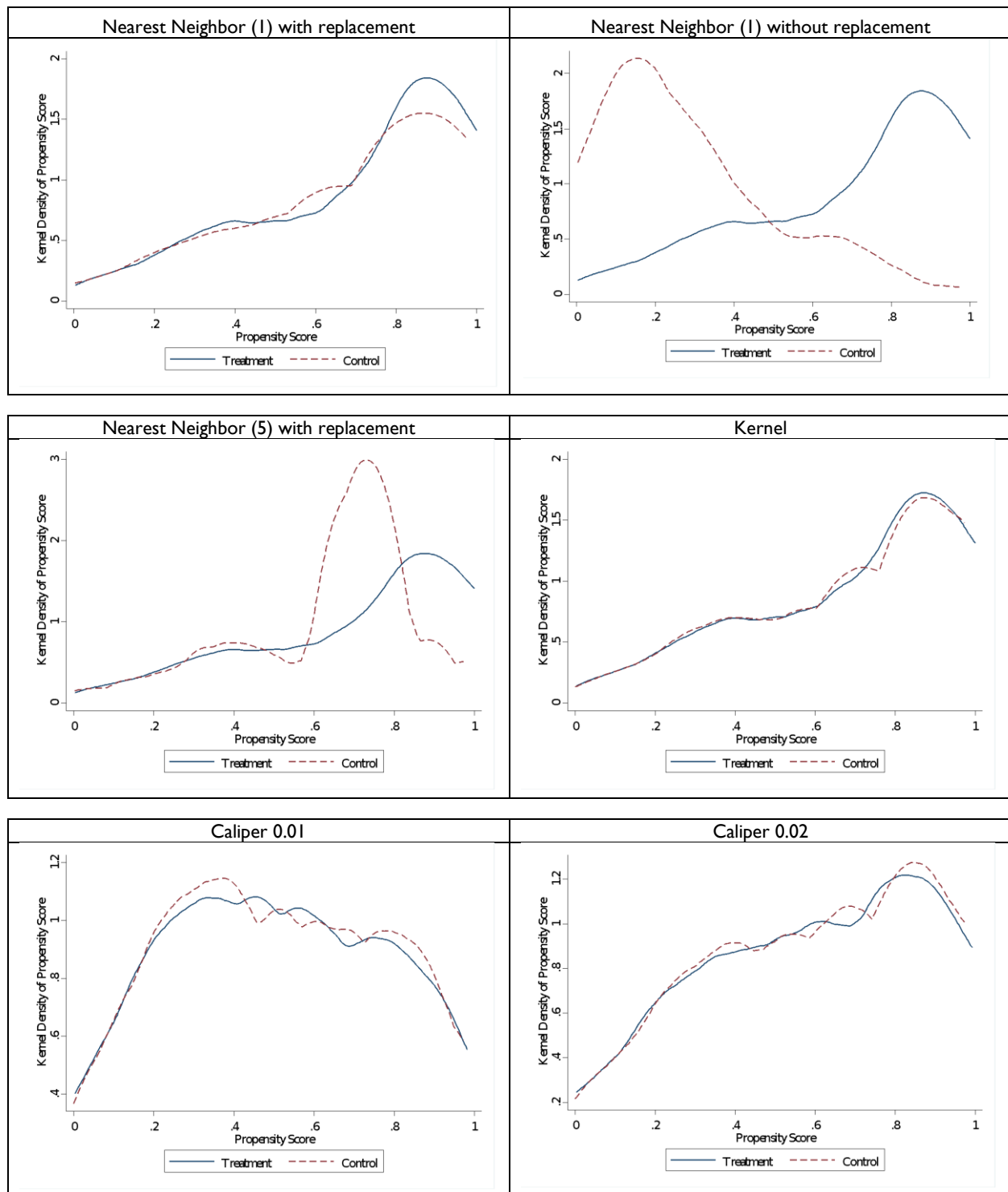
Several covariates seem to differ significantly between treatment and control. The statistical significance is given by a t-test of the difference of the means. The stars indicate that the difference between treatment and control is statistically significant (* at 10%, **, at 5 %, and *** at 1%). Graphically, the imbalance is shown by the very different distributions in propensity scores for treatment and control groups in the unmatched sample.

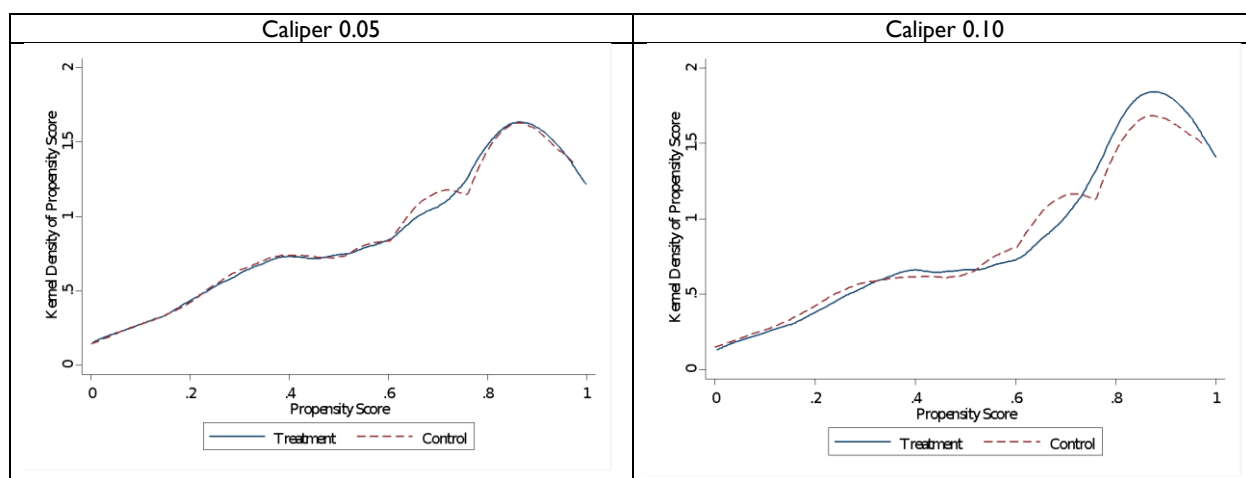
Figure A3. 1: Kernel Density of unmatched propensity score by treatment status – Cohort I and Control



To perform the matching, we used the `psmatch2` module available in Stata. It allows for different forms of matching algorithms. The propensity score is estimated out of a logit as suggested by Caliendo (2005). We used 8 types of algorithms: 1) Nearest Neighbor (1) with replacement, 2) Nearest Neighbor (1) without replacement, 3) Nearest Neighbor (5) with replacement, 4) Kernel, 5) Radius with a caliper of 0.01, 6) Radius with a caliper of 0.02, 7) Radius with a caliper of 0.05, and 8) Radius with caliper of 0.1. The graphic representation for the balance of the covariates between treatment and control by the different type of matching algorithms are presented in the different panels of Figure A3.2. Based on the results, the suggested matching algorithm would likely be either using a radius matching with caliper 0.02 or 0.05.

Figure A3. 2: Kernel Density of propensity score by treatment status and matching algorithm – Cohort I and Control





We settled on using a radius matching algorithm with a caliper of 0.05. There were two principal reasons behind this choice. First, as the corresponding graph in Figure A3.2 shows, the treatment and comparison schools are well-matched. This is also demonstrated in Table 3.6, which presents the balance between treatment and comparison schools after the matching has been implemented. Compared to the balance with the unmatched sample in Table 3.5, the treatment and comparison schools appear more similar. While some statistically significant differences remain, some of this is to be expected from random variation given the large number of variables tested.

The second reason behind the choice of using the radius matching algorithm with caliper 0.05 was due to the number of successfully matched schools. While the balance is somewhat better using the caliper of 0.01, for example, a large number of treatment schools are dropped from the sample because they fall outside of the area of common support. In fact, 52 of the 82 treatment schools fall outside of the area of common support when this algorithm is used. With the caliper of 0.05, 8 of the treatment schools fall outside the area of common support.

Table A3. 6: Balance between treatment and comparison school characteristics at baseline – Cohort 1 and Matched Comparison Group

Variable				
	Treated	Control	T-test	P-value
Total Enrollment	368.75	362.31	0.07	0.95
Enrollment Grade 1	31.58	25.61	1.34	0.18
Enrollment Grade 2	26.21	22.93	1.01	0.31
Enrollment Grade 3	27.58	26.90	0.23	0.82
Teachers Grade 2	4.34	2.78	**	2.18
Classrooms in Grade 2 (1+)	0.55	0.78	**	-1.98
Classrooms in Grade 3 (1+)	0.56	0.80	**	-2.23
Nepali speakers % range	3.03	2.94		0.16
Number of girls present in classroom	7.56	7.02		0.63
Grade 2 is mono-grade classroom	0.53	0.59		-0.30
Teacher Assistant literacy instruction	0.25	0.11	*	1.86
Guidance to parents to help children become readers	0.73	0.85		-1.42
Ask parents to help with homework	0.89	0.98	**	-2.46
Active parent-teacher association	0.70	0.70		0.02
School has improvement plan	0.79	0.79		0.04
Annual program and budget	0.56	0.60		-0.22

Variable					
	Treated	Contr ol	T-test	P-value	
School has library facility	0.40	0.54	-0.80	0.42	
School provides report cards to parents	0.60	0.73	-0.99	0.32	
School has annual report and social audit	1.01	1.13	-0.95	0.34	
Number of working computers in school	4.78	4.31	0.20	0.84	
School have electricity	0.58	0.61	-0.18	0.85	
Source of water: Tap	0.30	0.21	0.88	0.38	
Average number of assets at home	4.95	5.40	-1.65	0.10	
Average Matra score grade 1	3.30	1.91	1.50	0.14	
Average Matra score grade 2	9.68	8.80	0.45	0.66	
Average Matra score grade 3	17.85	16.89	0.33	0.74	
Average Letter Sound score grade 1	11.17	8.69	1.02	0.31	
Average Letter Sound score grade 2	20.56	20.13	0.18	0.86	
Average Letter Sound score grade 3	28.47	28.48	0.00	1.00	
Average Invented Word score grade 1	0.72	0.28	**	2.33	0.02
Average Invented Word score grade 2	2.93	2.38	0.61	0.55	
Average Invented Word score grade 3	5.83	4.74	1.26	0.21	
Average Oral Reading score grade 1	1.84	0.95	*	1.68	0.10
Average Oral Reading score grade 2	7.49	6.78	0.48	0.63	
Average Oral Reading score grade 3	14.20	14.63	-0.13	0.89	
Average Reading Comprehension score grade 1	0.18	0.09	1.53	0.13	
Average Reading Comprehension score grade 2	0.73	0.71	0.18	0.86	
Average Reading Comprehension score grade 3	1.34	1.50	-0.39	0.69	
Average Listening Comprehension score grade 1	0.30	0.20	1.18	0.24	
Average Listening Comprehension score grade 2	0.58	0.56	0.12	0.90	
Average Listening Comprehension score grade 3	0.83	0.82	0.12	0.91	
Student was absent at least one day last week	0.32	0.28	0.82	0.41	
Total number of days student was absent last week	0.85	0.65	1.61	0.11	
Mother can read	0.49	0.47	0.48	0.64	
Father can read	0.72	0.76	-0.87	0.39	

Table A3. 7: Balance between Treatment and Comparison at Baseline. Individual Characteristics. Cohorts 1 and 2 and matched comparison groups

Variable	Cohort	Treatment	Comparison	T- Test	P- value	Effect Size
Correct Sound of Letters Per Minute	1	19.54	19.17	0.18	0.85	0.02
	2	21.54	22.04	-0.30	0.77	-0.03
Correct Matra Per Minute	1	10.21	9.54	0.50	0.62	0.04
	2	12.43	12.00	0.31	0.75	0.02
Correct Invented Words Per Minute	1	3.11	2.64	1.06	0.29	0.08
	2	3.76	3.30	0.88	0.38	0.07
Oral Reading Fluency	1	7.60	7.35	0.20	0.84	0.02
	2	9.11	9.30	-0.18	0.86	-0.01
Untimed Oral Reading Fluency (per minute)	1	7.18	7.02	0.13	0.90	0.01
	2	8.69	8.88	-0.18	0.86	-0.01
Matra % of questions correct.	1	10.20	9.53	0.50	0.62	0.04
	2	12.46	11.99	0.34	0.73	0.03
Letter sounds % of questions correct.	1	19.52	19.17	0.17	0.86	0.02

Variable	Cohort	Treatment	Comparison	T-Test	P-value	Effect Size
	2	21.57	22.04	-0.28	0.78	-0.02
Invented Words % of questions correct.	1	6.22	5.27	1.05	0.29	0.08
	2	7.54	6.60	0.91	0.37	0.07
Oral Reading % of questions correct.	1	12.17	11.88	0.14	0.89	0.01
	2	14.77	15.01	-0.14	0.89	-0.01
Read Comp % of questions correct.	1	11.62	11.95	-0.14	0.89	-0.01
	2	14.64	15.40	-0.39	0.70	-0.03
Untimed Oral Reading % of questions correct.	1	22.18	22.29	-0.03	0.97	0.00
	2	26.63	27.42	-0.29	0.77	-0.02
Untimed Read Comp % of questions correct.	1	17.38	17.81	-0.14	0.89	-0.01
	2	21.20	22.10	-0.34	0.73	-0.03
Listening Comp % of questions correct.	1	18.31	17.44	0.34	0.73	0.03
	2	18.15	18.65	-0.26	0.80	-0.02
Matra Student scored zero	1	0.53	0.55	-0.41	0.68	-0.04
	2	0.47	0.46	0.19	0.85	0.01
Letter sound Student scored zero	1	0.16	0.12	1.46	0.15	0.12
	2	0.10	0.09	0.39	0.70	0.02
Invented Words Student scored zero	1	0.73	0.76	-0.88	0.38	-0.06
	2	0.68	0.70	-0.62	0.54	-0.04
Oral Reading Student scored zero	1	0.62	0.62	-0.02	0.98	0.00
	2	0.57	0.56	0.49	0.63	0.03
Reading Comp Student scored zero	1	0.73	0.71	0.38	0.71	0.04
	2	0.67	0.65	0.54	0.59	0.04
Untimed Oral Read Student scored zero	1	0.62	0.62	0.05	0.96	0.00
	2	0.57	0.55	0.49	0.62	0.03
Untimed Read Comp Student scored zero	1	0.71	0.69	0.44	0.66	0.04
	2	0.65	0.62	0.81	0.42	0.06
Listening Comp Student scored zero on section.	1	0.63	0.63	0.02	0.99	0.00
	2	0.62	0.59	0.56	0.58	0.05
Is the student female?	1	0.58	0.57	0.51	0.61	0.02
	2	0.54	0.55	-0.75	0.45	-0.02
grade==First	1	0.32	0.33	-0.43	0.67	-0.02
	2	0.34	0.32	1.06	0.29	0.04
grade==Second	1	0.34	0.32	1.01	0.31	0.04
	2	0.32	0.34	-1.52	0.13	-0.04
grade==Third	1	0.34	0.35	-0.42	0.67	-0.01
	2	0.34	0.34	-0.15	0.88	0.00
Nepali (L1 Learner)	1	0.43	0.43	-0.03	0.98	-0.01
	2	0.47	0.38	1.32	0.19	0.17

Table A3. 8: Balance between Treatment and Comparison at Baseline. Individual Characteristics. LI Learners, Cohorts 1 and 2 and matched comparison groups

Variable	Grade	Cohort	Treatment	Comparison	T-Test	P-value	Effect Size
Correct Sound of Letters Per Minute	1	1	16.90	11.50	1.66	0.10	0.24
	2		25.22	22.69	0.52	0.60	0.13
	3		32.96	33.46	-0.13	0.89	-0.02
	1	2	15.64	13.81	0.64	0.53	0.13
	2		25.27	27.25	-0.74	0.46	-0.11
	3		38.25	37.71	0.17	0.86	0.03
Correct Matra Per Minute	1	1	5.84	3.71	1.20	0.23	0.20
	2		13.39	12.70	0.19	0.85	0.04
	3		21.88	21.38	0.18	0.85	0.02
	1	2	5.48	4.69	0.60	0.55	0.08
	2		15.05	15.93	-0.36	0.72	-0.05
	3		27.87	25.87	0.75	0.45	0.09
Correct Invented Words Per Minute	1	1	1.18	0.89	0.64	0.52	0.08
	2		4.08	3.53	0.40	0.69	0.08
	3		6.81	6.08	0.75	0.45	0.09
	1	2	1.22	1.09	0.36	0.72	0.04
	2		4.28	4.44	-0.17	0.87	-0.02
	3		9.32	7.84	1.14	0.25	0.17
Oral Reading Fluency	1	1	3.44	2.45	0.83	0.41	0.13
	2		10.66	10.23	0.11	0.91	0.03
	3		18.22	18.40	-0.06	0.95	-0.01
	1	2	2.43	2.33	0.14	0.89	0.02
	2		10.74	12.53	-0.93	0.35	-0.12
	3		23.55	23.99	-0.17	0.86	-0.02
Untimed Oral Reading Fluency	1	1	3.02	2.11	0.86	0.39	0.13
	2		10.05	9.89	0.05	0.96	0.01
	3		17.39	17.80	-0.13	0.90	-0.02
	1	2	2.21	1.97	0.38	0.71	0.04
	2		9.86	11.95	-1.13	0.26	-0.15
	3		22.87	23.14	-0.11	0.92	-0.01
Oral Reading Comprehension Percentage of questions correct.	1	1	5.91	3.94	0.88	0.38	0.15
	2		17.61	17.72	-0.02	0.99	0.00
	3		28.47	32.45	-0.67	0.50	-0.13
	1	2	3.77	4.11	-0.27	0.79	-0.03
	2		18.19	21.69	-1.09	0.28	-0.14
	3		38.35	43.39	-1.14	0.25	-0.17
Untimed Oral Reading Comprehension Percentage of questions correct.	1	1	9.82	6.25	1.01	0.31	0.17
	2		26.31	26.05	0.03	0.98	0.01
	3		42.27	46.32	-0.61	0.54	-0.10

Variable	Grade	Cohort	Treatment	Comparison	T-Test	P-value	Effect Size
	1	2	5.81	6.44	-0.32	0.75	-0.04
	2		26.84	32.90	-1.29	0.20	-0.17
	3		54.78	59.95	-1.03	0.31	-0.13
Listening Comprehension Percentage of questions correct.	1	1	15.80	15.10	0.17	0.86	0.03
	2		26.01	27.06	-0.27	0.79	-0.04
	3		36.40	35.36	0.37	0.71	0.03
	1	2	13.96	16.02	-0.45	0.65	-0.09
	2		21.98	29.66	-2.34	0.02	-0.25
	3		37.25	36.62	0.21	0.84	0.02

Table A3. 9: Balance between Treatment and Comparison at Baseline. Individual Characteristics. L2 Learners, Cohorts 1 and 2 and matched comparison groups

Variable	Grade	Cohort	Treatment	Comparison	T-Test	P-value	Effect Size
Correct Sound of Letters Per Minute	1	1	6.52	8.06	-1.21	0.23	-0.15
	2		16.37	17.32	-0.45	0.65	-0.06
	3		22.56	23.21	-0.24	0.81	-0.03
	1	2	8.09	8.36	-0.19	0.85	-0.03
	2		17.52	19.89	-1.04	0.30	-0.14
	3		25.95	28.50	-0.74	0.46	-0.12
Correct Matra Per Minute	1	1	1.50	1.66	-0.24	0.81	-0.03
	2		7.59	6.26	0.77	0.45	0.11
	3		12.76	12.78	-0.01	0.99	0.00
	1	2	1.81	2.05	-0.34	0.74	-0.04
	2		7.87	8.07	-0.11	0.91	-0.02
	3		17.40	18.38	-0.29	0.77	-0.05
Correct Invented Words Per Minute	1	1	0.42	0.41	0.03	0.98	0.00
	2		2.16	1.59	1.00	0.32	0.12
	3		4.39	3.67	0.93	0.36	0.10
	1	2	0.42	0.45	-0.13	0.90	-0.01
	2		2.03	1.79	0.40	0.69	0.05
	3		5.53	5.24	0.26	0.80	0.04
Oral Reading Fluency	1	1	0.62	0.68	-0.19	0.85	-0.02
	2		4.54	4.01	0.51	0.61	0.06
	3		9.58	9.52	0.03	0.98	0.00
	1	2	0.88	0.86	0.03	0.98	0.00
	2		4.95	5.04	-0.08	0.94	-0.01
	3		12.90	14.11	-0.43	0.67	-0.07
Untimed Oral Reading Fluency	1	1	0.59	0.61	-0.07	0.94	-0.01
	2		4.23	3.72	0.53	0.60	0.06

Variable	Grade	Cohort	Treatment	Comparison	T-Test	P-value	Effect Size
	3		9.16	9.18	-0.01	0.99	0.00
	1	2	0.75	0.78	-0.08	0.94	-0.01
	2		4.59	4.60	0.00	1.00	0.00
	3		12.52	13.72	-0.44	0.66	-0.07
Oral Reading Comprehension Percentage of questions correct.	1	1	0.74	0.99	-0.76	0.45	-0.06
	2		6.45	6.30	0.10	0.92	0.01
	3		13.42	12.64	0.26	0.80	0.03
	1	2	1.33	1.15	0.30	0.77	0.03
	2		7.73	8.36	-0.32	0.75	-0.04
	3		19.96	20.03	-0.02	0.99	0.00
Untimed Oral Reading Comprehension Percentage of questions correct.	1	1	1.23	1.67	-0.67	0.51	-0.06
	2		9.57	9.57	0.00	1.00	0.00
	3		19.64	20.44	-0.17	0.86	-0.02
	1	2	1.67	2.12	-0.41	0.68	-0.05
	2		11.57	12.07	-0.17	0.86	-0.02
	3		28.52	28.67	-0.03	0.98	0.00
Listening Comprehension Percentage of questions correct.	1	1	5.32	4.11	0.71	0.48	0.08
	2		13.49	12.14	0.36	0.72	0.06
	3		17.36	15.84	0.44	0.66	0.06
	1	2	5.42	5.09	0.21	0.84	0.02
	2		12.28	13.67	-0.43	0.67	-0.06
	3		19.95	19.74	0.07	0.95	0.01

ANNEX IV: ADDITIONAL ANALYSES

Table A4. I Table A4.1: Materials available to Students in the Classrooms. Cohorts 1 and 2

Cohort 1	Baseline			Midline			DiD (7=6-3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
All or most students have Nepali language textbook	0.78	0.62	-0.16	0.55	0.82	0.27	0.43**	0.93
All or most students have Nepali language workbook	0.71	0.51	-0.2	0.58	0.91	0.33	0.53***	1.22
All or most students have mother language textbook	0.00	0.06	0.06	0.11	0.12	0.01	-0.05	-0.16
All or most students have mother language workbook	0.02	0.04	0.02	0.12	0.13	0.01	-0.01	-0.03
Reading materials (not textbook) easily accessible inside classroom	0.19	0.21	0.02	0.44	0.91	0.47	0.45***	0.99
Curriculum of related subject	0.29	0.21	-0.08	0.35	0.54	0.19	0.27**	0.56
Teacher's Guidelines for Nepali Language	0.19	0.13	-0.06	0.23	0.85	0.62	0.68***	1.38
Teacher's Guidelines for Local Language	0.00	0.01	0.01	0.03	0.14	0.11	0.10**	0.36
Supplementary reading material avail.	0.19	0.09	-0.1	0.13	0.89	0.76	0.86***	1.72
Blackboard/Whiteboard	0.96	0.99	0.03	0.98	0.99	0.01	-0.02	-0.18
Chalk/Marker	0.88	0.82	-0.06	0.96	0.93	-0.03	0.03	0.14
Pen/Pencil	0.55	0.58	0.03	0.77	0.92	0.15	0.12	0.33
Notebook	0.06	0.17	0.11	0.26	0.26	0.00	-0.11	-0.25

Cohort 2	Baseline			Midline			DiD (7=6-3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
All or most students have Nepali language textbook	0.74	0.88	0.14	0.43	0.69	0.26	0.12	0.24
All or most students have Nepali language workbook	0.65	0.66	0.01	0.47	0.76	0.29	0.28	0.57
All or most students have mother language textbook	0	0.02	0.02	0.08	0.04	-0.04	-0.06	-0.25
All or most students have mother language workbook	0.01	0.04	0.03	0.13	0.03	-0.1	-0.13**	-0.47
Reading materials (not textbook) easily accessible inside classroom	0.28	0.13	-0.15	0.57	0.79	0.22	0.37***	0.81
Curriculum of related subject	0.25	0.34	0.09	0.38	0.25	-0.13	-0.22*	-0.47

Cohort 2	Baseline			Midline			DiD (7=6-3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
Teacher's Guidelines for Nepali Language	0.19	0.27	0.08	0.22	0.31	0.09	0.01	0.05
Teacher's Guidelines for Local Language	0	0.01	0.01	0.04	0.04	0	-0.01	-0.05
Supplementary reading material avail.	0.29	0.07	-0.22	0.27	0.69	0.42	0.64***	1.26
Blackboard/Whiteboard	0.97	0.95	-0.02	0.98	0.99	0.01	0.03	0.24
Chalk/Marker	0.9	0.94	0.04	0.96	0.92	-0.04	-0.08*	-0.38
Pen/Pencil	0.43	0.53	0.1	0.77	0.71	-0.06	-0.16	-0.36
Notebook	0.05	0.12	0.07	0.29	0.08	-0.21	-0.28**	-0.71

Table A4. 2: Teacher Practices Indexes. Cohorts 1 and 2

Cohort 1	Baseline			Midline			DiD (7=6-3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
Student-Centered Teaching Practices Index	15.26	14.71	-0.55	15.12	20.02	4.9	5.45***	0.91
Student-Centered Teaching Quality Index	6.45	7.5	1.05	8.14	9.61	1.47	0.42	0.21
Student-Centered Classroom Index	3.64	4.19	0.55	4.35	5.36	1.01	0.46	0.33

Cohort 2	Baseline			Midline			DiD (7=6-3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
Student-Centered Teaching Practices Index	16.52	15.59	-0.93	15.62	14.58	-1.04	-0.11	-0.02
Student-Centered Teaching Quality Index	6.82	7.29	0.47	8.56	7.8	-0.76	-1.23**	-0.73
Student-Centered Classroom Index	4.07	3.86	-0.21	4.65	4.91	0.26	0.47	0.35

Table A4. 3: Teacher Support, Reading Assignments and Attitudes about Reading. Cohorts 1 and 2

Cohort I	Baseline			Midline			DiD (7=6-3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
Additional Support to Learners and Communication with Parents								
Individualized remedial support outside class	0.23	0.28	0.05	0.16	0.16	0	-0.05	-0.14
Individualized remedial support inside class	0.51	0.49	-0.02	0.42	0.5	0.08	0.10	0.22
Additional practice time inside class	0.49	0.54	0.05	0.44	0.27	-0.17	-0.22	-0.48
Peer pairing or small group work	0.25	0.38	0.13	0.12	0.39	0.27	0.14	0.32
Whole class revision	0.17	0.23	0.06	0.04	0.07	0.03	-0.03	-0.18
Additional reading materials or project work outside class	0.03	0.12	0.09	0.06	0.13	0.07	-0.02	-0.03
Additional Support: Parent-teacher conference or communication	0.29	0.14	-0.15	0.04	0.23	0.19	0.34***	0.99
Conducts at least 1 formal meeting w/ parents per term	0.44	0.28	-0.16	0.55	0.58	0.03	0.19	0.36
Sends at least 1 student progress report to parents per term	0.39	0.35	-0.04	0.49	0.41	-0.08	-0.04	-0.06
Reading Assignments								
Gives daily reading assignment to complete outside school	0.79	0.68	-0.11	0.58	0.7	0.12	0.23	0.48
Attitudes								
All learners can learn to read.	0.53	0.67	0.14	0.74	0.63	-0.11	-0.25	-0.54
All learners can learn to write.	0.59	0.76	0.17	0.72	0.66	-0.06	-0.23	-0.5
Children acquire reading skills by exposure, without being taught to read.	0.41	0.4	-0.01	0.17	0.4	0.23	0.24	0.55
Give learners time each day to read freely materials of their own choice.	0.98	0.94	-0.04	0.89	0.96	0.07	0.11	0.42
Learners must be able to recite a text before they can read it.	0.4	0.42	0.02	0.1	0.15	0.05	0.03	0.09
Better to teach R&V separately	0.68	0.87	0.19	0.83	0.57	-0.26	-0.45***	-0.98
Learners cannot write an original passage until at least grade 3 or 4.	0.67	0.69	0.02	0.6	0.49	-0.11	-0.13	-0.28
Important to give learner time each day to write on topics of own choice.	0.99	0.92	-0.07	1	0.99	-0.01	0.06*	0.86
It is important to correct ALL the errors in sentences learners produce.	0.96	0.9	-0.06	0.99	0.94	-0.05	0.01	0.05
Reading stories to learners helps them develop their reading skills	0.66	0.78	0.12	0.67	0.74	0.07	-0.05	-0.11
Young learners must memorize a text before they can understand it.	0.29	0.23	-0.06	0.25	0.27	0.02	0.08	0.18

Cohort 1	Baseline			Midline			DiD (7=6-3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
Silent reading should be avoided, as it can't check if learner reading	0.64	0.72	0.08	0.78	0.68	-0.1	-0.18	-0.41
A learner writes "well" is not make any grammatical or spelling mistake.	0.67	0.66	-0.01	0.57	0.63	0.06	0.07	0.14
Some students learn to read more slowly as not understand language well.	0.94	0.86	-0.08	0.74	0.72	-0.02	0.06	0.13
If a student can read quickly, that means he/she is a good reader.	0.58	0.71	0.13	0.7	0.85	0.15	0.02	0.05

Cohort 2	Baseline			Midline			DiD (7=6-3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
Additional Support to Learners and Communication with Parents								
Individualized remedial support outside class	0.29	0.33	0.04	0.13	0.07	-0.06	-0.10	-0.36
Individualized remedial support inside class	0.51	0.64	0.13	0.41	0.46	0.05	-0.08	-0.14
Additional practice time inside class	0.47	0.53	0.06	0.43	0.36	-0.07	-0.13	-0.24
Peer pairing or small group work	0.25	0.36	0.11	0.09	0.23	0.14	0.03	0.06
Whole class revision	0.09	0.19	0.1	0.08	0.08	0	-0.10	-0.41
Additional reading materials or project work outside class	0.03	0.06	0.03	0.05	0.08	0.03	0.00	0
Additional Support: Parent-teacher conference or communication	0.18	0.22	0.04	0.09	0.04	-0.05	-0.09	-0.36
Conducts at least 1 formal meeting w/ parents per term	0.48	0.3	-0.18	0.61	0.33	-0.28	-0.10	-0.2
Sends at least 1 student progress report to parents per term	0.42	0.14	-0.28	0.46	0.25	-0.21	0.07	0.12
Reading Assignments								
Gives daily reading assignment to complete outside school	0.82	0.67	-0.15	0.51	0.67	0.16	0.31*	0.61
Attitudes								

Cohort 2	Baseline			Midline			DiD (7=6-3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
All learners can learn to read.	0.55	0.69	0.14	0.75	0.66	-0.09	-0.23	-0.51
All learners can learn to write.	0.55	0.66	0.11	0.81	0.77	-0.04	-0.15	-0.34
Children acquire reading skills by exposure, without being taught to read.	0.35	0.36	0.01	0.16	0.3	0.14	0.13	0.34
Give learners time each day to read freely materials of their own choice.	0.98	0.94	-0.04	0.97	0.93	-0.04	0.00	0
Learners must be able to recite a text before they can read it.	0.34	0.29	-0.05	0.08	0.24	0.16	0.21	0.55
Better to teach R&W separately	0.73	0.78	0.05	0.81	0.76	-0.05	-0.10	-0.22
Learners cannot write an original passage until at least grade 3 or 4.	0.7	0.69	-0.01	0.67	0.62	-0.05	-0.04	-0.06
Important to give learner time each day to write on topics of own choice.	0.98	0.95	-0.03	1	0.97	-0.03	0.00	0
It is important to correct ALL the errors in sentences learners produce.	0.92	0.86	-0.06	0.95	0.89	-0.06	0.00	0
Reading stories to learners helps them develop their reading skills	0.75	0.81	0.06	0.64	0.86	0.22	0.16	0.39
Young learners must memorize a text before they can understand it.	0.28	0.34	0.06	0.18	0.25	0.07	0.01	0.05
Silent reading should be avoided, as it can't check if learner reading.	0.57	0.72	0.15	0.76	0.84	0.08	-0.07	-0.17
A learner writes "well" is not make any	0.67	0.66	-0.01	0.59	0.79	0.2	0.21	0.45

Cohort 2	Baseline			Midline			DiD (7=6-3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
grammatical or spelling mistake.								
Some students learn to read more slowly as not understand language well.	0.95	0.75	-0.2	0.84	0.85	0.01	0.21***	0.58
If a student can read quickly, that means he/she is a good reader.	0.63	0.64	0.01	0.67	0.79	0.12	0.11	0.27

Table A4. 4: School Management Committee. Cohorts 1 and 2

Cohort 1	Baseline			Midline			DiD (7=6-3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
Received school management capacity building training in last two years	0.26	0.36	0.1	0.15	0.35	0.2	0.10	0.25
Management Index	7.05	7.18	0.13	7.69	8.72	1.03	0.90*	0.49
SMC met at least once per month in last year	0.68	0.52	-0.16	0.53	0.55	0.02	0.18	0.36
Early Grade Literacy is high priority for SMC	0.71	0.47	-0.24	0.63	0.68	0.05	0.29**	0.61
PTA meets at least every two months this year	0.32	0.26	-0.06	0.46	0.3	-0.16	-0.10	-0.23
School engages PTA/community for book drives & book donations	0.39	0.48	0.09	0.18	0.49	0.31	0.22	0.47
School works with PTA to manage resources for EGR improvement programs	0.55	0.55	0	0.62	0.75	0.13	0.13	0.26
School provided guidance to parents to help children read	0.84	0.78	-0.06	0.87	0.83	-0.04	0.02	0.06

Cohort 2	Baseline			Midline			DiD (7=6-3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
Received school management capacity building training in last two years	0.43	0.37	-0.06	0.16	0.24	0.08	0.14	0.35
Management Index	7.8	7.1	-0.7	8.11	8.05	-0.06	0.64	0.31
SMC met at least once per month in last year	0.57	0.57	0	0.62	0.46	-0.16	-0.16	-0.32
Early Grade Literacy is high priority for SMC	0.68	0.45	-0.23	0.63	0.49	-0.14	0.09	0.18
PTA meets at least every two months this year	0.32	0.24	-0.08	0.45	0.24	-0.21	-0.13	-0.27
School engages PTA/community for book drives & book donations	0.31	0.41	0.1	0.2	0.36	0.16	0.06	0.13
School works with PTA to manage resources for EGR improvement programs	0.59	0.53	-0.06	0.7	0.66	-0.04	0.02	0.04
School provided guidance to parents to help children read	0.86	0.69	-0.17	0.94	0.77	-0.17	0.00	0.06
School asks parents to help with homework and read to children	0.96	0.94	-0.02	0.93	0.95	0.02	0.04	0.21
Received school management capacity building training in last two years	0.43	0.37	-0.06	0.16	0.24	0.08	0.14	0.35
Management Index	7.8	7.1	-0.7	8.11	8.05	-0.06	0.64	0.31
SMC met at least once per month in last year	0.57	0.57	0	0.62	0.46	-0.16	-0.16	-0.32

Table A4. 5: At Home Reading Activities. Cohorts 1 and 2

	Baseline			Midline			DiD (7=6-3)	Effect Size
	Comp	Treat	Diff	Comp	Treat	Diff		
Cohort 1								
Subscribe to children's magazines	0.0	0.0	0.0	0.0	0.1	0.0	0.03	0.1
You or someone in your household reads to your child at least once a week	0.7	0.6	-0.1	0.8	0.8	0.0	0.08	0.2

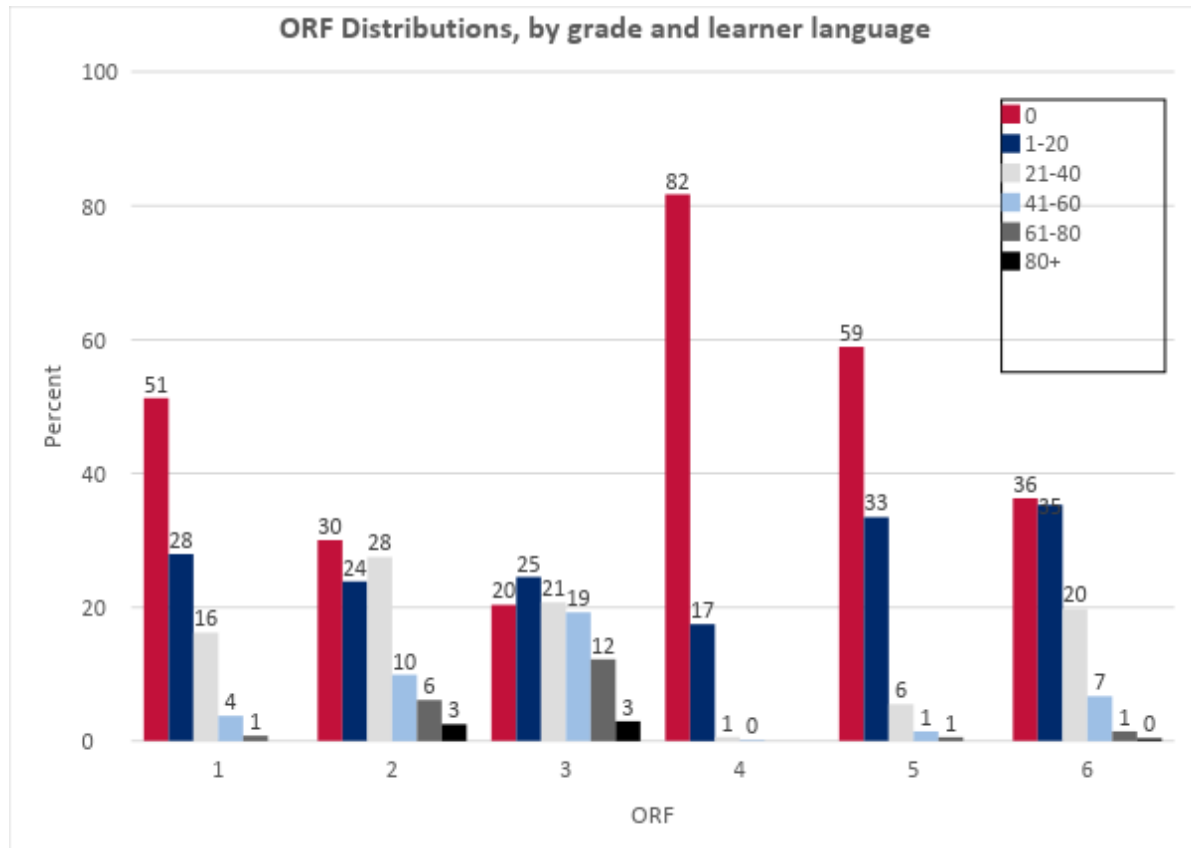
	Baseline			Midline			DiD (7=6-3)	Effect Size
	Comp	Treat	Diff	Comp	Treat	Diff		
Your child reads to you or someone in your household at least once a week	0.5	0.5	0.1	0.9	0.8	-0.1	-0.12	-0.3
Cohort 2								
Subscribe to children's magazines	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.1
You or someone in your household reads to your child at least once a week	0.6	0.6	0.0	0.7	0.7	-0.1	-0.02	0.0
Your child reads to you or someone in your household at least once a week	0.6	0.5	-0.1	0.9	0.8	-0.2	-0.09	-0.2

Table A4. 6: EGRP Effects for Boys and Girls and Difference Between the Effects (DIDID)

	Male Students			Female Students			Diff-in-Diff-in-Diff		
	Baseline Diff (1)	Midline Diff (2)	DiD (3=2-1)	Baseline Diff (4)	Midline Diff (5)	DiD (6=5-4)	DIDID (7=6-3)	Adjusted DIDID	Effect Size
Correct Sound of Letters Per Minute									
Grade 1	2.8	7.6	4.8**	0	5.5	5.5***	0.7	1.9	0.14
Grade 2	2.1	2.6	0.5	-0.5	4.3	4.8	4.3	3.8	0.21
Grade 3	1.7	7.5	5.8	-2.6	5.4	8.0***	2.2	2.3	0.11
Correct Matra Per Minute									
Grade 1	1.1	4.6	3.5**	0.4	4.1	3.7**	0.2	1.0	0.09
Grade 2	2.1	1.7	-0.4	0.6	2.6	2.0	2.4	2.3	0.13
Grade 3	1.9	4.8	2.9	-1.3	4.6	5.9*	3.0	2.8	0.13
Correct Invented Words Per Minute									
Grade 1	0.2	1.2	1.0**	0	1.4	1.4***	0.4	0.4	0.11
Grade 2	0.8	0.6	-0.2	0.6	1.9	1.3	1.5	1.5	0.23
Grade 3	1.2	3.1	1.9	0.3	2.6	2.3*	0.4	0.3	0.04
Oral Reading Fluency									
Grade 1	0.6	3.1	2.5**	0.1	3.1	3.0***	0.5	0.7	0.09
Grade 2	1.5	2.6	1.1	0	4.5	4.5	3.4	3.6	0.23
Grade 3	1.2	8.5	7.3*	-1.4	7	8.4***	1.1	1.2	0.06
read_comp Percentage of questions correct.									
Grade 1	1	5.9	4.9**	0.3	5.9	5.6***	0.7	0.9	0.06
Grade 2	1.4	2.6	1.2	-0.4	5.3	5.7	4.5	4.3	0.17
Grade 3	-0.3	8.5	8.8	-3	9.4	12.4***	3.6	3.6	0.12
list_comp Percentage of questions correct.									
Grade 1	-1.8	10.3	12.1***	2.9	12.8	9.9***	-2.2	-2.5	-0.1
Grade 2	-0.3	6.7	7.0	1.7	14.3	12.6**	5.6	3.2	0.1

	Male Students			Female Students			Diff-in-Diff-in-Diff		
	Baseline Diff (1)	Midline Diff (2)	DiD (3=2-1)	Baseline Diff (4)	Midline Diff (5)	DiD (6=5-4)	DIDID (7=6-3)	Adjusted DIDID	Effect Size
Grade 3	4.5	16.9	12.4**	-1.9	2.3	4.2	-8.2	-9.5	-0.27

Figure A4. 1: Oral Reading Fluency Distributions, by Grade and Learner Language. Cohort 1



ANNEX V: SAMPLE

Before NORC was asked to conduct the IE of EGRP, RTI had already decided on the sample approach and calculated the sample size to be used. A representative sample of schools from treatment and control districts was selected for the baseline. That sample of schools was re-visited at midline, and will be re-visited again at endline, forming a panel of schools.

As mentioned, the sample design and calculations were done by the IP, RTI, and we include their information below.

OVERVIEW

The impact evaluation is concerned with how the Early Grade Reading Program will improve learning outcomes for pupils. The population of interest are the children in cohorts 1 & 2 who are L1 and L2 learners. As a result, the sample design is concerned with creating a sample of pupils that is representative of the L1 and L2 learners within cohorts 1 & 2.

Note that impact evaluation is measured at the cohort level. Using probability proportional to size sampling (PPS) across each cohort will result in a sample which is representative of each cohort. While we will adjust the sample to ensure we have enough L1 and L2 learners, the sampling technique controls for other differences in the cohort such as District, socio-economic status, eco-belt and other factors through randomization; thus eliminating the need to sample for these other differences. The impact will not be measured at the district level; this issue will be addressed in the performance evaluation which will measure the implementation of the EGRP model, not impact.

RESEARCH QUESTIONS

- EGRP improved the reading outcomes of pupils who speak Nepali as a first language (L1 learners) in cohorts 1 & 2
- EGRP improved the reading outcomes of pupils who do not speak Nepali as a first language (L2 learners) in cohorts 1 & 2

This impact is evaluated through a difference-in-difference analysis model; looking at the improvement of the pupils in the categories described above controlling for the learning gain of pupils not at a school participating in the EGRP.

Note we are concerned with the learning outcomes of L1 and L2 learners, as it is not possible to classify schools as L1 or L2 types because most schools have a mix of learners. All published results will be disaggregated by cohort and L1/L2 learner type.

SAMPLE DESIGN

The sample determinations are made such that we statistically significantly detect a difference of 6 wpm for reading fluency with 80% confidence.

The original calculation used the following assumptions, based on previous studies:

- Grade 2 mean= 15 words per minute, with standard deviation = 28 words per minute
- Grade 3 mean= 28 words per minute, with standard deviation = 24 words per minute
- The intraclass correlation coefficient or ICC for the school clusters = 0.25
- Power of the test = 80%
- MDES is the minimum detectable effect size. The MDES is the smallest impact of the activity on the outcome variable that the evaluation will be able to detect. EGRP selected a MDES of 6 words per minute per year

Based on those parameters, the sample size was estimated as 86 treatment schools in each treatment cohort (1 and 2), with 10 students per grade, from grades 1 to 3 in each school (amounting to 30 students per school and 2,580 students in total); and 90 comparison schools, with 10 students each from grades 1-3 per school (for a total of 2,700 students in total). Students are always selected randomly among those present in the classroom or classrooms, if the grade has more than one sections.

NORC requested a larger sample size, given that a MDES of 6 wpm seems ambitious, particularly among first graders. We originally requested an increase in sample to be able to identify a MDES equal to 4wpm but it was not possible to accommodate the request.

NORC then requested an increase in the number of comparison schools to 120 in order to be able to conduct the matching and avoid problems in finding common support among treatment and control schools. EGRP agreed to this larger sample for the comparison group.

The schools listed in the sample framework included many institutions with very few students in grades 1, 2 and 3. Drawing the sample without taking this fact into account would yield a sample smaller than desired, because some schools would have less than the requisite 10 students per grade. Therefore, it was agreed to:

- survey and assess 12 random students -rather than 10- per grade per school when possible
- drop schools with 5 or less pupils in G1, G2 or G3

The final sample was then 86 treatment schools in each cohort, 12 students per grade, in grades 1 to 3 per school (a total of 3,096 students) and 120 control schools, 12 students per grade in grades 1 to 3 per school (a total of 4,320 students).

Because the measurement of student performance for impact of the EGRP will be reported for cohort 1 & 2 by L1 and L2 learner, it is important to stratify by L1 and L2 learner in each cohort. That is, sample the desired amount of L1 & L2 learners to ensure desired statistical power. **Ten** pupils of each grades 1-3 will be selected in the sampled schools, a total of 30 pupils per school. The total sample size is shown below in table 1.

Table A5. 1: Sample Size EGRP

	Number of Schools	Learner TYPE	Total Grade 1 pupils	Grade 2 pupils per school	Grade 3 pupils per school	Total pupils to be sampled
Cohort 1	86	L1	430	430	430	2580
		L2	430	430	430	
Cohort 2	86	L1	430	430	430	2580
		L2	430	430	430	

Cohort 1 has approximately 32% & 68% L1 and L2 learners, respectively, while cohort 2 has 38% & 62% learners for L1 and L2. If we sampled in these proportions, our sample sizes for L1 learners would be smaller and lack statistical power. Thus, we will oversample L1 learners to achieve approximately 50% L1 learners in the sample.

By categorizing schools as percent of learners who are Nepali speakers, we are able to adjust the number of schools required to achieve the desired proportion of L1 and L2 learners. The **Table A4.2: Sample Design EGRP** shows the approximate number of pupils within schools categorized by percentage of Nepali speakers in the schools. Column A shows the categories of schools and columns C and D show the approximate number L1 and L2 learners by these school categories. The percent of total row show that the proportion of learners in cohorts 1 & 2 is unbalanced and there are more L2 learners in both cohorts. We need to oversample the number of L1 learners in order to achieve an approximately 50-50 split of L1 and L2 learners in the sample. This is achieved by oversampling more schools with higher L1 learners and less schools with L2 learners. This adjustment is shown in column G. The final desired number of schools to be sampled is shown in column H.

The number of control schools is dependent, like cohorts 1 and 2, on the percentage of L1 and L2 learners within the entire control “cohort”. As a result, it may also be necessary to oversample to achieve the following:

- An appropriate number of L1 and L2 learners
- An oversample of schools such that school matching can occur.

Table A5. 2: Sample Design EGRP

	Percentage of Nepali Speakers in School	Sum of Total number of Pupil (Grade 1-3)	Approx percent L1 Learners	Approx percent L2 Learners	Proportion of Total Number of Pupils	Schools to be sampled proportionally	Adjustment	Number of Schools to be Sampled	Approx Number of L1 Pupils Sampled	Approx Number of L2 Pupils Sampled
Cohort 1	0-20	67515	6752	60764	51%	44	-16	28	83	744
	20-40	18884	5665	13219	14%	12	-6	6	56	130
	40-60	21562	10781	10781	16%	14	-2	12	179	179
	60-80	15483	10838	4645	12%	10	12	22	462	198
	80-100	9894	8905	989	7%	6	12	18	496	55
	TOTAL	133338	42940	90398	100%	86	0	86	1275	1305
	PERCENT OF TOTAL		32%	68%					49%	51%
Cohort 2	0-20	83680	8368	75312	35%	30	-9	21	63	569
	20-40	45985	13796	32190	19%	17	-6	11	95	221
	40-60	60141	30071	30071	25%	22	0	22	324	324
	60-80	34196	23937	10259	14%	12	8	20	426	183
	80-100	15193	13674	1519	6%	5	7	12	336	37
	TOTAL	239195	89845	149350	100%	86	0	86	1245	1335
	PERCENT OF TOTAL		38%	62%					48%	52%

SMALL SCHOOLS DETERMINATION

The school list from which the sample will be drawn reports many schools with few pupils in grades 1, 2 and 3 such that if the sample was drawn without consideration of this issue, the sample would be smaller than desired; the average number of pupils sampled per grade would be approximately 8.5, a 15% drop in the sample. As a result, a proactive decision needs to be made regarding how to keep the sample size at the desired level. The following options are available:

- Sampling 12 pupils per grade/school and kept all the schools in the sample list, we would have an average of 9.6 pupils per school/grade – this is acceptable
- If we drop schools with 5 or less pupils in G1, G2 OR G3, we'd have 9.6 pupils per school/grade average, but inference would be reduced to the schools remaining in the list
- If we drop schools with 6 or less pupils in G1, G2 OR G3 we'd have 9.8 pupils per school/grade average, but inference would be reduced to the schools remaining in the list

SAMPLING PROCEDURE

Stage 1: School Selection

- **Irrespective of district**, school lists will be grouped (i.e. stratified) by percentage of Nepali Speakers in schools (0-20, 20-40, etc.). Using probability proportional of size (PPS) sampling, the number of schools selected in each will reflect the numbers shown in table 2, column H. Additionally, replacement schools will be selected in each language category equal to 20% of the desired sample, rounded up.

Stage 2: Pupil Selection

- Pupils will be lined up by grade from tallest to shortest, **irrespective of gender and L1 or L2 language status**. Then depending on the number of pupils per grade, one pupil will be selected at intervals along the line. For example, if there are 20 pupils in grade 1, select every other pupil for a total of 10. This systematic sampling should be done for each grade.
- Because it will be necessary to link the teacher dataset to the pupil data, if a school has multiple classes for a given grade, one class per grade will be selected randomly and pupils selected will be from those selected classes only. The teacher interviewed will be teacher of the selected class, ensuring linkage between teacher and pupil datasets.

CONTROL SAMPLE DESIGN COHORT I (created by NORC to complement RTI treatment sample design)

Given that all schools in treatment districts will receive EGRP, we need to create a control sample using out of district schools. A group of control districts was selected by RTI to match the characteristics of the treatment districts in general. The dimensions that were taking into account for the selection were landscape/climate, socio-cultural settings, and economic activity. The selected control districts to match Cohort I treatment districts are: Doti, Myagdi, Kapilvastu, Bara, Sunsari, and Kavre.

We will follow a sample design very similar to the one use for the treatment schools. Because we will need to match control and treatment schools, an oversample of schools to facilitate matching is need. A

forty percent increase in the sample size –to 120 schools- seems to balance statistical and budget concerns.

Because the measurement of student performance for impact of the EGRP will be reported by L1 and L2 learner groups, it is important to stratify by L1 and L2 learner like we do in the treatment sample. **Ten** pupils of each grades 1-3 will be selected in the sampled schools, a total of 30 pupils per school. The total sample size is shown below in table 3

Table A5. 3: Sample Size EGRP- Controls

	Number of Schools	Learner TYPE	Total Grade 1 pupils	Grade 2 pupils per school	Grade 3 pupils per school	Total pupils to be sampled
Controls	120	L1	600	600	600	3600
		L2	600	600	600	

For control schools, unfortunately we do not have the number of L1 and L2 learners. We will use therefore the proportion of L1 and L2 population in each VDC/Municipality as a proxy. We will assume that the number of L1 and L2 learners is identical to the proportion of L1 and L2 population.

As it is the case with treatment schools, by categorizing schools as percent of learners who are Nepali speakers, we are able to adjust the number of schools required to achieve the desired proportion of L1 and L2 learners. The **Table A4.3: Sample Design Controls** shows the approximate number of pupils within schools categorized by percentage of Nepali speakers in the schools (using the population proxy). Column A shows the categories of schools and columns C and D show the approximate number L1 and L2 learners by these school categories. The "Percent of total" row shows that the proportion of learners in control schools is unbalanced and there are many more L2 learners. We need to oversample the number of L1 learners in order to achieve an approximately 50-50 split of L1 and L2 learners in the control sample. This is achieved by oversampling more schools with higher L1 learners and less schools with L2 learners. This adjustment is shown in column G. The final desired number of schools to be sampled is shown in column H.

Table A5. 4: Sample Design Comparison Schools

	Percentage of Nepali Speakers in School	Sum of Total number of Pupil (Grade 1-3)	Approx . percent LI Learners	Approx. percent L2 Learners	Proportion of Total Number of Pupils	Schools to be sampled proportionally	Adjustment	Number of Schools to be Sampled	Approx. Number of LI Pupils Sampled	Approx. Number of L2 Pupils Sampled
Cohort I	0-20	169318	16932	152386	73%	87	-37	50	150	1353
	20-40	17862	5359	12503	8%	9	-4	5	47	109
	40-60	22105	11053	11053	9%	11	-4	7	111	111
	60-80	13143	9200	3943	6%	7	6	13	268	115
	80-100	10772	9695	969	5%	6	39	45	1203	134
	TOTAL	233200	52238	180854	100%	120	0	120	1778	1822
	PERCENT OF TOTAL		22%	78%					49%	51%