USAID
FROM THE AMERICAN PEOPLE

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

Prepared under Contract No.: GS-10F-0033M/AID-OAA-M-13-00010, Tasking N7617.011.01

USAID READING AND ACCESS ENDLINE EVALUATION REPORT

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

JUNE 2020

Prepared under Contract No.: GS-10F-0033M/AID-OAA-M-13-00010, Tasking N7617.011.01

Submitted to:

USAID

Submitted by:

Dr. Alicia Menendez and Gregory Haugan

Contractor:

NORC at the University of Chicago
4350 East West Highway, 8th Floor
Bethesda, MD 20814
Attention: Varuni Dayaratna
Tel: 301- 634-9414; E-mail: Dayaratna-varuni@norc.org

DISCLAIMER

The authors' 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

ACRONYMS

ADS	Automated Directives System
ABE	Assistance to Basic Education
ACR	All Children Reading
CDC	Curriculum Development Centre
CLA	Central Level Agency
CWPM	Correct Words per Minute
DEC	Development Experience Clearinghouse
DEO	District Education Officer
DiD	Difference-in-Difference
DoE	Department of Education
EGR	Early Grade Reading
EGRA	Early Grade Reading Assessment
EGRP	Early Grade Reading Program
FY	Fiscal Year
GoN	Government of Nepal
ERO	Education Review Office
IE	Impact Evaluation
MoEST	Ministry of Education, Science, and Technology
MT	Mother Tongue
NCED	National Center for Education Development
NEGRP	National Early Grade Reading Program
NORC	NORC at the University of Chicago
RM	Reading Motivator
RP	Resource Person
SOW	Scope of Work
SMC	School Management Committee
SRM	Supplementary Reading Materials
TG	Teacher Guide
TLMs	Teaching and Learning Materials
TOC	Theory of Change
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 Government of Nepal's National Early Grade Reading Program (NEGRP) and the USAID-funded Early Grade Reading Program (EGRP) in Nepal.

PROJECT BACKGROUND

EGRP is a program of technical support to the NEGRP, implemented from March 2015 through October 2021 by RTI International and its partner organizations, Another Option, Plan International Nepal, Room to Read, and SIL LEAD. The program has two overarching goals: 1) Improve early grade reading performance of students in Grades 1-3; and 2) Build the GON's capacity to deliver an NEGRP that can be replicated nationwide.

Program EGRP 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 of early grade reading interventions. Component 3 works with local NGOs, school management and parent-teacher associations to conduct advocacy campaigns, capacity development trainings, reading materials development, and related activities.

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

In Cohort 1, NEGRP activities started in 2016. All public schools (called community schools in Nepal) in the six districts received the full NEGRP package, which we call Nepali L1 interventions and consists of:

- Distribution of Nepali teaching and learning materials (TLMs) such as teachers' guides, student workbooks, decodable readers, letter and word cards, and various charts; as well as supplementary reading materials.
- Ten-day in-service teacher training on the use of TLMs in 2016 and continuing training during the following year, which included head teacher and school management committee (SMC) member orientation;
- Teacher coaching, mentoring, and support implemented through reading motivators (RMs), who are teachers or resources persons within the GON system;
- Parent and community level engagement activities only in the first two years; and
- Public Service Announcements on the radio and newspapers to promote early grade reading in the community.

In Cohort 2, activities started only partially in the 2016-17 school year. At midline (2018) Cohort 2 was still in light intensity implementation mode, having received only some components of the NEGRP. This included 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 distribution; and public service announcements on the radio and newspapers to promote early grade reading in the community. The NEGRP was rolled out at a high intensity in Cohort 2 districts in the 2018-19 school year, receiving the same set of NEGRP interventions received by Cohort 1 schools.

In summary, Cohort 1 received the full intervention during four academic years and Cohort 2 received two years of light intensity intervention and two academic years of the full intervention.

Additional activities specifically targeting Nepali L2 learners ("NEGRP Nepali L2 interventions") were considered for inclusion in Cohort 2 districts. However, they were not implemented until academic year 2020-21 and therefore are not studied in this evaluation.

EVALUATION METHODOLOGY OVERVIEW

The main questions this IE seeks to answer focus on the extent to which the NEGRP improved the reading outcomes of native (L1) and non-native (L2) Nepali speakers. The IE also aimed to investigate the extent to which the Nepali L2 component of the program—if implemented—generated additional impacts for L2 learners. In addition, the IE seeks to answer questions regarding the extent to which the NEGRP resulted in changes in teachers' reading instruction practices in the classroom, and the extent to which it generated changes in school management support for EGR.

The NEGRP 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 comparison districts (Doti, Myagdi, Kapilvastu, Bara, Sunsari, and Kavre) 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 endline in the treatment group with the changes between baseline and endline 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

Our evaluation is guided by 5 main evaluation questions.

EQ1: To what extent did NEGRP (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 NEGRP (Nepali L1 program) improve the reading outcomes of pupils who speak Nepali as a second language (L2 learners) in cohorts 1 and 2?	At endline the NEGRP had positive effects on all measured reading skills for L1 learners in cohorts 1 and 2. Smaller effects were found for L2 learners and in some cases there are no effects, particularly for cohort 1.

For all EGRA subtasks and across grades the effect of NEGRP at endline is positive and, in many cases, statistically significant for L1 learners. There are some subtasks for which the effect, although positive, is not statistically significant at conventional levels. In general, the lack of statistical significance seems to be the result of a sample underpowered to detect effects of that size[1].

The effect of the NEGRP is positive for grade 1 L2 learners for all subtasks. However, the effects for grades 2 and 3 tend to be small and are not statistically significant.

In the table below, we summarized the effect of NEGRP on oral reading fluency for both cohorts. The changes due to the program are positive with the exception of grade 2 L2 learners in cohort 1.

Adjusted Difference in Difference Effects of NEGRP on Oral Reading Fluency								
	Cohort 1				Cohort 2			
	L1 learners		L2 learners		L1 learners		L2 learners	
	CWPM	Effect Size	CWPM	Effect Size	CWPM	Effect Size	CWPM	Effect Size
Grade 1	2.8	0.28	1.5***	0.42	5.0***	0.46	2.6***	0.51
Grade 2	5.3	0.27	-0.1	-0.01	11.1***	0.51	2.4	0.21
Grade 3	11.9*	0.44	3.7*	0.19	12.0***	0.49	8.5***	0.42

Note: CWPM=correct words per minute. 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.

Despite these positive NEGRP effects, reading performance remains low, particularly for L2 learners. For example, the average oral reading fluency among L1 treated students is 33 correct words per minute in grade 3 in both cohorts. Similarly to baseline and midline, the average reading scores for L2 learners are much lower than those of L1 learners. L2 learners' scores are approximately one full grade behind those of L1 learners. Among L2 learners the average oral reading fluency is around 15 and 19 correct words per minute in cohort 1 and 2, respectively.

Although NEGRP tends to benefit all learners, it is clear that the impact of the program on ORF is lower for L2 than for the L1 learners, particularly in grades 2 and 3, and the gap between the two groups was not reduced.

EQ3: To what extent did the NEGRP Nepali L2 intervention improve the reading outcomes of pupils who speak Nepali as a second language (L2 Learners) in cohort 2?	It is not possible to answer this question as the NEGRP Nepali L2 intervention was not implemented until the academic year 2020 21, after the endline was conducted.

EQ4: To what extent has the NEGRP Nepali L1 program changed teachers' reading instruction practices in the classroom?	The teaching reading instruction practice index shows a positive impact of the NEGRP for both cohorts.

[1] The impact is always statistically significant when analyzing all grades together.

We created two indexes to measure teachers' reading instruction practices in the classroom. The first index –Index I- includes 30 items describing desirable actions during an early grade reading lesson. For example, this included teaching and practicing letter sounds, reading independently, introducing vocabulary, using teaching and learning materials appropriately, etc. The Teacher Reading Instruction Practices Index—Index II— uses calculation guidelines from USAID. This index includes a subset of questions used in Index I, but requires specific combinations of teaching practices that reflect categories such as phonemic awareness instruction, fluency modeling, reading comprehension exercises, etc. Both indexes show an improvement in teaching practices among teachers from both cohorts.

EQ5: To what extent has the NEGRP Nepali L1 program changed the school leadership and management index (as defined in the monitoring index), demonstrating active support for EGR?	The NEGRP Nepali L1 program has generated a modest improvement of 1.2 points (out of 14) in the management index for cohort 2. However, at endline there is no impact in the index among schools in cohort 1.

USAID/Nepal and the EGRP team defined a School Leadership and Management Index. The index includes 14 items related to the school priorities, actions devoted to promote reading, parental involvement, student reading performance monitoring, etc. This information was collected through interviews with head teachers and SMC members and classroom observations, and each item was weighted equally, resulting in an index that extends from 0 to 14. In cohort 1, the index was positively impacted by the program at midline. The impact at midline was 0.9 points and was statistically significant but was no longer present at endline. Cohort 2 shows a statistically significant effect for NEGRP on the index at endline of 1.2 points.

CONCLUSIONS

Based on the facts presented above and additional findings included in the report, we can conclude the following:

The NEGRP had positive effects at the endline among learners in cohorts 1 and 2. The effects of the program are similar in each cohort, giving us greater confidence in the findings. Additionally, the findings for cohort 2 at endline are similar to the effects found at midline for cohort 1, where the program had already been fully rolled out. In contrast, the lack of findings for cohort 2 at midline, where the program had not yet been fully rolled out, confirms a key assumption of the analytical methodology: that in the absence of the NEGRP interventions, the treatment and comparison groups move in parallel along a similar trajectory. When combining this finding with the impacts found for cohort 2 at endline, by which time implementation had been fully rolled out, we can confidently attribute causality to the NEGRP for the improvements in reading outcomes seen in the treatment groups.

Overall, reading performance indicators improved for treatment learners. However, **there is still room for improvement**. Most grade 1 learners are non-readers and by grade 3 around a quarter of them are still not able to read a single word from a connected paragraph. Oral reading fluency is still low for all grades and very few learners reach the GON's reading benchmark of 45 cwpm and 80% reading comprehension.

In cohorts 1 and 2, both L1 and L2 learners benefited from the program. This is highly desirable given that the performance of both groups of students is far below the levels that the GoN considers to be the minimum reading standards. **However, the program benefitted L1 learners more than L2 learners.** As noted at baseline and midline, 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. NEGRP was able to improve performance among L2 learners but not enough to reduce the disadvantage they experience as non-Nepali speakers. L2 learners not only lag behind L1 learners in terms of their reading skills, but there was also a lag in overall oral Nepali language comprehension among L2 learners.

The NEGRP has benefited students with both low and high performance. An improvement in reading performance was found across groups of learners with different reading abilities. NEGRP reduced the number of zero scores among learners and also increased the percentage of learners that reach the benchmark of 45 correct words per minute and 80 percent oral reading comprehension that the GoN has adopted.

Examining the channels through which the program functioned, **there is no evidence that the program has led to changes in parents' at-home support** for their children's reading development. However, parent support for reading development seems quite high for all groups. **SMC support for reading activities shows a very modest improvement for cohort 2 and no improvement for cohort 1.**

There is evidence that the program has had a **positive effect on teachers' reading instruction**, as captured by the classrooms observation exercise. The percentage of teachers conducting desirable reading instruction activities in class has increased in both treatment cohorts and it is higher than in comparison groups. At the same time, it is important to mention that we recommend, in Section 6, a different and more rigorous approach to assess the quality of teaching.

Support supervision of teachers is still not universal. Although treatment teachers have higher probability of receiving support, there is still a significant fraction of teachers that reported receiving no supervision at all.

The program was quite successful at ensuring access to materials, including students' access to Nepali-language workbooks, and additional children's reading materials, and teachers' access to teaching guidelines, materials, and curriculum. Almost all teachers reported using these resources. Thus, it is **likely that the positive effects of the program functioned via a combination of improved teaching practices plus broad access to and use of learning and teaching materials.**

RECOMMENDATIONS

A number of recommendations stem from our findings:

Special attention to L2 learners: Similar to what we found at midline, the disadvantage in early grade reading skills of L2 learners relative to L1 learners was evident at endline. NEGRP was able to improve performance among L2 learners but not enough to reduce the disadvantage they experience as non-Nepali speakers. 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 supporting non-native 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 non-Nepali language learners in their classrooms.

Improve teacher support supervision: A larger fraction of teachers in both cohorts received more frequent support supervision than comparison groups; however, there are still many teachers who do not receive any supervision at all. Evidence suggests that including follow-up classroom visits and teacher support increases learning gains (see for example, 2018 World Development Report). We recommend exploring this challenge and how to effectively scale support supervision within the education system to ensure sustainability of the program.

Improve SMC role: SMC support for reading activities does not show substantial improvement. This component of the program requires revision and in-depth assessment to understand its challenges and effectiveness.

Parental engagement: the approach used to measure parental engagement was to ask about the importance of learning reading in early grades, reading activities with children at home, and parents' opinions about their educational responsibilities. Parents seem to be well aware of the importance of reading and their role in enabling the process. Most parents also think that teaching how to read is a joint endeavor between the school and the home and that even illiterate parents can help their children. These parents' opinions suggest that raising parental awareness about the importance of early reading is not a priority. Independently of whether or not parents' actual behavior reflects what they report, they seem to be well informed about the issue already. We recommend that in the future, qualitative research is conducted through focus group discussions with parents, to learn more about their actual behaviors rather than opinions, and to identify the difficulties they may face when trying to support their children's learning process. This type of research can inform strategies to guide parents in future programs.

1. 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 provides technical assistance to the Ministry of Education, Science, and Technology's (MoEST) National Early Grade Reading Program (NEGRP). The EGRP-Nepal, implemented by RTI International, the MoEST and its Central Level 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 NEGRP 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, EGRP-Nepal, practitioners, researchers, the donor community and NGOs operating in the education sector in Nepal. The evaluation findings will be used to inform programmatic decisions and guide roll-out of NEGRP to additional districts, 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 endline evaluation of the NEGRP activities. We show learners' reading performance at different points in time: baseline (2016), midline (2018) and endline (2020), and details of the program fidelity of implementation at endline and over time.

1.1. 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 (CDC), the Center for Education and Human Resource Development (CEHRD, new entity formed by merging previous Department of Education, National Center for Educational Development and Non-Formal Education Center), and the Education Review Office (ERO) —to develop and test an early grade reading program that is effective, replicable, cost-efficient, and sustainable.

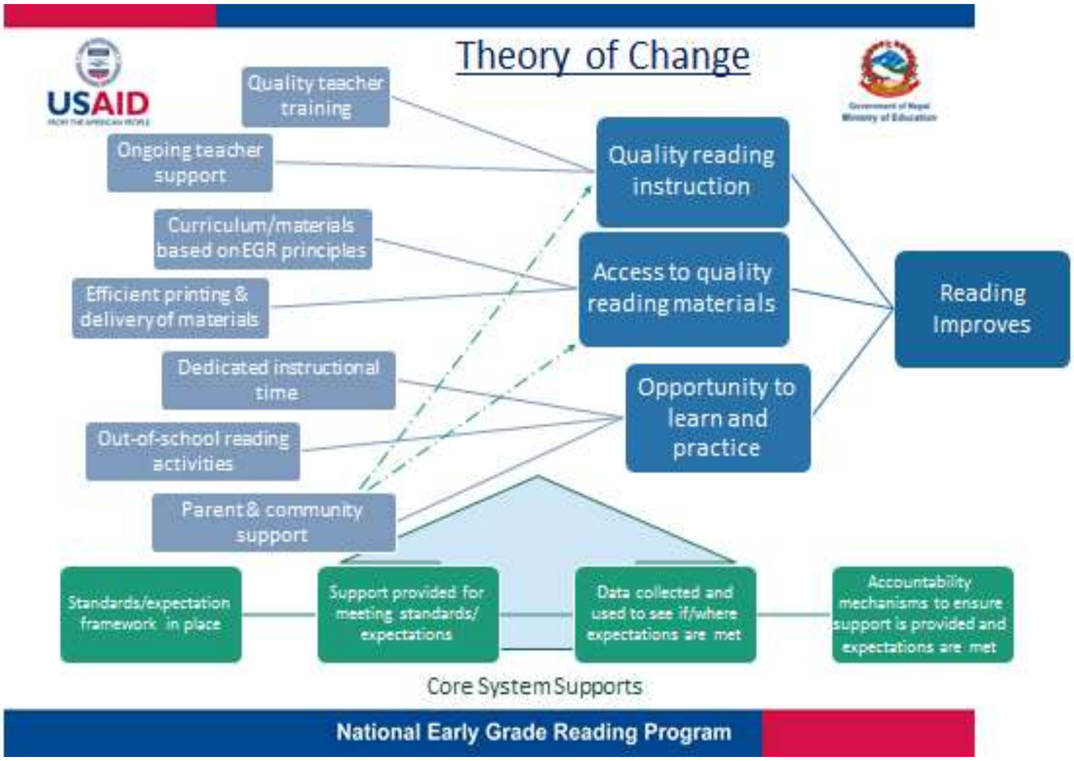
The NEGRP has two principal goals: 1) To improve early grade reading performance of students in Grades 1-3; and 2) To build the GoN's capacity to deliver an early grade reading program that can be replicated nationwide.

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 improved reading outcomes. Specifically, USAID/Nepal hypothesizes that the NEGRP will improve the reading skills of both L1 and L2 learners.

Figure 1: NEGRP Theory of Change


Theory of Change
USAID
Quality teacher training
Ongoing teacher support
Curriculum/materials based on EGR principles
Efficient printing & delivery of materials
Dedicated instructional time
Out-of-school reading activities
Parent & community support
Quality reading instruction
Access to quality reading materials
Opportunity to learn and practice
Reading Improves
Standards/expectation framework in place
Support provided for meeting standards/ expectations
Data collected and used to see if/where expectations are met
Accountability mechanisms to ensure support is provided and expectations are met
Core System Supports
National Early Grade Reading Program

Focusing on grades 1, 2, and 3, NEGRP was rolled out in 2 cohorts. Cohort 1 includes 6 districts[2] - 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, NEGRP activities started in 2016. All public schools, called community schools in Nepal in the six districts received the full NEGRP package, which we call Nepali L1 interventions and consists of:

- Distribution of Nepali Teaching and Learning Materials (TLMs): teachers' guides, learners' readers, decodables and workbooks, letter and word cards, and various charts;
- Ten-day in-service teacher training on the use of TLMs in 2016 and continuing training during the following year, which included head teacher and school management committee (SMC) member orientation;
- Teacher coaching, mentoring, and support model implemented through reading motivators (RMs), who are teachers or resources persons within the GoN system (during the 2017-2018 school years) or through Head Teachers and Primary-in-Charges (during the 2019-2020 school year);
- Parent and community level engagement activities in the first 2 years; and
- Public Service Announcements on the radio and newspapers to promote early grade reading in the community.

[2] There are 77 districts in Nepal

In cohort 2, activities started only partially in the 2017-18 school year. At midline (2018) cohort 2 was still in light intensity implementation mode, having received only some components of the NEGRP, 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 NEGRP was rolled out in high intensity in cohort 2 districts in the 2018-19 school year, receiving the same set of NEGRP interventions received by cohort 1 schools.

Summarizing, cohort 1 received the full intervention during 4 academic years and cohort 2 received two years of light intensity intervention and two academic years of full intervention.

For a while, some additional activities specifically targeting Nepali L2 learners were considered for inclusion in cohort 2 districts. These activities are denominated NEGRP Nepali L2 interventions. The NEGRP team piloted some of these activities in areas not included in this evaluation. The final decision was to delay NEGRP Nepali L2 activities until the academic year 2020-21 and therefore are not studied in this evaluation.

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 NEGRP 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 NEGRP Nepali L1 program improve the reading outcomes of pupils who speak Nepali as a second language (L2 learners) in cohorts 1 and 2?

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

Note: it is not possible to answer this question as it was decided to delay NEGRP Nepali L2 intervention until the academic year 2020-21.

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

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

Q5. To what extent has NEGRP 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 NEGRP 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 changes in EGRA outcomes of the treatment group with the average changes in outcomes among a statistically matched comparison group of schools.
Q2. To what extent did NEGRP Nepali L1 program improve the reading outcomes of pupils who speak Nepali as a second language (L2 learners) in cohorts 1 and 2?		
Q3. To what extent did the NEGRP Nepali L2 intervention improve the reading outcomes of pupils who speak Nepali as a second language (L2 Learners) in cohort 2?		NEGRP Nepali L2 interventions did not take place before the endline and therefore it is not possible to answer this question.
Q4. To what extent has the NEGRP 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.
Q5. To what extent has NEGRP Nepali L1 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

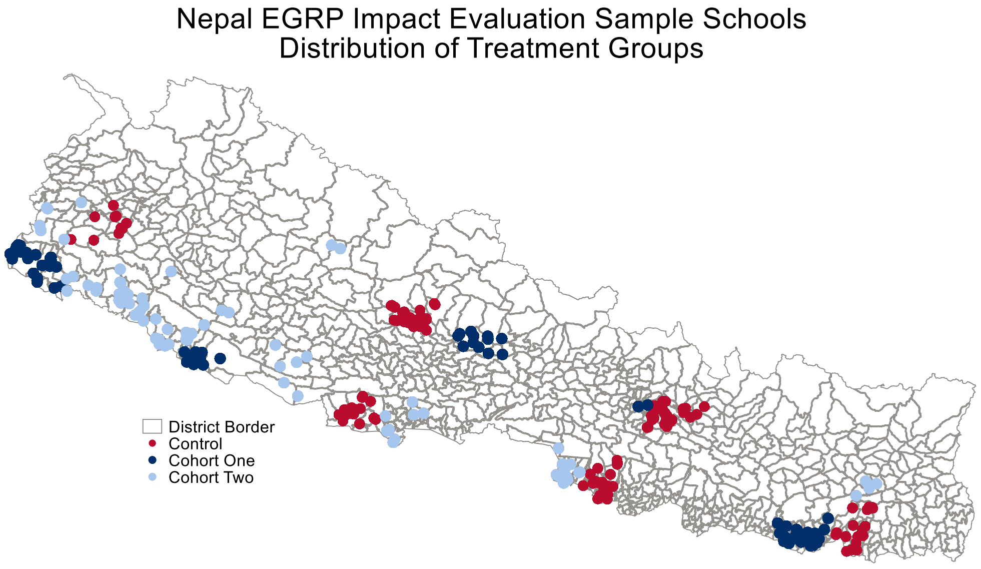
This evaluation uses a quasi-experimental design to measure impact. Using statistical techniques, NORC created a credible group of comparison schools against which to measure changes in schools that received the NEGRP. In this section, we explain the approach.

2.2.1. TREATMENT ASSIGNMENT

The NEGRP focuses on grades 1, 2, and 3 and was rolled out in 2 cohorts of districts. Cohort 1 includes 6 districts while Cohort 2 covers 10 additional districts. The EGRP team, the MoEST and USAID/Nepal decided that all community schools within a treatment district would receive NEGRP interventions and, therefore, comparison schools would necessarily need to be found in other districts. To this end, the EGRP team 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 Cohort2 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


Nepal EGRP Impact Evaluation Sample Schools
Distribution of Treatment Groups
□ District Border
● Control
● Cohort One
● Cohort Two

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 groups of schools (one for each cohort) are 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 endline in the treatment group with the changes between baseline and endline 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 basic 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 handled by RTI and its partners in Nepal. CAMRIS International, the USAID's Monitoring, Evaluation, and Learning (MEL) contractor provided quality assurance oversight of the data collection process. The NORC evaluation team provided support in preparation and during training of enumerators. The collection of data used in this evaluation followed 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 the 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: Endline data was collected at the end of the 2019-20 academic year, in February-March 2020. The NORC evaluation team received these data on May 7, 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 proficiency[3]. 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 (Grade 2 level)	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

[3] 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.

Early Literacy Skill	Sub test	Measurement
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 (Education Management efficiency Survey or EMES instruments) were developed by the government of Nepal and EGRP collaboratively and administered by the EGRP team. The NORC evaluation team contributed to the final version of the endline tools.

- 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 Nepali subject teacher, preference for grade 2 teacher
- SMC Questionnaire: administered to the SMC chair or most active member in schools 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 Nepali subject.

All instruments can be found in Annex VII.

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

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 and endline data collections 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, Midline and Endline Samples

	Treatment Schools						Comparison Schools		
	Cohort 1			Cohort 2					
	Baseline	Midline	Endline	Baseline	Midline	Endline	Baseline	Midline	Endline
Schools visited	86	85	86	86	86	86	120	120	119
Parents	86	85	86	86	85	86	120	120	119

	Treatment Schools						Comparison Schools		
	Cohort 1			Cohort 2					
	Baseline	Midline	Endline	Baseline	Midline	Endline	Baseline	Midline	Endline
Teachers	86	85	86	86	86	85	120	120	119
Head teachers	85	85	86	86	86	85	120	120	119
SMC member	86	85	86	85	86	86	120	120	119
School Inventory	86	85	86	86	86	86	120	120	119
Classroom Inventory	85	84	86	86	86	85	120	118	117
Classroom Observations	86	84	85	85	85	85	120	119	117
Grade 1 learners	839	782	840	870	821	851	1110	993	1,072
Grade 2 learners	870	822	828	842	811	854	1057	1014	1,044
Grade 3 learners	885	827	849	882	829	833	1132	984	1,065
Total learners	2594	2431	2,517	2594	2461	2,538	3299	2991	3,181

Table 3 shows the number of schools visited in each group at baseline, midline and endline, 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 PILOTING

The EGRA instruments used in the NEGRP impact evaluation at baseline, midline and endline were based on the tools developed through the national EGRA conducted in 2014 through Ed Data II.[4] A three-day workshop to adapt the EMES instruments was held in November 2013 in Kathmandu, followed by two rounds of pre-testing and revisions before instrument finalization. Workshop participants included Ministry of Education (MoE), Department of Education (DoE), and Curriculum Development Center (CDC), National Center for Education Development (NCED), District Education Officers (DEOs), Resource Persons (RPs), Education Training Centre (ETC) instructors, head teachers, and representatives from USAID, World Bank, Save the Children, and Room to Read.

Similarly, a workshop to adapt the EGRA instrument, the student interview, and an assessment instrument for teachers was held January 2014 in Kathmandu. Representatives from the MOE, DoE, CDC, and ERO (Education Review Office), as well as international/ nongovernmental organizations, attended the workshop. Over the course of the three days, attendees drafted and agreed upon the subtasks of the EGRA instrument, adapted the student interview, and created the teacher instrument. A half-day of field testing at a local school was included.

[4] USAID Education Data for Decision Making (EdData II) (2014). *Nepal Early Grade Reading Assessment (EGRA) Study.* Research Triangle Park, NC: RTI International.

Some further adaptations were conducted for the purposes of the NEGRP impact evaluation at baseline, in collaboration with the Education Review Office (ERO). For example, additional passages were developed for the ORF and reading comprehension subtasks; while the order of the individual items in the other subtasks was shuffled. It was also decided to extend the time for the reading passage so that it could more accurately measure comprehension, with the understanding that L2 learners may read more slowly than L1 learners. Subsequently, a rapid pilot test was conducted to understand the level of difficulty of the items, and discrimination analysis was conducted to understand the item discrimination capacity and measure the consistency of the tools. The items were then reviewed and approved by the ERO Subject Committee before use in the assessment.

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), the 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 were made as long as they did not impede comparability to the baseline data collection. The EGRA instruments remained the same at the endline, but some items of the EMES instruments were revised for example, the classroom observation tool, based on the insights gained during previous data collections.

2.2.5. ASSESSOR TRAINING

Training for the NEGRP Endline Assessment was held in Kathmandu, Nepal from Sunday January 5 to Friday January 17, 2020. Training was facilitated by ERO staff, the EGRP field office staff with support from RTI Headquarters, NORC at the University of Chicago, and FEDUC, EGRP's local data collection subcontractor.

Training and piloting were conducted in three separate phases. Phase I consisted of 3 days of classroom training plus 1 day of field practice and included 50 senior data collectors, 25 of which were to be selected as team leaders and 25 of which were to be selected as EMES administrators. Phase II consisted of 4 days of classroom training plus 2 days of field practice, and included three separate tracks: EGRA administrator track, EMES administrator track, and team leader track. Finally, a 2-day dry run was conducted on January 16-17, prior to the formal launch of data collection on January 23.

NORC observed all classroom trainings as well as 2 school-based pilots. English translators were present for the training and accompanied NORC, USAID, CAMRIS, and the RTI HQ team to pilot schools, allowing for close monitoring of activities. Following the first day of training, NORC developed a cloud-based training feedback log so that issues or concerns observed during the training could be logged and monitored by the EGRP team in real-time. Overall, the EGRP team was able to address the great majority of issues in a timely fashion, and to NORC's satisfaction.

Key strengths of the training include:

- The development of in-depth standard operating procedures (SOPs) which accompanied most data collection tools.

- The large number of days allocated for classroom training and field practice.
- EGRP's responsiveness to feedback from NORC, CAMRIS and USAID, and willingness to quickly address/remedy observed issues.
- Logistics and organization.

2.2.6. FIELDING THE SURVEY AND QUALITY ASSURANCE

Data collection for the endline took place in January/February, 2020. 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 three EGRA assessors, one EMES surveyor and one team supervisor were deployed in each school. The supervisor 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. 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/4G 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.

2.2.7. DATA GENERATION, CLEANING, AND FINALIZATION

As noted above, the EGRA endline 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, 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.

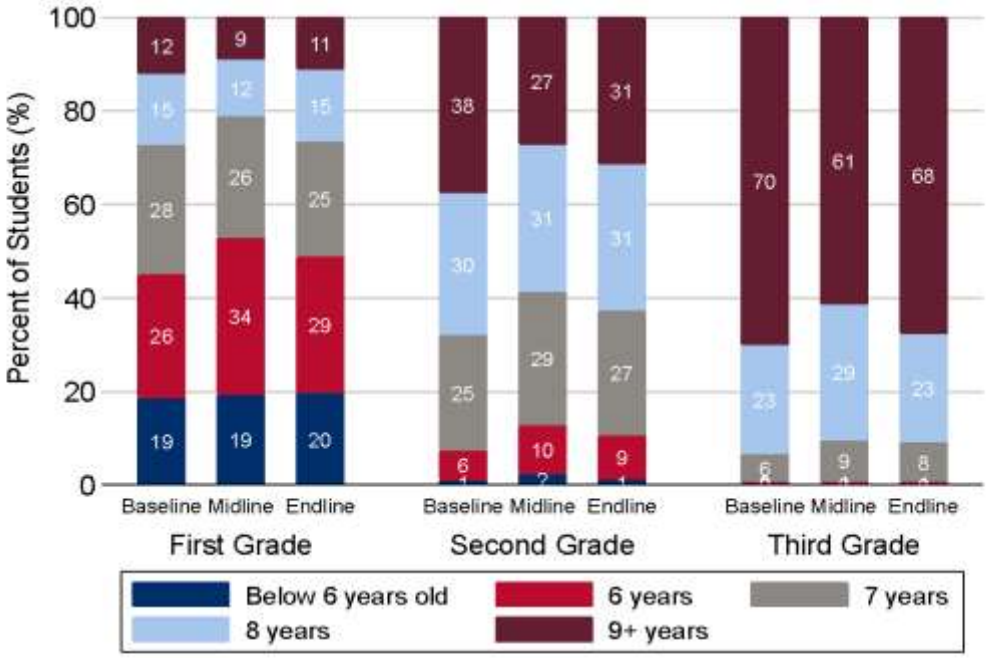
The dataset was delivered to NORC via a secure server. The final version of the data set was received by NORC on May 7th 2020.

3. FINDINGS

3.1 LEARNER CHARACTERISTICS

More than half the learners assessed were girls in the baseline (55.2%), midline (55.2%), and endline (55.6%) samples. Figure 3 presents learners' age distribution in 5 categories: below 6 years of age, 6 years, 7 years, 8 years, and 9 years 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 baseline, midline, and endline samples are 8.1, 7.8, and 8.0, respectively.

Figure 3: Age Distribution by Grade and Survey Round



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		Endline	
	Learners	Teachers	Learners	Teachers	Learners	Teachers
Nepali (L1)	32.6%	49.0%	33.5%	47.0%	36.1%	44.3%
Non-Nepali (L2)	67.4%	51.0%	66.5%	53.0%	63.9%	55.7%
Maithali	19.5%	16.5%	21.4%	16.1%	20.6%	15.1%
Bhojpuri	26.9%	15.7%	27.8%	17.5%	27.6%	15.6%
Tharu	11.7%	14.9%	15.1%	14.2%	13.8%	19.7%
Tamang	2.3%	2.6%	1.6%	2.9%	1.3%	5.6%
Awadhi	12.5%	0.0%	13.1%	5.9%	13.2%	4.3%
Others	27.0%	50.2%	21.0	43.4%	23.5	39.6%
Observations	8487	292	7883	291	8236	291

Note: Sample weights applied to recover population representativeness

In Table 5 we summarize 3 reading indicators –the percentage of non-readers (zero cwpm), the average oral reading fluency (cwpm), and the percentage reaching the reading benchmark. We show these indicators by treatment group and detailed by grade at different points in time, baseline, midline and endline.

Table 5. Percentage of non-readers, average oral reading fluency and percentage reaching the reading benchmark, by group, grade and wave

	Grade 1			Grade 2			Grade 3		
% Non readers (zero cwpm)	**Baseline**	**Midline**	**Endline**	**Baseline**	**Midline**	**Endline**	**Baseline**	**Midline**	**Endline**
Treatment	85.8%	79.1%	73.1%	57.7%	52.6%	44.6%	38.5%	31.2%	23.8%
Cohort 1	84.9%	70.1%	69.8%	61.7%	46.8%	48.7%	44.5%	29.5%	27.0%
Cohort 2	86.3%	84.3%	74.9%	55.4%	56.0%	42.2%	35.3%	32.2%	22.0%
Comparison	89.5%	90.8%	88.0%	62.0%	68.4%	61.2%	41.8%	39.0%	41.1%
ORF (cwpm)	Baseline	Midline	Endline	Baseline	Midline	Endline	Baseline	Midline	Endline
Treatment	1.5	2.3	3.6	7.1	8.4	12.6	15.7	17.5	23.2
Cohort 1	1.7	4.2	3.4	7.0	10.8	10.8	13.3	19.4	21.2
Cohort 2	1.5	1.2	3.5	7.1	7.0	13.7	17.0	16.5	24.4
Comparison	1.2	0.8	0.9	6.4	4.6	6.2	14.1	12.5	13.8
% Reaching Benchmark	Baseline	Midline	Endline	Baseline	Midline	Endline	Baseline	Midline	Endline
Treatment	0.1%	0.2%	0.6%	0.9%	2.0%	3.2%	5.9%	7.6%	9.4%
Cohort 1	0.2%	0.5%	0.8%	1.6%	3.4%	3.7%	5.1%	9.6%	10.7%

	Grade 1			Grade 2			Grade 3		
% Non readers (zero cwpm)	Baseline	Midline	Endline	Baseline	Midline	Endline	Baseline	Midline	Endline
Cohort 2	0.1%	0.0%	0.4%	0.5%	1.2%	3.0%	6.3%	6.6%	8.6%
Comparison	0.0%	0.5%	0.0%	0.8%	0.2%	1.0%	5.3%	3.2%	4.3%

Note: Sample weights applied to recover population representativeness

As expected, learners' reading performance goes up with grade. Learners in grade 3 perform better than their counterparts in grade 2 who, in turn, perform better than learners in grade 1. For example, at endline the percent of non-readers among treatment learners is 73 percent in grade 1, 45 percent in grade 2 and 24 percent in grade 3. It can also be seen that the group that received NEGRP activities, improved its performance from baseline to endline in all 3 indicators; the percentage of learners not able to read a single word went down among treated learners, and oral reading fluency (cwmp) and percentage reaching the Nepal reading benchmark of 45 cwpm and 80% comprehension went up for all grades. In contrast, the comparison group performance remained the same during the period. However, the treatment and comparison groups are not strictly comparable because the populations they represent are different. To be able to compare the groups and estimate the effect that NEGRP had on the treatment learners, we need to take into account those differences. This is what we do in the next subsection, 3.2 NEGRP Impact on EGRA Scores at Endline.

In Annex IV (Tables A4.8-10) we present the evolution of these indicators in more detail for L1 and L2 learners, and for boys and girls separately.

3.2 NEGRP IMPACT ON EGRA SCORES AT ENDLINE

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

EQ1: To what extent did NEGRP (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 NEGRP (Nepali L1 program) improve the reading outcomes of pupils who speak Nepali as a second language (L2 learners) in cohorts 1 and 2?	At endline the NEGRP had positive effects on all measured reading skills for L1 learners in cohorts 1 and 2. Smaller effects were found for L2 learners and in some cases there are no effects, particularly for cohort 1.

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 NEGRP interventions beginning in 2016. During the 2018 midline, large effects were found on all measured reading skills for cohort 1. Cohort 2 had only received a light intensity version of NEGRP by the 2018 midline, and no impacts were found for cohort 2 at midline. Thus, the endline represents the first opportunity to measure the impacts of the high intensity set of interventions for cohort 2, two years after it first began.

3.2.1. COHORT 1

We show the effect of the NEGRP on each skill measured by the EGRA. Table 6 and 7 show results by grade for L1 and L2 learners respectively. The tables show the mean score for each EGRA subtask at baseline and endline, 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 NEGRP we compare the change between baseline and endline in the treatment group to the change between baseline and endline 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.).[5]

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 endline, the averages are 22.4 and 33.1 correct words for comparison and treatment groups respectively, amounting to a difference between the groups of 10.7 words. The simple DiD between the groups is, therefore, 10.9 words (10.7-(-0.2)). When we adjust to take into account learners' basic characteristics, the adjusted DiD is 11.9 words, showing a clear advantage of the cohort 1 treatment group over their comparison counterparts. The effect is statistically significant at the 90 percent confidence level and its size is equivalent to 0.44 of a standard deviation. This effect is large; to give a sense of its magnitude, an improvement of 11.9 words due to the NEGRP is almost as large as the difference between average ORF in Grade 2 (ORF=10 cwmp) and in Grade 3 (ORF=22.4 cwpm) for the comparison group at endline, for example.

Table 6: Effect of NEGRP on EGRA Subtasks, cohort 1, Nepali (L1) Learners, by Grade

	Baseline			Endline			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	11.5	16.9	5.4	13.2	18.1	4.9	-0.5	0.5	0.04
Grade 2	22.7	25.2	2.5	19.7	29.6	9.9	7.4*	6.3	0.34
Grade 3	33.5	33.0	-0.5	30.1	40	9.9	10.4*	11.0*	0.52
Correct Matra Per Minute [Max=100]									
Grade 1	3.7	5.8	2.1	4.6	9	4.4	2.3	3.1	0.25
Grade 2	12.7	13.4	0.7	11.6	19.9	8.3	6***	6.6**	0.34
Grade 3	21.4	21.9	0.5	24.7	32.3	7.6	7.1	7.9	0.32
Correct Invented Words Per Minute [Max= 50]									
Grade 1	0.9	1.2	0.3	0.8	2.3	1.5	1.2**	1.5**	0.38
Grade 2	3.5	4.1	0.6	3	5.9	2.9	2.3*	2.1	0.30
Grade 3	6.1	6.8	0.7	7.7	11.2	3.5	2.8*	3.3*	0.33
Oral Reading Fluency									
Grade 1	2.4	3.4	1.0	2.7	5.8	3.1	2.1	2.8	0.28

[5] 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.

	Baseline			Endline			DiD	Adjusted DiD	Adj. Effect Size
	Comp (1)	Treat (2)	Diff (3 2 1)	Comp (4)	Treat (5)	Diff (6 5 4)	(7 6 3)	(8)	(9)
Grade 2	10.2	10.7	0.5	10	16.8	6.8	6.3**	5.3	0.27
Grade 3	18.4	18.2	-0.2	22.4	33.1	10.7	10.9*	11.9*	0.44
Reading Comprehension, Percentage Correct									
Grade 1	3.9	5.9	2.0	5.2	10.3	5.1	3.1	4.2	0.25
Grade 2	17.7	17.6	-0.1	15.2	26.6	11.4	1.5***	10.0**	0.37
Grade 3	32.5	28.5	-4.0	37	44.4	7.4	11.4	13.0	0.38
Listening Comprehension, Percentage Correct									
Grade 1	15.1	15.8	0.7	12.9	23.1	10.2	9.5	11.3*	0.45
Grade 2	27.1	26.0	-1.1	21.6	33.6	12	13.1**	12.2**	0.42
Grade 3	35.4	36.4	1.0	35	48.5	13.5	12.5***	3.1***	0.40

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 6 shows that for all EGRA subtasks and across grades the effect of NEGRP at endline is positive and, in many cases, statistically significant for L1 learners. There are some subtasks for which the effect, although positive, is not statistically significant at conventional levels. In general, the lack of statistical significance seems to be the result of a sample underpowered to detect effects of that size[6].

Despite the positive NEGRP effects, reading performance remains low. For example, the average oral reading fluency among L1 treated students is less than 6 cwpm in grade 1, less than 17 in grade 2, and 33 in grade 3.

Table 7 replicates the information presented in Table 6 but now focusing on L2 learners. There are several findings worth noting. First, similarly to baseline and midline, the average scores for L2 learners are much lower than those displayed in Table 6 for L1 learners. L2 learners' scores are approximately one full grade behind those of L1 learners. This is true for all reading subtasks in which learners were assessed. The difference is also very large in the Listening Comprehension subtask, suggesting deficiencies in overall oral Nepali language comprehension among L2 learners rather than problems in reading skills exclusively.

Second, the effect of the NEGRP is positive for grade 1 L2 learners for all subtasks. However, the effects for grades 2 and 3 tend to be small and are not statistically significant. The absolute levels of reading competence remain very low, on average, for this group. For example, in grade 3, L2 learners that received NEGRP treatment read on average only 15.2 words per minute and only answered 29.4 percent of the listening comprehension questions correctly.

Finally, comparing Tables 6 and 7, it is clear that the impact of the program is lower for L2 learners than for L1 learners, particularly in grades 2 and 3. NEGRP was able to improve performance among L2

[6] The impact is always statistically significant when analyzing all grades together.

learners but not enough to reduce the disadvantage they experience as non-Nepali speakers, further increasing the gap between the two groups.

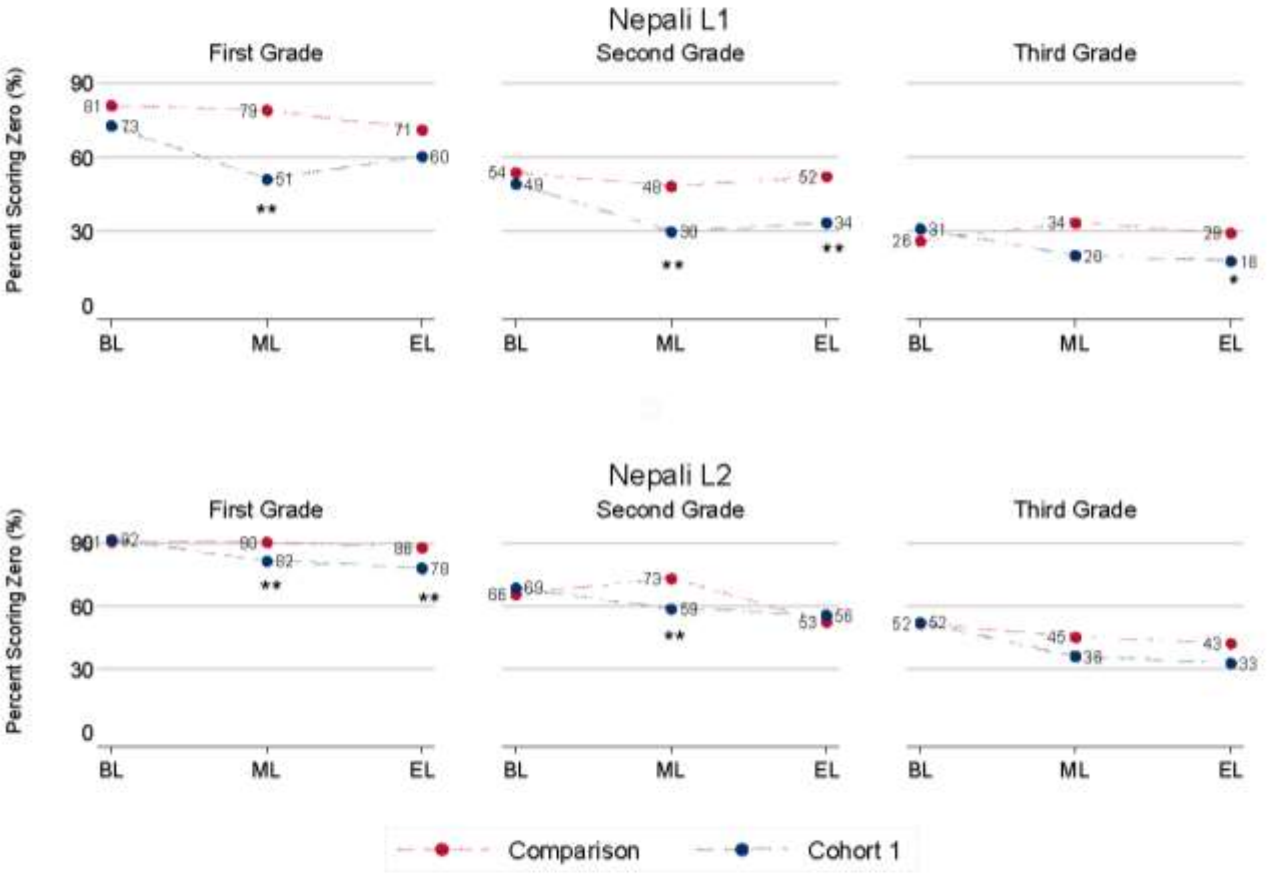
Table 7: Effect of NEGRP on EGRA Subtasks, cohort 1, Non-Nepali (L2) Learners, by Grade

	Baseline			Endline			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 Sound of Letters Per Minute [Max=100]									
Grade 1	8.1	6.5	-1.6	5.8	9.7	3.9	5.5***	6.3***	0.6
Grade 2	17.3	16.4	-0.9	18.8	17.4	-1.4	-0.5	-0.7	-0.04
Grade 3	23.2	22.6	-0.6	23.1	27.6	4.5	5.1	4.8	0.24
Correct Matra Per Minute [Max=100]									
Grade 1	1.7	1.5	-0.2	1.4	3.3	1.9	2.1**	2.3**	0.37
Grade 2	6.3	7.6	1.3	9.4	9.4	0	-1.3	-1.5	-0.1
Grade 3	12.8	12.8	0.0	16	19.8	3.8	3.8	3.3	0.16
Correct Invented Words Per Minute [Max=50]									
Grade 1	0.4	0.4	0.0	0.1	1.1	1	1.0***	1.1***	0.48
Grade 2	1.6	2.2	0.6	2.3	3.1	0.8	0.2	0.2	0.04
Grade 3	3.7	4.4	0.7	4.4	6.6	2.2	1.5	1.3	0.16
Oral Reading Fluency [Max=100]									
Grade 1	0.7	0.6	-0.1	0.4	1.6	1.2	1.3***	1.5***	0.42
Grade 2	4.0	4.5	0.5	5.6	6.2	0.6	0.1	-0.1	-0.01
Grade 3	9.5	9.6	0.1	11.2	15.2	4	3.9*	3.7*	0.19
Reading Comprehension, Percentage Correct									
Grade 1	1.0	0.7	-0.3	0.4	2.4	2	2.3***	2.8***	0.42
Grade 2	6.3	6.5	0.2	7	9.6	2.6	2.4	2.4	0.14
Grade 3	12.6	13.4	0.8	14.7	22	7.3	6.5**	6.3*	0.24
Listening Comprehension, Percentage Correct									
Grade 1	4.1	5.3	1.2	6.8	15.8	9	7.8**	8.5**	0.39
Grade 2	12.1	13.5	1.4	16.2	17.6	1.4	0.0	-0.4	-0.01
Grade 3	15.8	17.4	1.6	27	29.4	2.4	0.8	-1.2	-0.04

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 NEGRP 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 correct words read per minute, part of the goal of the NEGRP is to target the weakest learners in order to reduce the prevalence of illiteracy or zero scores.

Figure 4: Percentage of zero score in Oral Reading Fluency, cohort 1, by grade and learner language

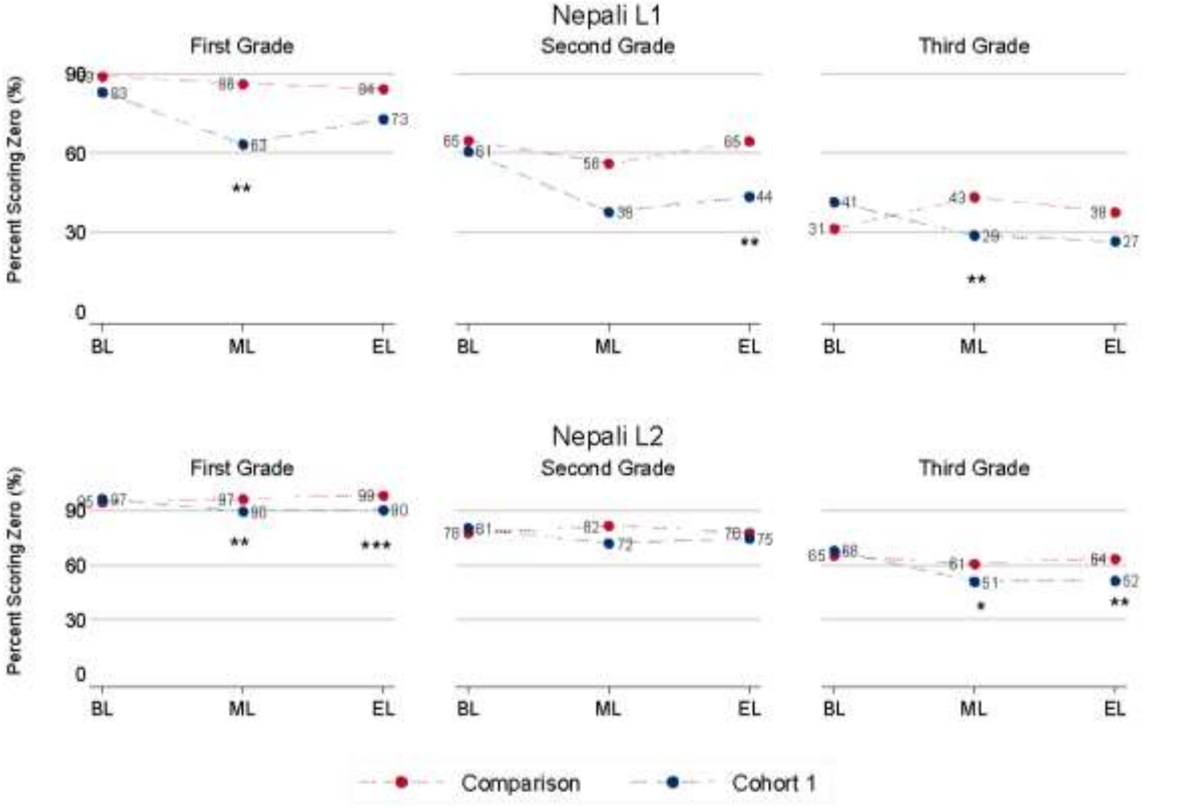


Note: Propensity score matching weights applied. DID compared to baseline *** p<0.01, ** p<0.05, * p<0.1

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 poor reading ability of students in the targeted schools. Among the L2 learners the percentages of zero scores (non-readers) are substantially higher. Even in third grade more than half of the L2 learners assessed were not able to read one word at baseline.

Between baseline and midline, the percentage of non-readers among L1 and L2 learners remained essentially unchanged in the comparison group, while in the intervention schools each grade level saw a reduction in zero scores. At endline, the percentage of zero scores among treatment L1 learners was slightly smaller than at midline. The impact of the NEGRP at endline for cohort 1 L1 learners is a reduction of zero scores of approximately13 percentage points in grade 2 and 16 percentage points in grade 3 (effects sizes of 0.24 and 0.42 std. dev., respectively). In the case of cohort 1 L2 learners in the treatment group, the percentage of students with zero scores continued on a downward trend from midline, but this was accompanied by a reduction of zero scores for L2 students in comparison schools, as well. The reduction of zero scores at endline was around 11 percentage points for grade 1 (effect size 0.29)[7]. There was no statistically significant impact of the program on reduction of zero scores in cohort 1 L2 students in grades 2 and 3. The stars in the figure indicate when the NEGRP effect at midline and at endline is statistically significant compared to baseline.

Figure 5: Percentage of zero scores in Reading Comprehension, Cohort 1, by grade and learner language



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

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 five percentage points (effect size is 0.20 of a std. dev.) in first grade, 17 percentage points (0.28 std. dev.) in second grade and 21 percentage points (0.49 std. dev.) in third grade. The stars in the figure indicate where the NEGRP effect is statistically significant. The effect size of the program is also sizeable for L2 learners, 11 percentage points for grade 1, six percentage points for grade 2, and 15 percentage points

[7] 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.

for grade 3 (effect sizes of 0.53, 0.16 and 0.26 std. dev. respectively). It is important to note that the proportion of L2 learners with zero comprehension scores is still very high at endline.

In Tables A4.9 and A4.10 in Annex IV, we show the complete distribution of cwpm achieved by L1 and L2 learners in cohort 1 and its comparison group, at baseline and endline. In general, we found an improvement in reading fluency in Cohort 1 at all reading levels due to NEGRP activities.

Oral Reading Fluency Benchmarks. The standards adopted by the GoN to assess progress in overall reading proficiency -defined as 'reading fluently with comprehension'- are 45 correctly identified words per minute from a connected text and 80 percent of the reading comprehension questions correctly answered.

In Figures A4.1 and A4.2 of Annex IV, we show the percentage of learners in treatment and comparison groups able to read 45 correctly identified words per minute and the percentage of learners able to answer 80% of the reading comprehension questions correctly, respectively. L1 learners reach the thresholds in much higher proportions than learners that have a different mother tongue than Nepali, particularly in grades 2 and 3. Not surprisingly, grade 3 learners do better than those in grade 2. However, L2 learners in grade 3 perform worse than L1 learners in 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 NEGRP, particularly among L1 students, was clearly evident. The change in the proportion of learners reaching the thresholds at midline was significantly larger for treated than for comparison learners (15 vs. 5 percent for grade2 and 28 vs. 5 percent for grade 3). At endline, L1 students in comparison schools were on more of an upward trend. A statistically significant effect for NEGRP is detected for grade 3 learners on the fluency benchmark but not on the comprehension benchmark. There was only limited evidence of a positive effect among L2 learners for either the fluency or comprehension benchmarks at midline, and no statistically significant findings for this group at endline at any grade level; the data once more suggests that additional and focused attention is required for L2 learners.

We calculated the percentage of learners able to reach the benchmark which includes both 45 cwpm and 80 percent of the reading comprehension questions answered correctly. Table 8 shows percentage estimates for grades 2 and 3 by the language of the learners at baseline, midline, and endline.

Table 8: Percentage able to read 45 cwpm and respond to 80 percent of reading comprehension questions correctly, cohort 1 and comparison, by grade and learner language.

	Grade 2			Grade 3		
	Baseline	Midline	Endline	Baseline	Midline	Endline
Cohort 1 (all)	1.6%	3.4%	3.7%	5.1%	9.6%	10.7%
L1 Students	5.2%	9.6%	9.1%	11.1%	24.1%	17.4%
L2 Students	0.0%	0.6%	0.9%	2.0%	2.7%	7.4%
Comparison (all)	0.8%	0.2%	1.0%	5.3%	3.2%	4.3%
L1 Students	2.1%	0.3%	1.4%	15.2%	6.2%	8.2%

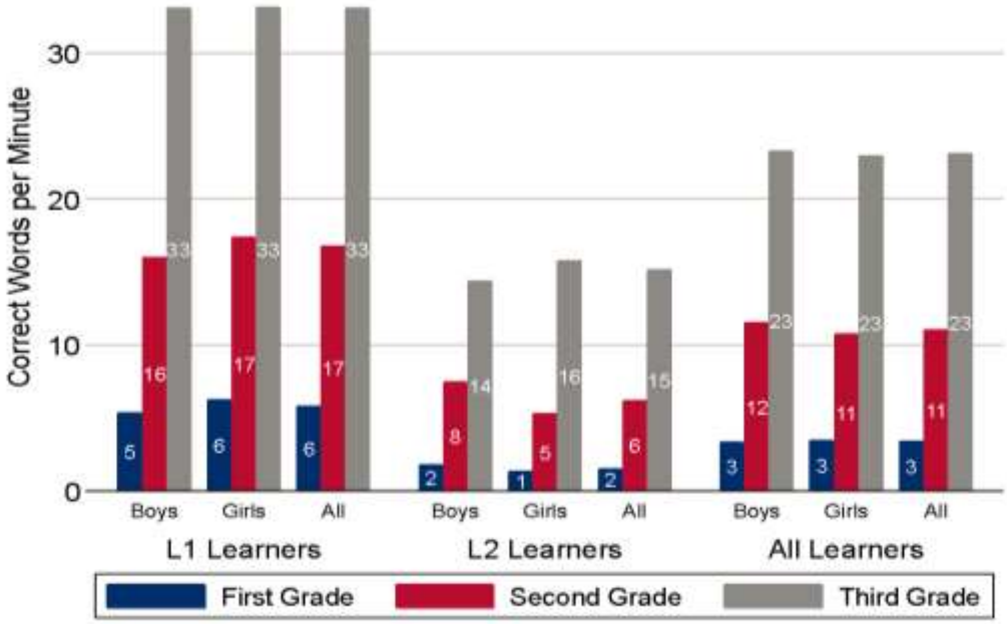
	Grade 2			Grade 3		
	Baseline	Midline	Endline	Baseline	Midline	Endline
L2 Students	0.5%	0.2%	0.8%	1.8%	2.1%	2.5%

Note: Sample weights applied to recover population representativeness. No matching applied.

The percentages show substantial progress between baseline and midline, and baseline and endline. L1 students made significant gains between baseline and midline, but gave up part of those gains at endline; in both grades 2 and 3, the percentage of students reaching both the fluency and comprehension benchmarks approximately doubled between baseline and midline, before falling slightly from the midline numbers at endline. In contrast, L2 learners are far from the levels of their L1 peers, but show progress with each round of testing. The percentage of grade 3 L2 learners in cohort 1 reaching the reading and comprehension benchmark has more than tripled between baseline and endline. Between baseline and midline, the percentage of grade 3 L2 learners reaching the benchmark was far below the percentage of grade 2 L1 learners reaching it, suggesting a more than one grade level difference between L1 and L2 learners. However, by the endline, this gap had narrowed somewhat.

Differential Impact by Gender: In Figure 6 we show the average correct words per minute at endline for the cohort 1 treatment group for girls and boys, and separately for L1 and L2 learners. Girls seem to perform slightly better than boys but none of the differences in means are statistically significant.

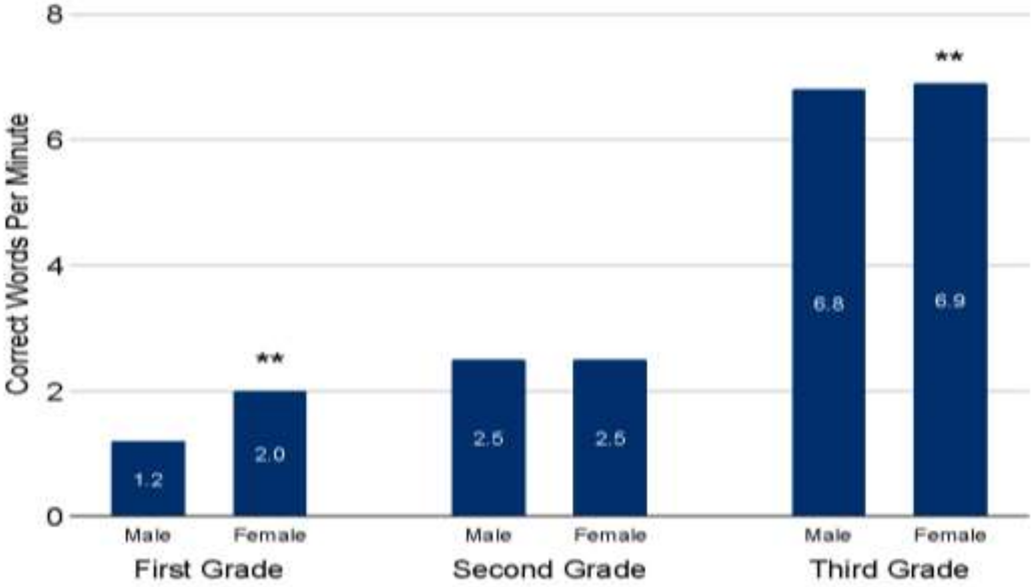
Figure 6: Mean Oral Reading Scores at Endline, cohort 1 Treatment, by Gender and Grade



NEGRP does not show differential effects for girls and boys. Figure 7 shows the NEGRP effect on oral reading by gender. All the estimated effects of the NEGRP on 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 Table A4.1 of Annex IV, we show more details. We present the NEGRP effects by gender for all the EGRA subtasks and the estimated differences in effects between boys and girls, which are not significant.

Figure 7: NEGRP Effect on Oral Reading Scores at Endline, cohort 1, 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 A4.3 and A4.4 of Annex IV. We find that both girls and boys have benefited from the NEGRP 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 slightly higher percentage of girls than boys reaching the national benchmarks, but this difference was statistically insignificant. Table 9 shows the percentage of girls and boys in grades 2 and 3 reaching both the fluency and comprehension thresholds, by Nepali speaking status. At endline, 7.5% of L1 boys and 10.4% of L1 girls in grade 2 reach both benchmarks, compared to 1.2% and 0.8% of L2 boys and girls, respectively. In grade 3, 17.6% of L1 boys and 17.3% of L1 girls reached both benchmarks at endline, compared to 5.1% and 8.7% for L2 boys and girls. The percentage of girls and boys reaching both benchmarks at endline does not consistently favor one gender or the other, and the progress between baseline and endline again suggests that the NEGRP benefited boys and girls approximately equally.

Table 9: Percentage able to read 45 cwpm and respond to 80 percent of reading comprehension questions correctly, cohort 1, by sex and learner language

	Grade 2			Grade 3		
BOYS	**Baseline**	**Midline**	**Endline**	**Baseline**	**Midline**	**Endline**
Cohort 1 (all)	1.3%	3.2%	3.5%	6.5%	7.4%	9.9%
L1 Students	3.8%	7.8%	7.5%	11.7%	17.1%	17.6%
L2 Students	0.0%	0.6%	1.2%	3.1%	3.1%	5.1%
Comparison (all)	0.7%	0.5%	0.9%	3.2%	3.4%	5.8%
L1 Students	0.8%	0.4%	1.7%	9.5%	5.2%	10.9%

	Grade 2			Grade 3		
L2 Students	0.6%	0.5%	0.4%	1.1%	2.9%	3.1%
GIRLS	**Baseline**	**Midline**	**Endline**	**Baseline**	**Midline**	**Endline**
Cohort 1 (all)	1.9%	3.6%	3.9%	4.1%	11.0%	11.2%
L1 Students	6.5%	11.2%	10.5%	10.6%	28.9%	17.3%
L2 Students	0.0%	0.6%	0.8%	1.3%	2.5%	8.7%
Comparison (all)	0.9%	0.1%	1.1%	6.7%	3.0%	3.4%
L1 Students	3.4%	0.3%	1.3%	18.8%	7.0%	6.2%
L2 Students	0.3%	0.0%	1.0%	2.3%	1.6%	2.1%

Note: Sample weights applied to recover population representativeness. No matching applied.

3.2.2. COHORT 2

Table 10 shows that for all EGRA subtasks and across grades the effect of NEGRP at endline is positive and, in most cases, statistically significant for L1 learners in cohort 2 districts. There are some subtasks for which the effect, although positive, is not statistically significant at conventional levels. In general, the lack of statistical significance seems to be the result of a sample underpowered to detect effects of that size[8].

These effects contrast with the midline findings where there was little evidence that the NEGRP had had any effect on learners' reading outcomes, because most aspects of the program had not been implemented at that point. At endline, after two academic years of full implementation of the NEGRP activities, the effects seen for L1 learners in cohort 2 are quite large, even if reading levels remain low, similar to what we found for cohort 1. For example, for L1 third graders in cohort 2, oral reading fluency is on average 32.9 and 22.6 cwpm for treatment and comparison groups, respectively. The NEGRP activities caused an increase of 12 words per minute after adjustments, an effect of 0.49 of the standard deviation.

Table 10: Effect of NEGRP on EGRA Subtasks, cohort 2, Nepali (L1) Learners, by Grade

	Baseline			Endline			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	12.8	19.8	7	5.2	7.8**	0.49
Grade 2	27.2	25.3	-1.9	21.9	32.5	10.6	12.5***	12.6***	0.63
Grade 3	37.7	38.2	0.5	32.2	40.1	7.9	7.4	8.0	0.38
Correct Matra Per Minute [Max=100]									
Grade 1	4.7	5.5	0.8	4.3	9.8	5.5	4.7***	6.1***	0.45
Grade 2	15.9	15	-0.9	14.2	24.1	9.9	10.8***	11.1***	0.51
Grade 3	25.9	27.9	2.0	24.9	32.7	7.8	5.8**	6.8*	0.29

[8] The impact is always statistically significant when analyzing all grades together.

	Baseline			Endline			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 Invented Words Per Minute [Max= 50]									
Grade 1	1.1	1.2	0.1	0.8	2.9	2.1	2.0***	2.4***	0.49
Grade 2	4.4	4.3	-0.1	3.7	7.5	3.8	3.9***	3.9***	0.47
Grade 3	7.8	9.3	1.5	7.6	10.4	2.8	1.3	1.8	0.19
Oral Reading Fluency									
Grade 1	2.3	2.4	0.1	2.4	6.5	4.1	4.0***	5.0***	0.46
Grade 2	12.5	10.7	-1.8	12.1	21.4	9.3	11.1***	11.1***	0.51
Grade 3	24	23.6	-0.4	22.6	32.9	10.3	10.7***	12.0***	0.49
Reading Comprehension, Percentage Correct									
Grade 1	4.1	3.8	-0.3	4.3	10.6	6.3	6.6***	8.5***	0.49
Grade 2	21.7	18.2	-3.5	17.6	31.6	14	17.5***	17.8***	0.64
Grade 3	43.4	38.3	-5.1	37.7	46.7	9	14.1**	15.9**	0.5
Listening Comprehension, Percentage Correct									
Grade 1	16	14	-2	11.9	16.1	4.2	6.2	8.7	0.38
Grade 2	29.7	22	-7.7	24.3	27.2	2.9	10.6**	11.9***	0.41
Grade 3	36.6	37.3	0.7	29.5	39.2	9.7	9.0*	8.7	0.27

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.

Similar to what we found for cohort 1, the effects among L2 learners, are also positive although smaller than for L1 learners and sometimes not statistically significant. In these districts, the performance of L2 learners is also lower than that of L1 learners. Table 11 shows the results. For example, for the treatment group, grade 3 L2 learners read on average 19.3 cwpm, while L1 learners read on average 32.9, an advantage of 13.6 cwpm. NEGRP has made a positive difference for L2 learners in increasing their reading skills. For example, the treated grade 3 L2 learners read on average 19.3 cwpm while the comparison group only reads 12.1 cwpm. This is an adjusted difference of 8.5 cwpm and an effect size of 0.42 of a standard deviation. L2 learners that received NEGRP intervention activities perform around the same levels of L1 learners that were not exposed to the program, suggesting that without the program they would be at a greater disadvantage than they are at endline.

Despite the original intention of including additional inventions specifically targeting Nepali L2 learners in cohort 2, this was not implemented and, therefore, we did not detect outsized impacts on non-native Nepali-speaking learners.

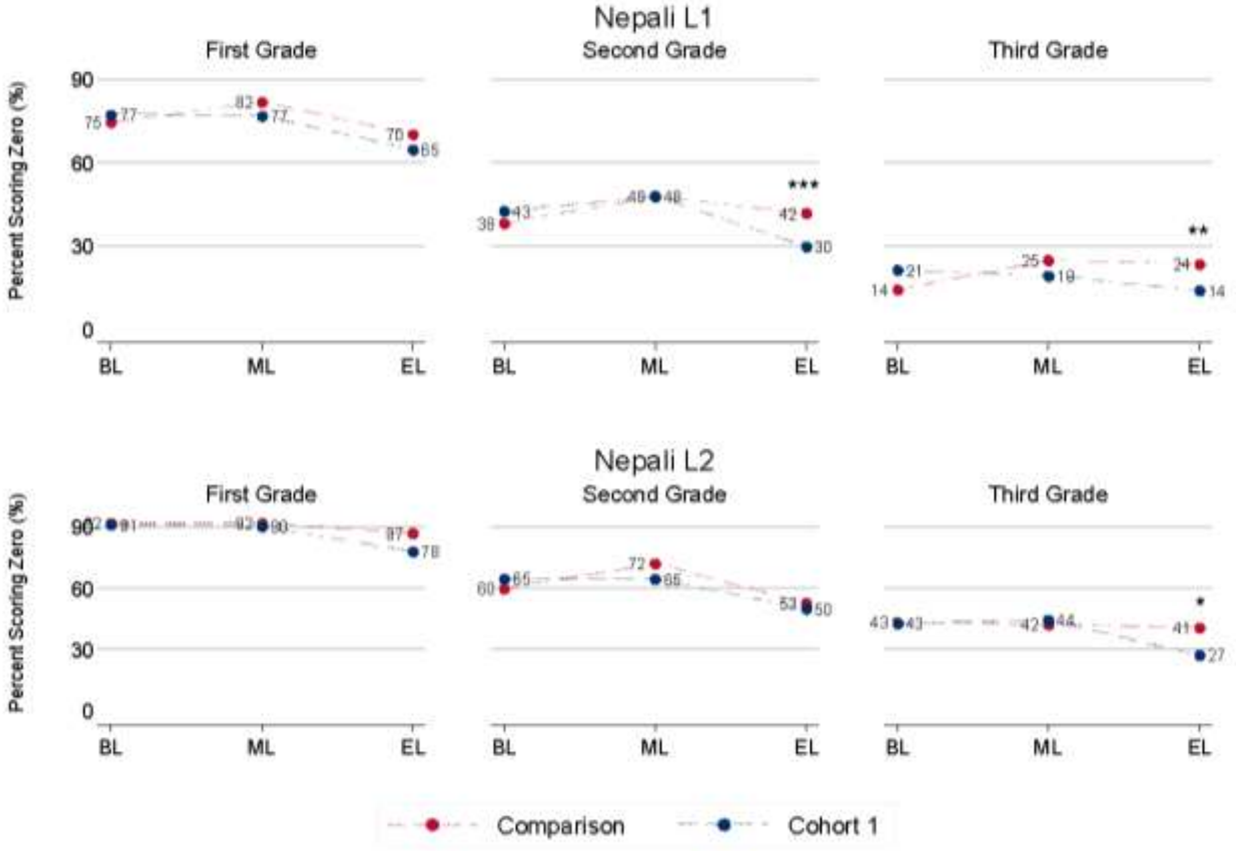
Table 11: Effect of NEGRP on EGRA Subtasks, cohort 2, Non-Nepali (L2) Learners, by Grade

	Baseline			Endline			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	7.3	11.6	4.3	4.6***	5.1***	0.43
Grade 2	19.9	17.5	-2.4	20.3	21.4	1.1	3.5	3.5	0.21
Grade 3	28.5	25.9	-2.6	23.2	30.7	7.5	10.1***	10.0***	0.52
Correct Matra Per Minute [Max=100]									
Grade 1	2	1.8	-0.2	1.5	4.4	2.9	3.1***	3.5***	0.43
Grade 2	8.1	7.9	-0.2	10.3	12	1.7	1.9	2.0	0.13
Grade 3	18.4	17.4	-1	15.7	22	6.3	7.3**	7.1**	0.35
Correct Invented Words Per Minute [Max= 50]									
Grade 1	0.4	0.4	0	0.1	1.2	1.1	1.1***	1.3***	0.46
Grade 2	1.8	2	0.2	2.5	3.1	0.6	0.4	0.5	0.09
Grade 3	5.2	5.5	0.3	4.4	6.7	2.3	2.0	1.9	0.24
Oral Reading Fluency [Max=100]									
Grade 1	0.9	0.9	0	0.3	2.5	2.2	2.2***	2.6***	0.51
Grade 2	5	5	0	6.1	8.3	2.2	2.2	2.4	0.21
Grade 3	14.1	12.9	-1.2	12.1	19.3	7.2	8.4***	8.5***	0.42
Reading Comprehension, Percentage Correct									
Grade 1	1.1	1.3	0.2	0.2	3.9	3.7	3.5***	3.9***	0.46
Grade 2	8.4	7.7	-0.7	8.9	12.7	3.8	4.5*	4.6*	0.24
Grade 3	20	20	0	17.5	28.1	10.6	10.6***	10.4**	0.38
Listening Comprehension, Percentage Correct									
Grade 1	5.1	5.4	0.3	7.2	11	3.8	3.5	3.9	0.21
Grade 2	13.7	12.3	-1.4	17.1	21.4	4.3	5.7	5.7	0.20
Grade 3	19.7	20	0.3	31.4	28.9	-2.5	-2.8	-2.7	-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 addition, we estimated the effect of NEGRP on the percentage of students not able to read a single word from the oral reading passage. Figure 8 shows the percentage of L1 and L2 learners with zero scores in the oral reading fluency subtask by grade. Among the L1 learners, the improvement is statistically significant for grades 2 and 3 while for L2 learners we only detect a statistically significant effect in grade 3. The fraction of none readers remains high despite some improvements, particularly among L2 learners. Half of L2 learners cannot read a single word by the end of grade 2 and more than a quarter still are unable to read at the end of grade 3.

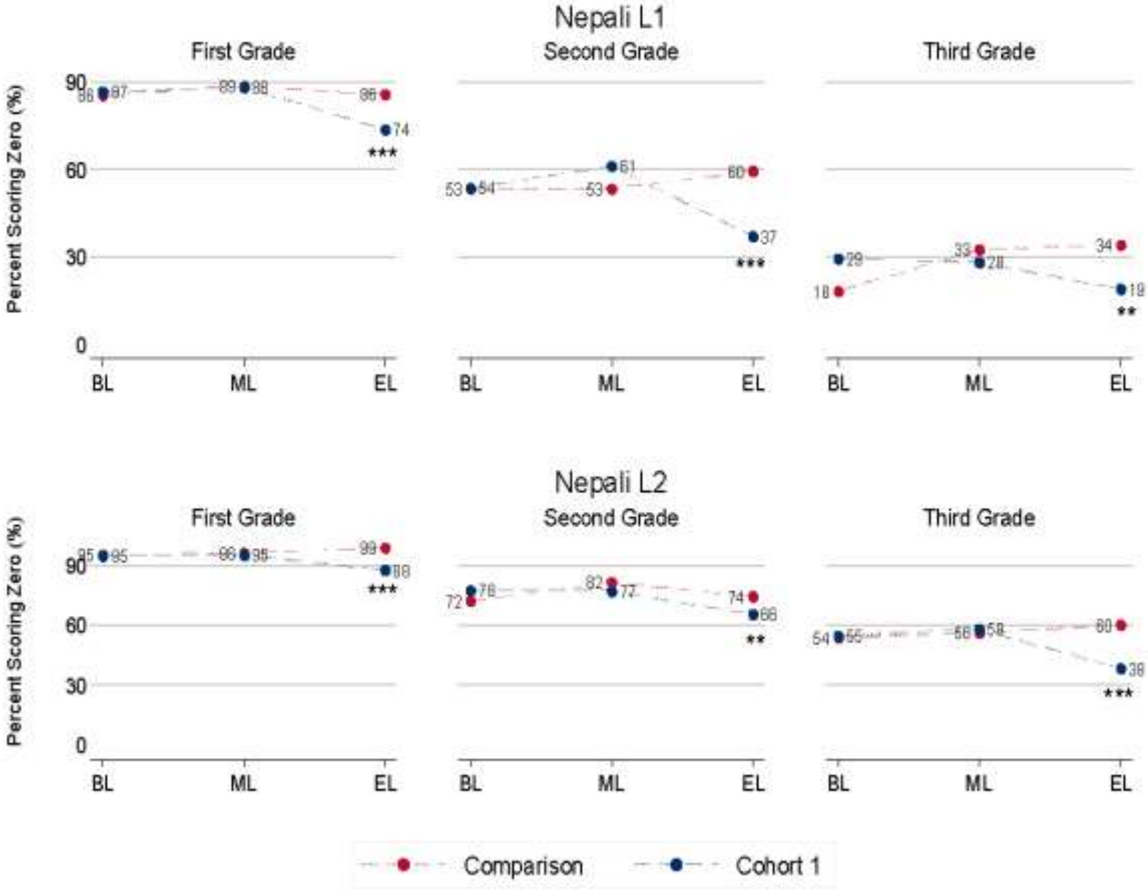
Figure 8: Percentage of zero score in Oral Reading Fluency, cohort 2, by grade and learner language



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

Learners in cohort 2 also show some improvement due to NEGRP in reading comprehension. Figure 9 shows the percentage of learners unable to correctly answer questions about the reading passage at baseline, midline and endline. Although the fraction of learners with zero scores is still high, by endline, we find that there was a statistically significant reduction of zero scores among those receiving NEGRP activities across all grade levels for both L1 and L2 learners.

Figure 9: Percentage of zero scores in Reading Comprehension, Cohort 2, by grade and learner language



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

In Tables A4.11 and A4.12 in Annex IV, we show the complete distribution of cwpm achieved by L1 and L2 learners in Cohort 2 and its comparison group at baseline and endline. In general, we found an improvement in reading fluency in Cohort 2 at all reading levels due to NEGRP activities.

Oral Reading Fluency Benchmarks. As we mentioned, the reading benchmark adopted by the GoN is 45 correctly identified words per minute from a connected text and 80 percent of the reading comprehension questions correctly answered. In Figures A4.5 and A4.6 of Annex IV, 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. As was seen in cohort 1, L1 learners reach the benchmarks in much higher proportions than learners that have a different mother tongue than Nepali, particularly in grades 2 and 3, and 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, when NEGRP had not yet been implemented at high intensity, numbers are practically unchanged from baseline for both treatment and comparison groups. At endline, L1 and L2 students in treatment schools showed substantial improvement, although so did students in comparison schools. Although the improvement in treatment schools appears to outpace those in comparison schools, the effect of NEGRP is only statistically significant for L2 grade 2 learners for the fluency benchmark, and for L1 grade 2 learners for the comprehension benchmark.

We calculated the percentage of learners able to reach the benchmark which includes both 45 cwpm and 80 percent of the reading comprehension questions answered correctly. Table 12 shows percentage estimates for grades 2 and 3 by the language of the learners at baseline, midline, and endline. The percentage of learners meeting the benchmark remained essentially unchanged between baseline and midline, when the cohort had only received a very light version of the NEGRP. By endline, the percentages have increased. Similar to cohort 1, the percentage of L2 learners reaching the benchmark is substantially lower than that of L1 learners. At endline, the percentage of L1 learners in grade 2 reaching the benchmark is similar to that of L2 learners in grade 3: 4.9 vs. 4.3 percent, respectively.

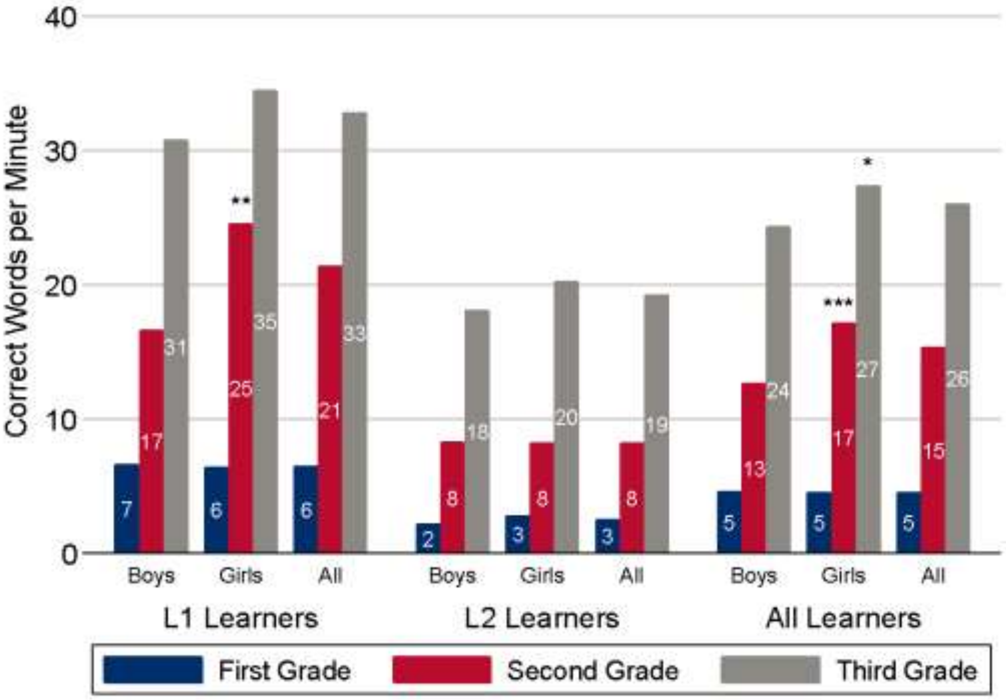
Table 12: Percentage able to read 45 cwpm and respond to 80 percent of reading comprehension questions correctly, cohort 2 and comparison, by grade and learner language

	Grade 2			Grade 3		
	Baseline	Midline	Endline	Baseline	Midline	Endline
Cohort 2 (all)	0.5%	1.2%	3.0%	6.3%	6.6%	8.6%
L1 Students	1.0%	2.5%	4.9%	10.7%	11.8%	13.9%
L2 Students	0.1%	0.0%	1.1%	2.9%	2.2%	4.3%
Comparison (all)	0.8%	0.2%	1.0%	5.3%	3.2%	4.3%
L1 Students	2.1%	0.3%	1.4%	15.2%	6.2%	8.2%
L2 Students	0.5%	0.2%	0.8%	1.8%	2.1%	2.5%

Note: Sample weights applied to recover population representativeness. No matching applied

Differential Impact by Gender: In Figure 10 we show the average correct words per minute at endline for the cohort 2 treatment group for girls and boys, separately for L1 and L2 learners. Similar to what was seen in cohort 1, girls seem to perform slightly better than boys, although only the differences in means is statistically significant only for L1 second graders.

Figure 10: Mean Oral Reading Scores at Endline, cohort 2 Treatment, by Gender and Grade

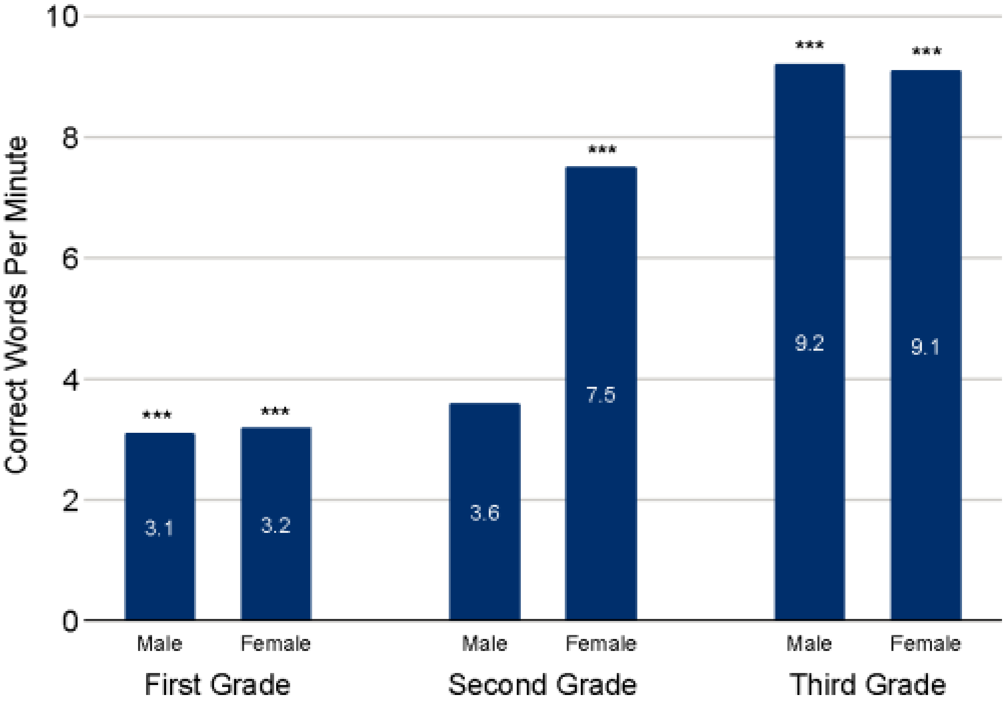


Note: Propensity score matching weights applied. Test for difference in means across genders in the same grade and language groups *** p<0.01, ** p<0.05, * p<0.1

NEGRP does not show differential effects for girls and boys. Figure 11 shows the NEGRP effect on oral reading by gender. Estimated effects of the NEGRP on girls are positive and significant (significance levels shown by stars) across all three grade levels, and positive and significant for boys in 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 Table A4.2 of Annex IV, we show more details. We present the NEGRP effects by gender for all the EGRA subtasks and the estimated differences in effects between boys and girls, which are not significant. Additionally, Figures A4.7 and A4.8 show the impact on reaching the fluency and comprehension benchmarks, respectively, by grade and gender.

Figure 11: NEGRP Effect on Oral Reading Scores at Endline, cohort 2, by Gender and Grade



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

Table 13 shows the percentage of cohort 2 girls and boys in grades 2 and 3 reaching both the fluency and comprehension thresholds, by Nepali speaking status. At endline, 4.0% of L1 boys and 5.7% of L1 girls in grade 2 reach both benchmarks, compared to 1.6% and 0.7% of L2 boys and girls, respectively. In grade 3, 12.7% of L1 boys and 15.0% of L1 girls reached both benchmarks at endline, compared to 5.3% and 3.4% for L2 boys and girls. The percentage of girls and boys reaching both benchmarks at endline tends to slightly favor girls in the L1 learner group, and to slightly favor boys in the L2 group. However, in terms of progress between baseline and endline, the results again suggest that the NEGRP benefited boys and girls approximately equally, regardless of grade or Nepali speaking status.

Table 13: Percentage able to read 45 cwpm and respond to 80 percent of reading comprehension questions correctly, cohort 2, by sex and learner language

	Grade 2			Grade 3		
BOYS	**Baseline**	**Midline**	**Endline**	**Baseline**	**Midline**	**Endline**
Cohort 2 (all)	0.7%	0.5%	2.7%	4.5%	4.6%	8.5%
L1 Students	1.3%	1.0%	4.0%	6.9%	8.3%	12.7%
L2 Students	0.3%	0.0%	1.6%	2.7%	1.6%	5.3%
Comparison (all)	0.7%	0.5%	0.9%	3.2%	3.4%	5.8%
L1 Students	0.8%	0.4%	1.7%	9.5%	5.2%	10.9%
L2 Students	0.6%	0.5%	0.4%	1.1%	2.9%	3.1%
GIRLS	**Baseline**	**Midline**	**Endline**	**Baseline**	**Midline**	**Endline**

	Grade 2			Grade 3		
BOYS	**Baseline**	**Midline**	**Endline**	**Baseline**	**Midline**	**Endline**
Cohort 2 (all)	0.4%	1.8%	3.2%	7.8%	8.2%	8.7%
L1 Students	0.8%	4.0%	5.7%	14.2%	14.5%	15.0%
L2 Students	0.0%	0.0%	0.7%	3.1%	2.7%	3.4%
Comparison (all)	0.9%	0.1%	1.1%	6.7%	3.0%	3.4%
L1 Students	3.4%	0.3%	1.3%	18.8%	7.0%	6.2%
L2 Students	0.3%	0.0%	1.0%	2.3%	1.6%	2.1%

Note: Sample weights applied to recover population representativeness. No matching applied

In the next section we analyze the mechanisms behind the positive impact of NEGRP among cohort 1 and 2 learners.

3.3. MECHANISMS BEHIND THE IMPACT OF NEGRP

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

The NEGRP 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 learners' 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 matched comparison schools. The sample for these categories is much smaller than the learner sample. We only have one observation per school per year in each of these categories (see Table 3 for exact sample sizes for each category by year and cohort).

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 NEGRP 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

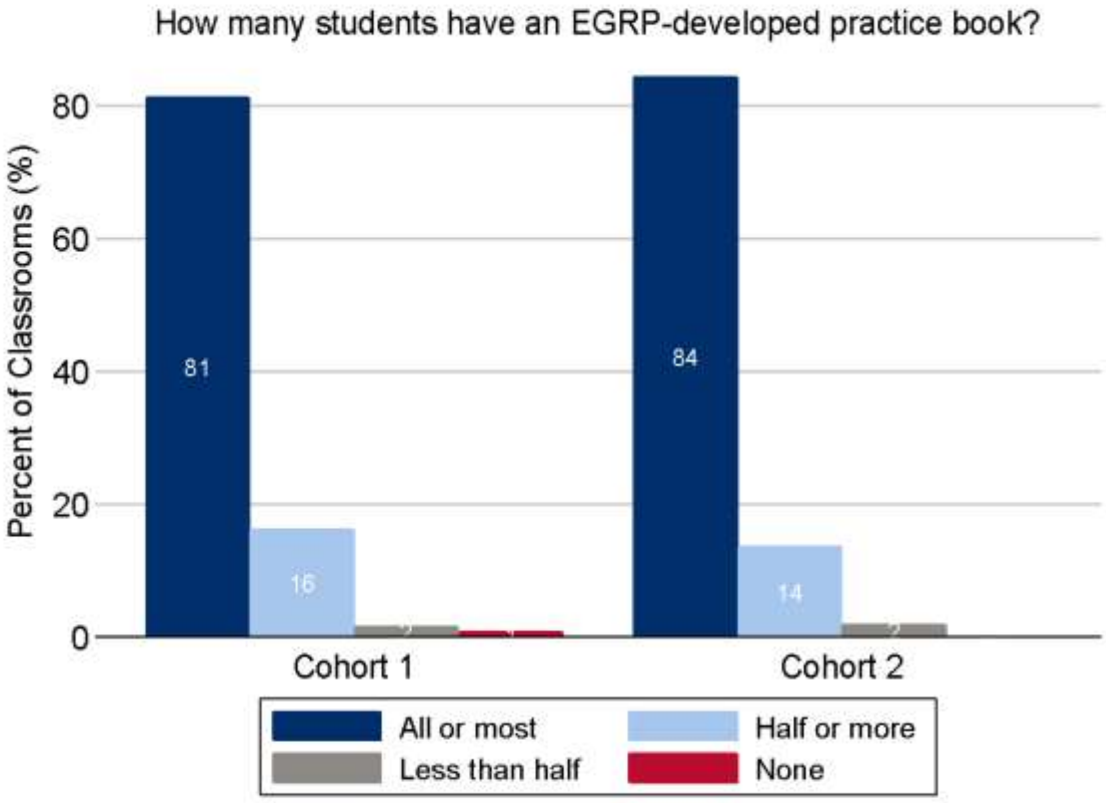
Through classroom inventories, enumerators recorded whether learners had NEGRP Nepali reading materials, such as the NEGRP developed workbook or practice book. The fraction of learners that had workbooks were recorded in five categories: none, very few, less than half, half or just over half, all or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

Figure 12 shows that in slightly **over 80 percent of the classrooms in cohorts 1 and 2, all or almost all children had their own workbook**. In an additional 15 percent of the classrooms, more than half of learners had their own workbook but some had to share, and in a small number of classrooms less than half the students had their own workbook.

In Figure 13 we show the proportion of classrooms where **supplemental reading materials** are available. In this case we see that all groups have increased the percentage of classrooms with additional reading resources; however, **cohort 1 and cohort 2 have seen a substantial and significantly higher increase**. By midline, availability of supplemental reading materials was widespread in cohort 1 and 2 schools, and by endline was nearly universal. Just four percent of the cohort 1 classes visited and seven percent in cohort 2 were lacking supplemental reading materials at endline.

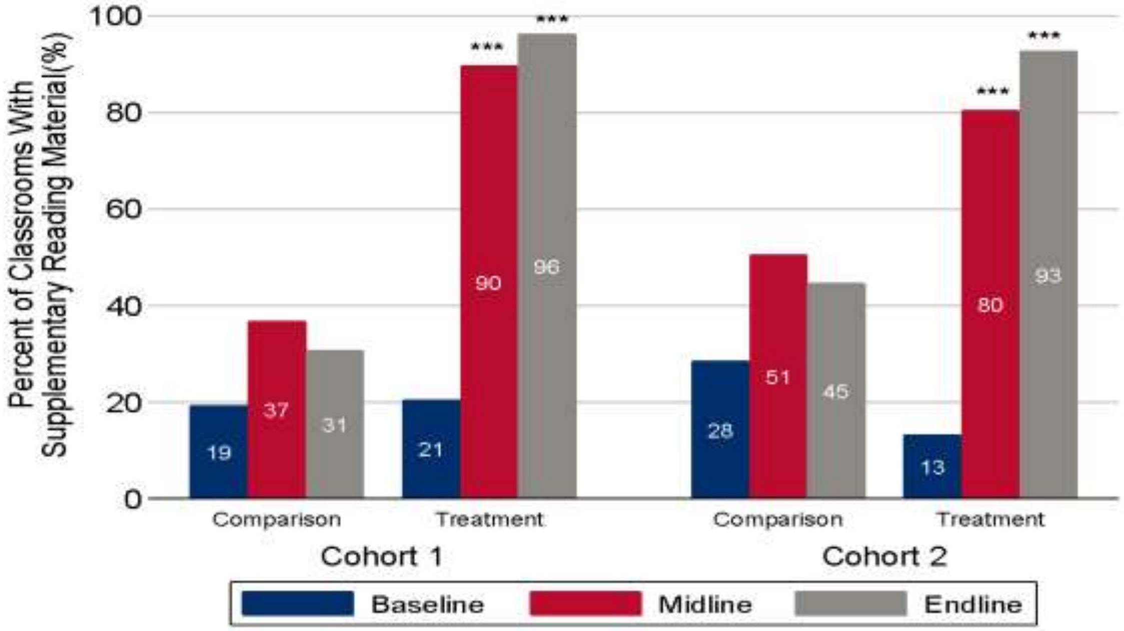
Table A4.3 in Annex IV provides additional details on the impact of NEGRP on the materials available to students in their classrooms.

Figure 12: Prevalence of NEGRP-developed practice books in Cohort 1 and 2 classrooms



Note: Sample weights applied to recover population representativeness.

Figure 13: Supplementary Reading Materials available in the Classroom



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

3.3.2. TEACHER TRAINING, SUPPORT, AND INSTRUCTIONAL PRACTICES

Teachers training is an important component of the NEGRP activity. NEGRP reported delivering reading training to 6,753 teachers in Cohort 2 during the year before endline data collection. However, the number of teachers trained during that period could not be identified based on teacher self-reporting at endline, due to a glitch in data collection instrument.

Table 14 shows in detail the number of teachers trained from 2016 to 2019 by NEGRP and the number of teachers that received refresher training.

Table 14: Teachers Trained through NEGRP

Year	Teachers	Cohort	Refresher Training
FY16	3,109	Cohort 1, 2 teachers per school	
FY17	1,568	Cohort 1, 1 teacher per school	
FY18	6,753	Cohort 2, 2 teachers per school	
FY19	2,881	Cohort 2, 1 teacher per school	4226
Total	14,311 Note: This is the number of unique teachers trained		4226 Note: These are already counted under the 14,311

The overall life-of-project training target is 14,780 unique teachers. ERGP missed the target by 469 educators. According to information we received from the program, not all planned training events could be completed in a few districts due to the release of G2G budgeted funds nearly at the end of the GoN's fiscal year (FY) and bottlenecks experienced by single provincial-level Education Training Centers

(ETCs) being required to roll out large-scale trainings at the same time. These obstacles resulted in the small (3 percent) underachievement of this target.

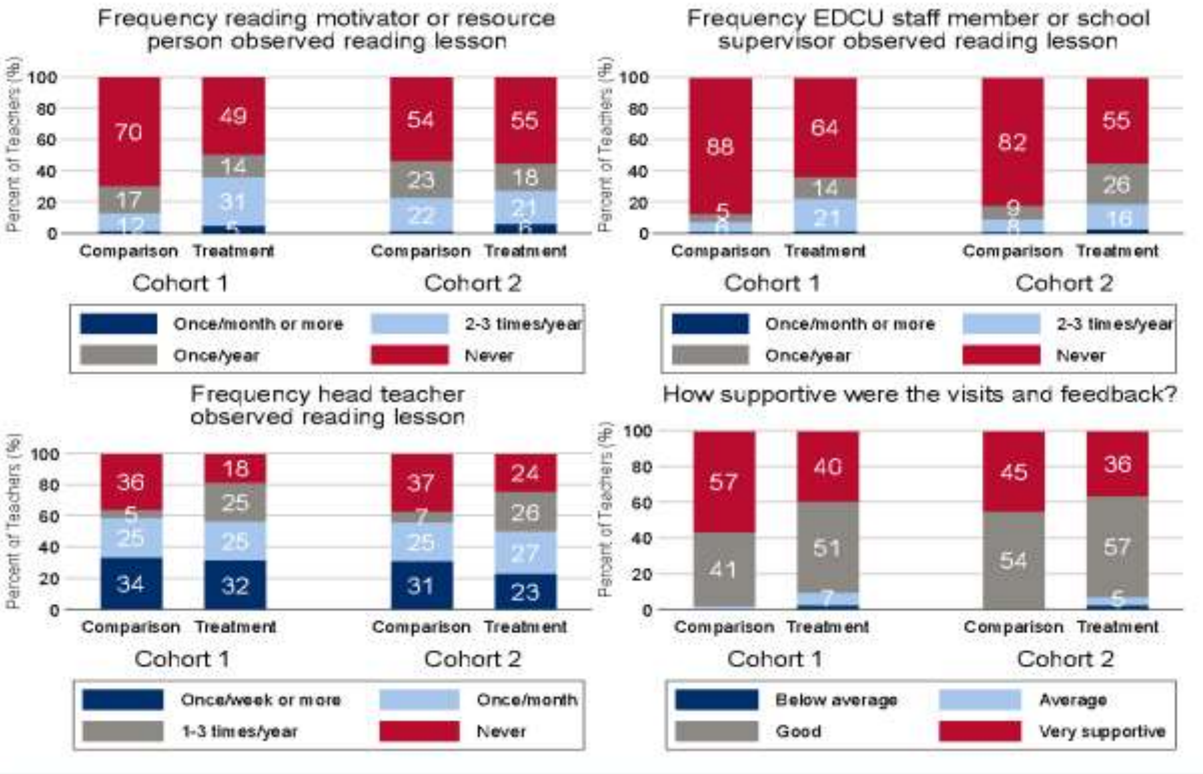
In addition, NEGRP had plans for refresher training for Cohort 2 teachers in the first half of 2020 however those activities were postponed because of a country lockdown due to COVID-19.

At endline, **teachers in cohorts 1 and 2 were more likely to receive support and supervision from different sources**. In Figure 14, we show that 36 percent of the treatment teachers in cohort 1 report that a motivator or resource person observed their reading class at least two to three times per year, compared to 13 percent of the matched comparison group for cohort 1. Among teachers in cohort 2, 27 percent reported observations from a motivator or resource person at least two to three times per year, compared to 23 percent of teachers in the matched comparison group for cohort 2. Still, a large proportion of the teachers in treatment schools in both cohorts report no supervision from a motivator or resource person at all.

Cohort 1 and 2 teachers also report more supervision from education officers than their peers in matched comparison groups, although more than half of them indicated that the district education officer never observed their reading classes. Cohort 1 and 2 teachers were also less likely to report the head teacher never observes their reading lessons, compared to the matched comparison groups.

Teachers who received support and feedback were generally happy with the quality of these visits. Over 90 percent of teachers who received visits said the visits were "very supportive" or "good".

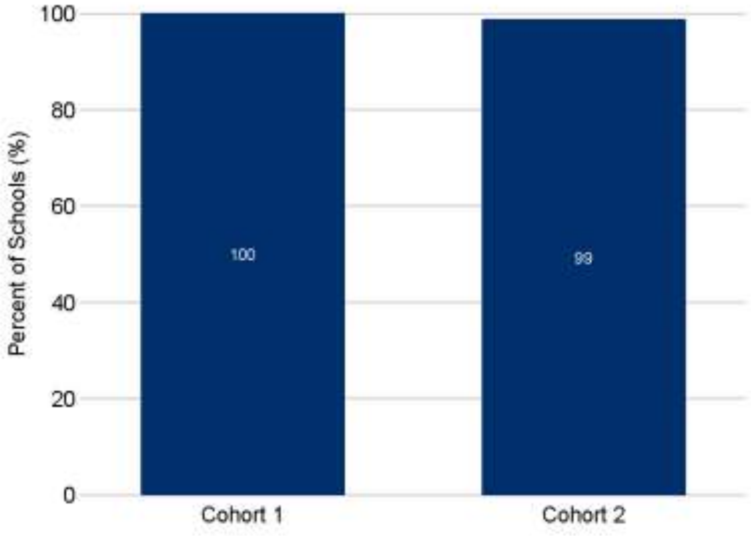
Figure 14: Support and Supervision of Teachers



Note: Propensity score matching weights applied.

Teachers were also asked if their schools have NEGRP developed materials. Figure 15 shows that every teacher in cohort 1 confirmed receiving the material, while 99 percent of the teachers in cohort 2 did so.

Figure 15: School has NEGRP Materials at Endline, as reported by Teachers



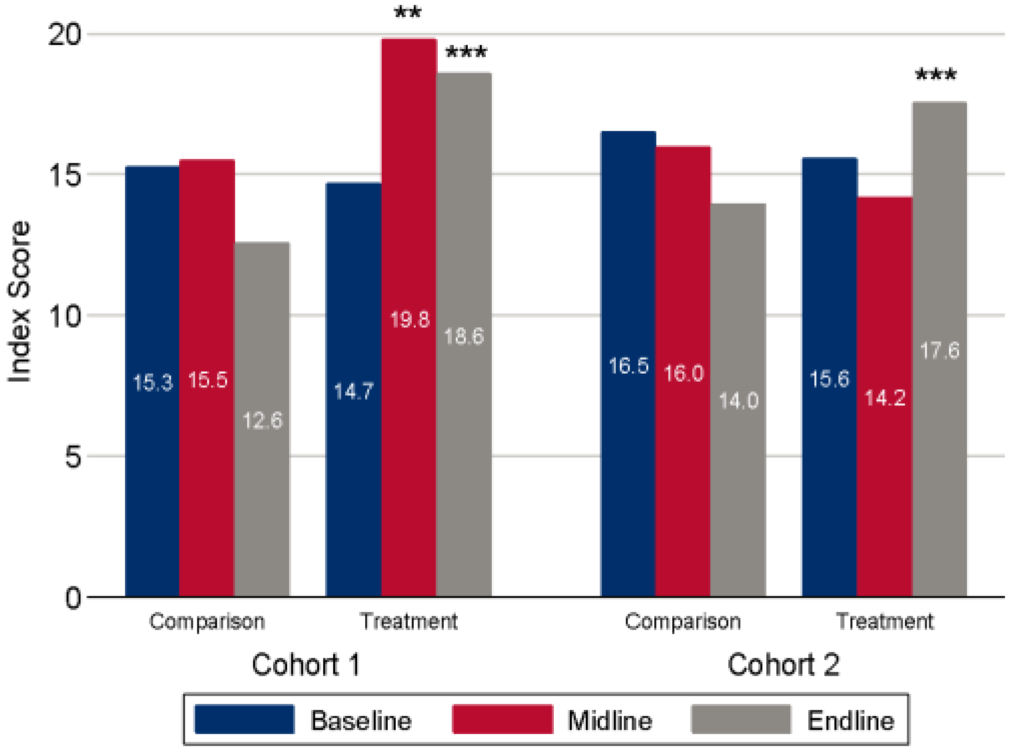
Note: Sample weights applied to recover population representativeness.

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 NEGRP Nepali L1 program changed teachers' reading instruction practices in the classroom?	The teaching reading instruction practice index shows a positive impact of the NEGRP for both cohorts.

We created two indexes to measure teachers' reading instruction practices in the classroom. The first index –Index I- includes 30 items describing desirable actions during an early grade reading lesson, for example, teaching and practicing letter sounds, reading independently, introducing vocabulary, using teaching and learning materials appropriately, etc. 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 complete list of items is included in Annex VI Construction of Indexes.

Figure 16: Teacher Reading Instruction Practices Index I



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

Figure 16 shows a positive effect for the NEGRP on the Teacher Reading Instruction Practices Index I for cohort 1 at midline and endline, and a positive effect for cohort 2 at endline. At endline, the DiD estimate shows a large and statistically significant effect of 6.6 points in the index for cohort 1 teachers, and a statistically significant effect of 4.5 points for cohort 2 teachers.

The evaluation team created an additional reading instruction index – the Teacher Reading Instruction Practices Index II- using calculation guidelines from USAID. This index includes a subset of questions used in Index I, but requires specific combinations of teaching practices that reflect categories such as phonemic awareness instruction, fluency modeling, reading comprehension exercises, etc. Index II takes values ranging from 0 to 13, giving one point for each of 13 practices. In Annex VI we present the index calculation details.

Table 15 shows the results for the mean Teacher Reading Instruction Practices Index II values for each cohort and comparison group at baseline, midline, and endline, as well as the percentage of teachers meeting USAID's effective teaching practices threshold (defined as scoring at least 9 out of 13 on the index). The table shows limited evidence that the NEGRP had increased teacher effectiveness by midline for teachers in cohort 1 schools, and shows strong evidence that it had done so by endline in both cohorts. Specifically, the NEGRP increased index scores by 2.9 points and increased the percentage of teachers meeting the effectiveness threshold by 38.7 percentage points in cohort 1 schools by endline, and increased index scores by 2.4 points and increased the percentage of teachers meeting the effectiveness threshold by 36.4 percentage points in cohort 2 schools by endline.

Table 15: Teacher Reading Instruction Practices Index II

	Index II			Percentage Meeting Index II Effective Practices Threshold of 9 points		
	Baseline	Midline	Endline	Baseline	Midline	Endline
Cohort 1 Treatment	7.9	10.0*	10.2***	45.9%	75.3%	81.8%**
Cohort 1 Comparison	7.9	8.3	7.3	47.3%	43.7%	44.5%
Cohort 2 Treatment	8.5	7.6	10.2***	55.7%	36.6%	81.0%***
Cohort 2 Comparison	8.6	8.5	7.8	59.9%	48.4%	48.9%

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

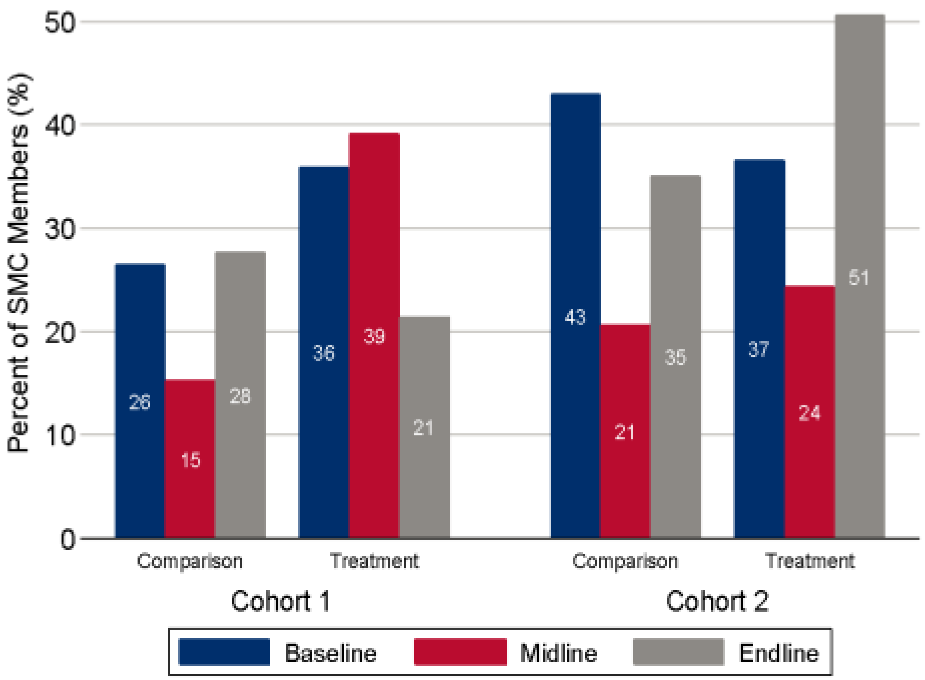
In addition, teachers self-reported their approach to supporting students with reading difficulties, their attitudes about how students learn to read and about how to teach early grade reading, and if they assign reading homework to learners. In general, we do not find differences in the support provided, attitudes, or the assignment of reading homework between the teachers in the cohort 1 and 2 treatment groups and their matched comparison groups. While some items show the desired trends, others do the opposite and few items appear statistically significant. However, the one consistent finding for both cohorts 1 and 2 is that teachers were significantly more likely to say they give daily reading assignments to complete outside of school, with the NEGRP showing an estimate increase of 58 percentage points for cohort 1 and 37 percentage points for cohort 2. The details of the impact of NEGRP on teachers are included in Tables A4.4 and A4.5 of Annex IV.

3.3.3. PARENTAL AND COMMUNITY ENGAGEMENT

Effect on School Leadership and Management

One of the components of the NEGRP 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 17: Training of School Management Committee Members



Note: Propensity score matching weights applied.

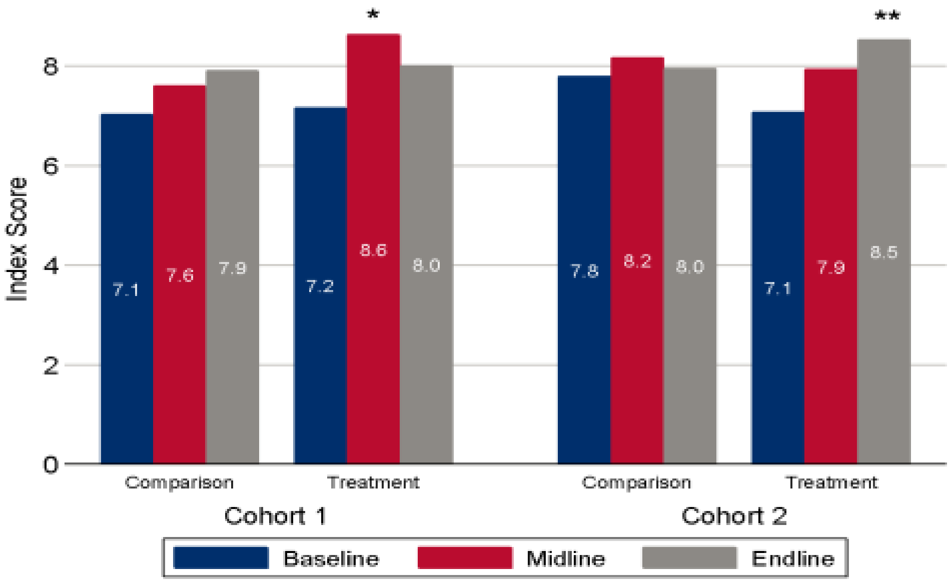
Figure 17 shows the percentage of SMC members that received training in the past two years. In each school, the question was administered to the chair of the SMC who is prioritized for training; in cases where the SMC chair was not available another SMC member was selected for the interview. The proportion of SMC members trained was higher in cohort 1 schools than in cohort 2 and comparison groups at midline, and higher for cohort 2 schools than in cohort 1 and comparison groups at endline. Since the question asks respondents to think about training in the last two years only, one possible explanation for this pattern is that training occurred early on in NEGRP cycle for cohort 1 schools, leading to higher reported training at midline, and then moved on to cohort 2 schools, resulting in the drop in the percentage for cohort 1 and the higher percentage for cohort 2 at endline. Nevertheless, none of these percentages is particularly high.

USAID/Nepal and the NEGRP team defined a School Leadership and Management Index. The index includes 14 items related to the school priorities, actions devoted to promote reading, parental involvement, student reading performance monitoring, etc. We provide the complete list of items in Annex VI Construction of Indexes. 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 Figure 18 the management index at baseline, midline, and endline for cohorts 1 and 2, along with their matched comparison groups. In cohort 1, the index was positively impacted by the program at midline. The impact at midline was 0.9 points and was statistically significant but was no longer present at endline. Cohort 2 shows a statistically significant effect for NEGRP on the index at endline of

1.2 points. Additional details about the impact of NEGRP on school management can be found in Table A4.6 of Annex IV Additional Analysis.

Figure 18: Management Index, Cohorts 1 and 2



Note: Propensity score matching weights applied. DID compared to baseline *** 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 NEGRP Nepali L1 program changed the school leadership and management index (as defined in monitoring index), demonstrating active support for EGR?	The NEGRP Nepali L1 program has generated a modest improvement of 1.2 points (out of 14) in the management index for cohort 2 However, at endline there is no impact in the index among schools in cohort 1.

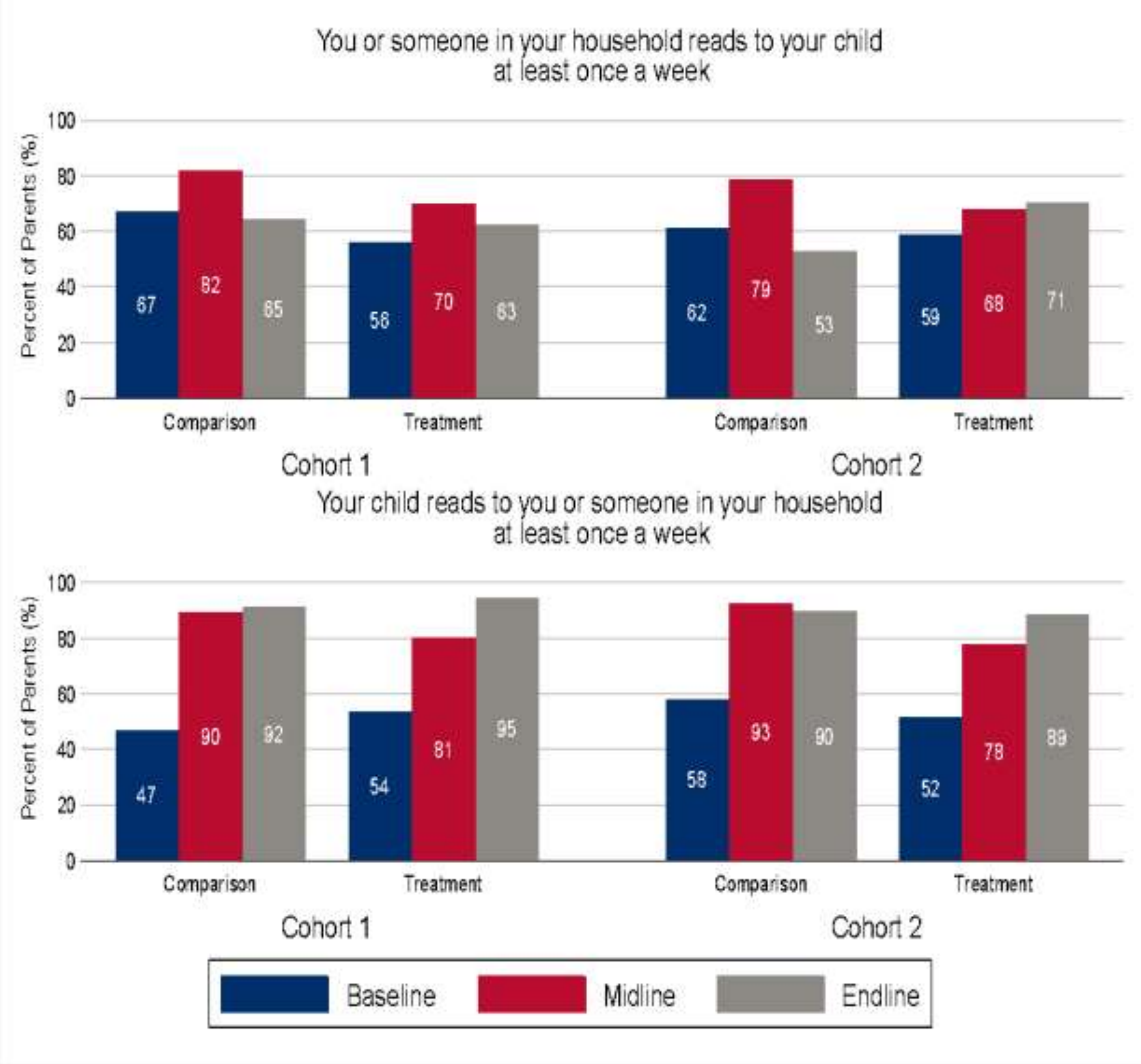
Effect on Parental Engagement and Awareness

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

Here, we explore the effects of the NEGRP 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 19 shows the averages at baseline and midline for cohorts 1 and 2 and their corresponding comparison groups. For all treatment groups, there is some increment between baseline and endline in the percentages of reading to the child and listening the child read at home at least once a week. The NEGRP, however, does not show an additional improvement over the comparison group on these indicators. It is possible that it is difficult to increase at home reading beyond the already achieved levels. For example, at endline in all the groups around 90 percent of children read to someone at home and around 60 percent of parents/caregivers read with the child at home at least once a week.

Table A4.7 in Annex IV shows additional details about the effect of the NEGRP on reading at home.

Figure 19: Reading at home, Cohorts 1 and 2



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

4. CONCLUSIONS

The NEGRP had positive effects at the endline among learners in cohorts 1 and 2. The effects of the program are similar in each cohort, giving us greater confidence in the findings. Additionally, the findings for cohort 2 at endline are similar to the effects found at midline for cohort 1, where the program had already been fully rolled out. In contrast, the lack of findings for cohort 2 at midline, where the program had not yet been fully rolled out, confirms a key assumption of the analytical methodology – that in absence of the NEGRP interventions, the treatment and comparison groups move in parallel along a similar trajectory – and combined with the impacts found for cohort 2 at endline, by which time implementation had been fully rolled out, we can confidently attribute causality to the NEGRP for the improvements in reading outcomes seen in the treatment groups.

Reading performance indicators have improved for treatment learners however, **there is still room for improvement**. At endline, most grade 1 learners are non-readers and by grade 3 around a quarter of them are still not able to read a single word from a connected paragraph. Oral reading fluency is still low for all grades and very few learners reach the reading benchmark of 45 cwpm and 80% reading comprehension. Table 16 below summarizes these indicators.

Table 16: Summary reading indicators at endline. Treatment learners by grade and cohort

% Non readers (zero cwpm)	Grade1	Grade 2	Grade 3
Cohort 1	69.8%	48.7%	27.0%
Cohort 2	74.9%	42.2%	22.0%
ORF (cwpm)			
Cohort 1	3.4	10.8	21.2
Cohort 2	3.5	13.7	24.4
% Reaching Benchmark			
Cohort 1	0.8%	3.7%	10.7%
Cohort 2	0.4%	3.0%	8.6%

Note: Sample weights applied to recover population representativeness

In cohorts 1 and 2, 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 benefits L1 learners more than L2 learners.** As noted at baseline and midline, 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. NEGRP was able to improve performance among L2 learners but not enough to reduce the disadvantage they experience as non-Nepali speakers. L2 learners not only lag behind L1 learners in terms of their reading skills, it is also clear that there is a deficiency in overall oral Nepali language comprehension among L2 learners. Similar to the findings at midline, the endline results point to the need for targeted and sustained interventions to support L2 learners beyond what was implemented through NEGRP technical support.

The NEGRP has benefited students with both low and high performance. An improvement in reading performance was found across groups of learners with different reading abilities. NEGRP reduced the number of zero scores among learners and also increased the percentage of learners that reach the benchmarks of 45 correct words per minute and 80 percent oral reading comprehension that the GoN has adopted.

Examining the channels through which the program has functioned, **this impact evaluation did not find evidence that the program has led to changes in parents' at-home support** for their children's reading development, which seems quite high for all groups. **SMC support for reading activities shows a very modest improvement for cohort 2 and no improvement for cohort 1.**

There is evidence that the program has had a **positive effect on teachers' reading instruction**, as captured by the classrooms observation exercise. The percentage of teachers conducting desirable reading instruction activities in class has increased in both treatment cohorts and it is higher than in comparison groups. It is important to mention that we recommend in Section 6 a different and more rigorous approach to assess the quality of teaching.

Support supervision of teachers is still not universal. Although treatment teachers have higher probability of receiving support, there is still a significant fraction of teacher that report no supervision at all.

The program has been quite successful at ensuring access to materials, including students' access to Nepali-language 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 improved teaching practices with broad access and use of learning and teaching materials.**

5. LIMITATIONS

Representativeness of the Sample. The sample is only representative of the districts where NEGRP 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 for cohort 1. 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. For cohort 2, the lack of significant differences in the difference-in-differences estimates at midline, where most aspects of the treatment program had still yet to be implemented, strongly supports the parallel trends assumption that, in absence of the NEGRP, 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 endline there are a total of only 86 observations for cohort 1 and 85 or 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 two-week school break. In contrast, midline data collection took place at the end of the school year 2017-18, and endline data collection took place at the end of school year 2019-2020. It is unlikely that this would make much of a difference, but it should be kept in mind when comparing means between baseline and midline or between baseline and endline. However, as all groups –cohort 1, 2 and comparison- had the same data collection schedule, this has no consequences for the integrity of the evaluation.

6. RECOMMENDATIONS

A number of recommendations stem from our findings:

Special attention to L2 Learners: Similar to what we found at midline, the disadvantage in early grade reading skills of L2 learners relative to L1 learners is evident, and the NEGRP, while benefiting everyone, has not done enough to help L2 learners to overcome the disadvantages they face when being educated in a language they have limited command of. 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 supporting non-native 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 non-Nepali language learners in their classrooms.

Improve teacher support supervision: A larger fraction of teachers in both cohorts received more frequent support supervision than comparison groups; however, there are still many teachers that do not receive any supervision at all. In particular, support supervision by district education officers is very low. Evidence suggests that including follow-up classroom visits and teacher support increases learning gains (see for example, 2018 World Development Report). We recommend exploring this challenge and how to effectively scale support supervision within the education system to ensure sustainability of the program.

Improve SMC role: SMC support for reading activities does not show much improvement. This component of the program requires revision and in-depth assessment to understand its challenges and effectiveness.

Parental engagement: the approach used to measure parental engagement was to ask about the importance of learning reading in early grades, reading activities with children at home, and parents' opinion about their educational responsibilities. Parents seem to be well aware of the importance of reading and their role in enabling the process. Most parents also think that teaching how to read is a joint endeavor between the school and the home and that even illiterate parents can help their children. These parents' opinions suggest that raising parental awareness about the importance of early reading is not a priority. Independently of whether or not parents' actual behavior reflects what they report, they seem to be well informed about the issue already. We recommend that in future work qualitative research is conducted through focus group discussions with parents, to learn more about their actual behaviors rather than opinions, and to identify the difficulties they may face when trying to support their children's learning process. This type of research can inform strategies to guide parents in future programs.

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

WORLD BANK, 2018, World Development Report 2018 (WDR 2018)—LEARNING to Realize Education's Promise. Washington DC. (https://www.worldbank.org/en/publication/wdr2018)

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 contribution to the knowledge base on what works in improving early grade literacy in linguistically complex settings.

SUMMARY INFORMATION

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

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

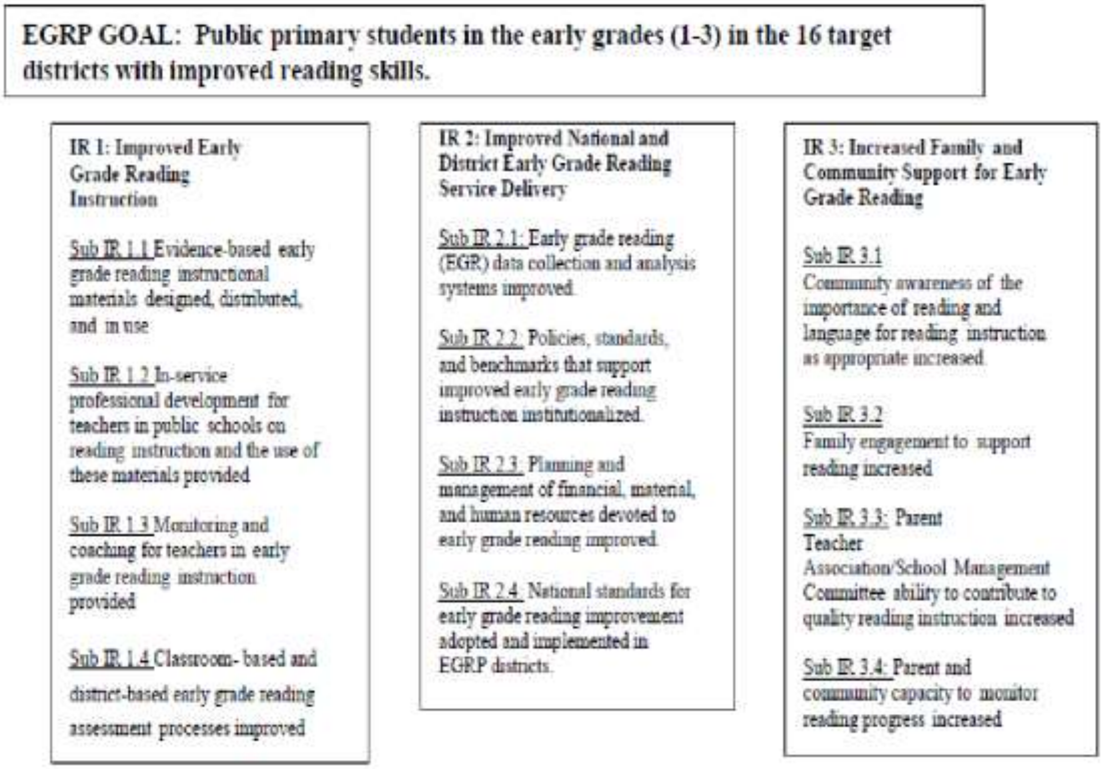
GON service strengthened: The Contractor will have supported the Government of Nepal through Phase One of the NEGRP and completed the design and demonstration of a national model that the Government of Nepal can then implement nationwide within its budget.

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.

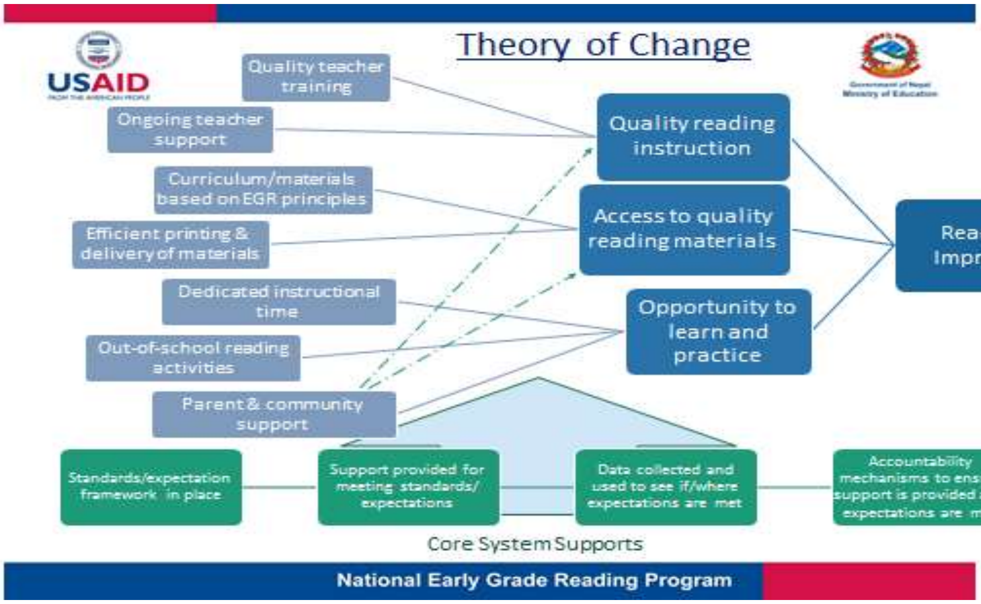
IR 3: Increased Family and Community Support for Early Grade Reading

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



Theory of Change

Quality teacher training

Ongoing teacher support

Curriculum/materials based on EGR principles

Efficient printing & delivery of materials

Dedicated instructional time

Out-of-school reading activities

Parent & community support

Quality reading instruction

Access to quality reading materials

Opportunity to learn and practice

Reading Improves

Standards/expectation framework in place

Support provided for meeting standards/ expectations

Data collected and used to see if/where expectations are met

Accountability mechanisms to ensure support is provided and expectations are met

Core System Supports

National Early Grade Reading Program

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 schools. The exact econometric approach to the comparison will be decided once we have the data and can then assess the different possibilities.
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	

Questions	Suggested Data Sources (*)	Suggested Data Collection Methods	Data Analysis Methods
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?	EGRA	Assessment	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.

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 201mah, 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 201maa, 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.[9]

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.

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.

[9] See **ADS 201mah, USAID Evaluation Report Requirements** and the Evaluation Report Review Checklist from the Evaluation Toolkit for additional guidance.

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.

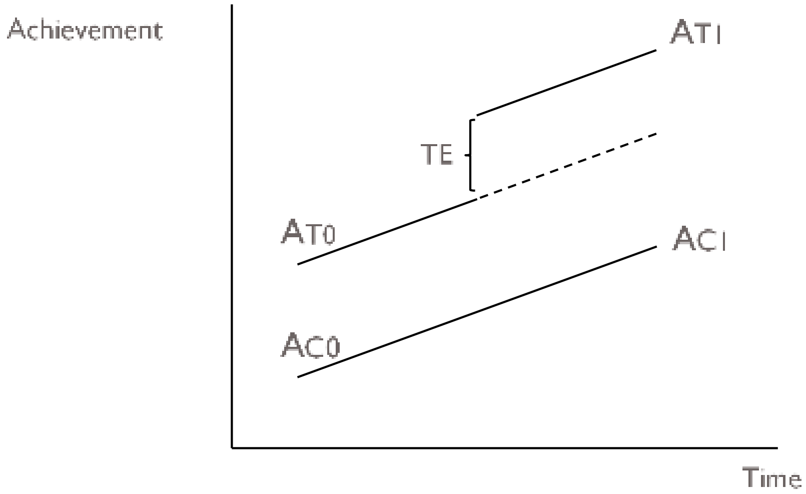
Identifying a credible comparison group is a critical aspect of an impact evaluation and there are several approaches to do so. Our impact evaluation is based on quasi-experimental methods where a comparison group is formed by statistical methods, rather than by random assignment.

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.

	Treatment Cohort 1					Control				
Variable	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

	Treatment Cohort 1					Control				
Variable	Mean	S.E	Min	Max	N	Mean	S.E	Min	Max	N
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
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 1 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*

Variable	Odds ratio [S.E]
	[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]
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

Variable	Odds ratio [S.E]
	[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>chi2	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
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

Variable	Log Likelihood Ratio	Prob>chi2	DF
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
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

Group	Hit Miss Rate	Pseudo R squared
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.

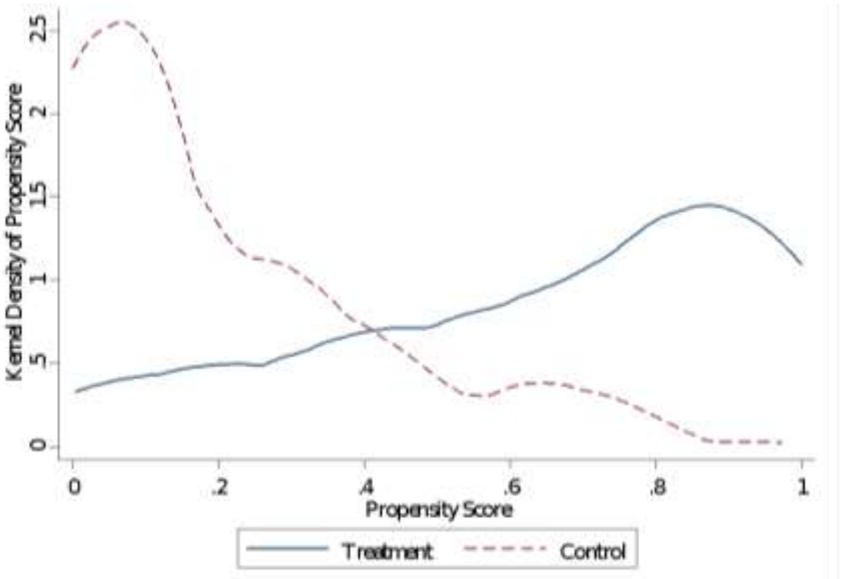
Table A3.5: Balance between treatment and control characteristics of the unmatched sample

Variable	Unmatched				
	Cohort 1	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
School has library facility	0.40	0.45		-0.49	0.62
School provides report cards to parents	0.62	0.35	***	2.83	0.01
School has annual report and social audit	1.22	1.18		0.27	0.79
Number of working computers in school	5.01	1.13	*	1.85	0.07
School have electricity	0.54	0.42		1.27	0.21
Source of water: Tap	0.37	0.52		-1.59	0.11
Average number of assets at home	5.01	4.88		0.50	0.62
Average Matra score grade 1	3.72	3.37		0.36	0.72
Average Matra score grade 2	10.02	10.76		-0.34	0.73
Average Matra score grade 3	17.37	21.20		-1.26	0.21
Average Letter Sound score grade 1	12.21	10.03		1.17	0.24
Average Letter Sound score grade 2	20.20	20.58		-0.13	0.89

Variable	Unmatched				
	Cohort 1	Control		T	P value
Average Letter Sound score grade 3	27.68	29.89		-0.69	0.49
Average Invented Word score grade 1	0.64	0.76		-0.53	0.60
Average Invented Word score grade 2	2.91	3.42		-0.65	0.51
Average Invented Word score grade 3	5.55	6.98		-1.19	0.23
Average Oral Reading score grade 1	1.73	1.66		0.14	0.89
Average Oral Reading score grade 2	7.65	7.72		-0.04	0.97
Average Oral Reading score grade 3	14.70	16.13		-0.49	0.62
Average Reading Comprehension score grade 1	0.17	0.15		0.34	0.73
Average Reading Comprehension score grade 2	0.76	0.76		0.01	1.00
Average Reading Comprehension score grade 3	1.44	1.51		-0.26	0.80
Average Listening Comprehension score grade 1	0.35	0.26		1.01	0.31
Average Listening Comprehension score grade 2	0.56	0.53		0.40	0.69

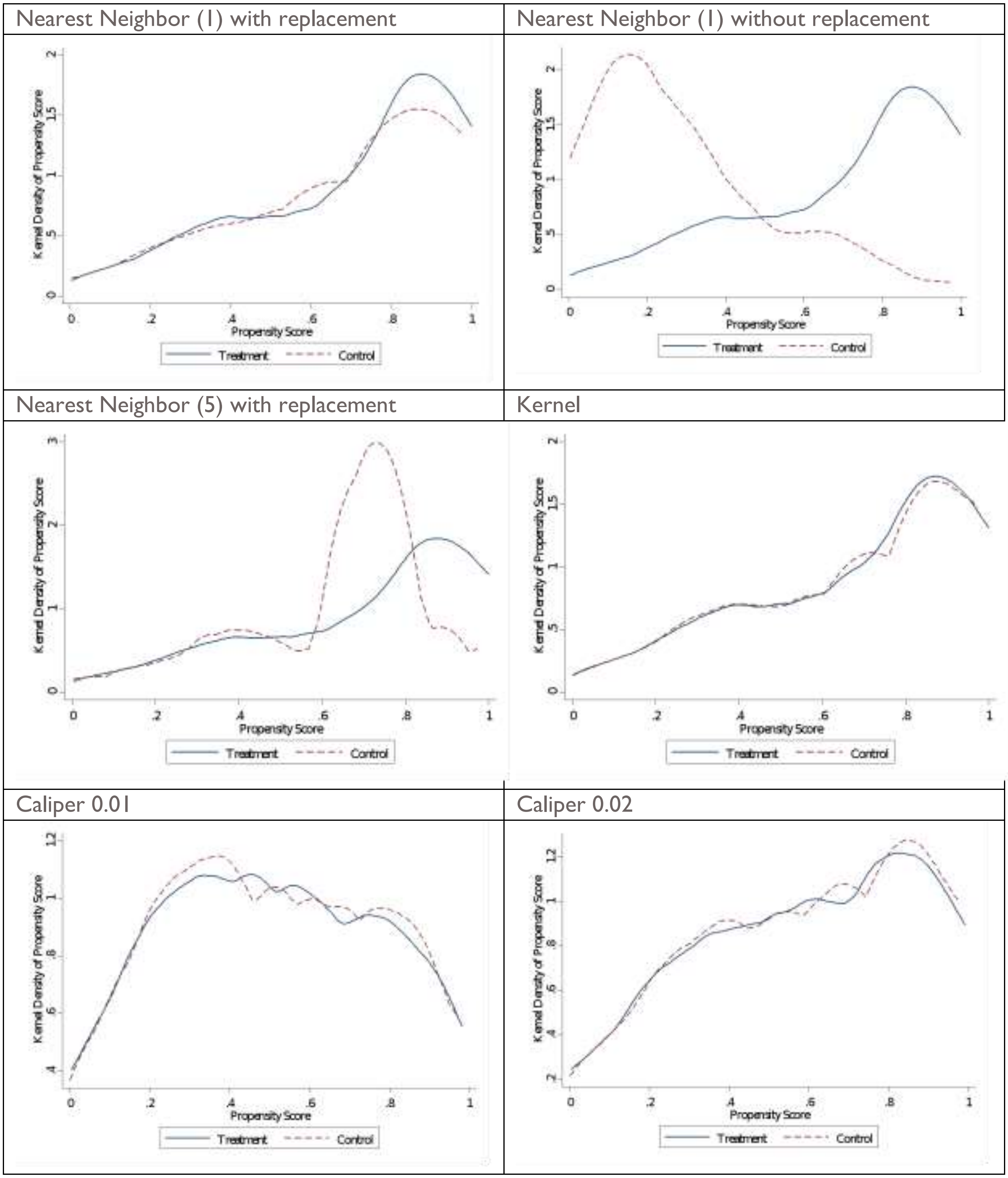
Several covariates seem to differ significantly between treatment and control. The statistical significance is given by a t-test of the difference of the means. The stars indicate that the difference between treatment and control is statistically significant (* at 10%, **, at 5 %, and *** at 1%). Graphically, 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 1 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 1 and Control



Nearest Neighbor (1) with replacement

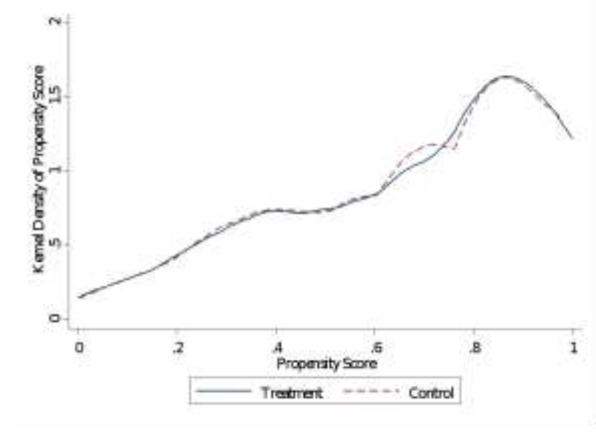
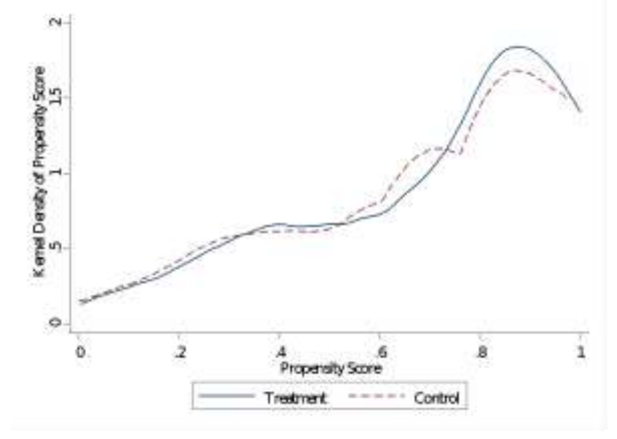
Nearest Neighbor (1) without replacement

Nearest Neighbor (5) with replacement

Kernel

Caliper 0.01

Caliper 0.02

Caliper 0.05	Caliper 0.10
	

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	0.03
Classrooms in Grade 2 (1+)	0.55	0.78	**	-1.98	0.05
Classrooms in Grade 3 (1+)	0.56	0.80	**	-2.23	0.03
Nepali speakers % range	3.03	2.94		0.16	0.87
Number of girls present in classroom	7.56	7.02		0.63	0.53
Grade 2 is mono-grade classroom	0.53	0.59		-0.30	0.77
Teacher Assistant literacy instruction	0.25	0.11	*	1.86	0.06
Guidance to parents to help children become readers	0.73	0.85		-1.42	0.16

Variable					
	Treated	Control		T test	P value
Ask parents to help with homework	0.89	0.98	**	-2.46	0.01
Active parent-teacher association	0.70	0.70		0.02	0.98
School has improvement plan	0.79	0.79		0.04	0.97
Annual program and budget	0.56	0.60		-0.22	0.82
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
	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

Variable	Cohort	Treatment	Comparison	T Test	P value	Effect Size
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. L1 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

Variable	Grade	Cohort	Treatment	Comparison	T Test	P value	Effect Size
	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
	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

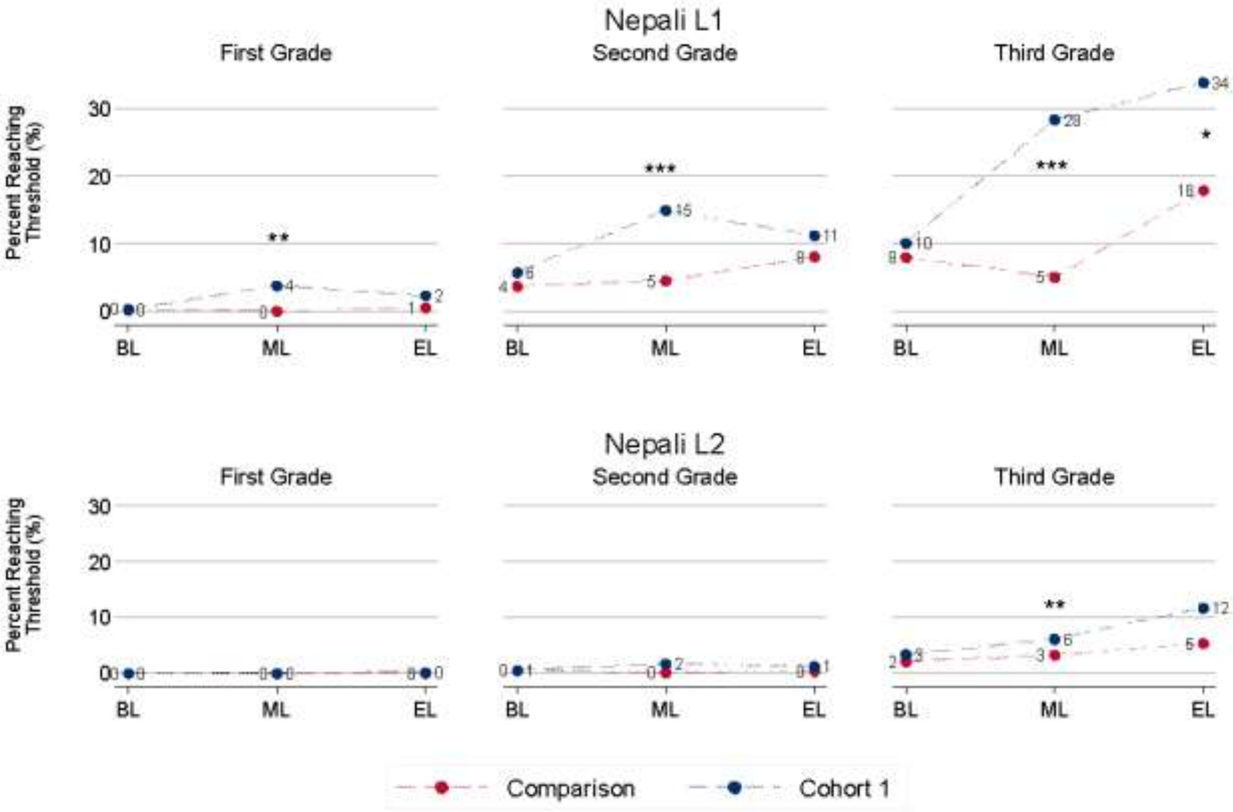
Variable	Grade	Cohort	Treatment	Comparison	T Test	P value	Effect Size
Correct Sound of Letters Per Minute	1	1	6.52	8.06	-1.21	0.23	-0.15
	2		16.37	17.32	-0.45	0.65	-0.06
	3		22.56	23.21	-0.24	0.81	-0.03

Variable	Grade	Cohort	Treatment	Comparison	T Test	P value	Effect Size
	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
	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	8.67	-0.03	0.98	0.00

Variable	Grade	Cohort	Treatment	Comparison	T Test	P value	Effect Size
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

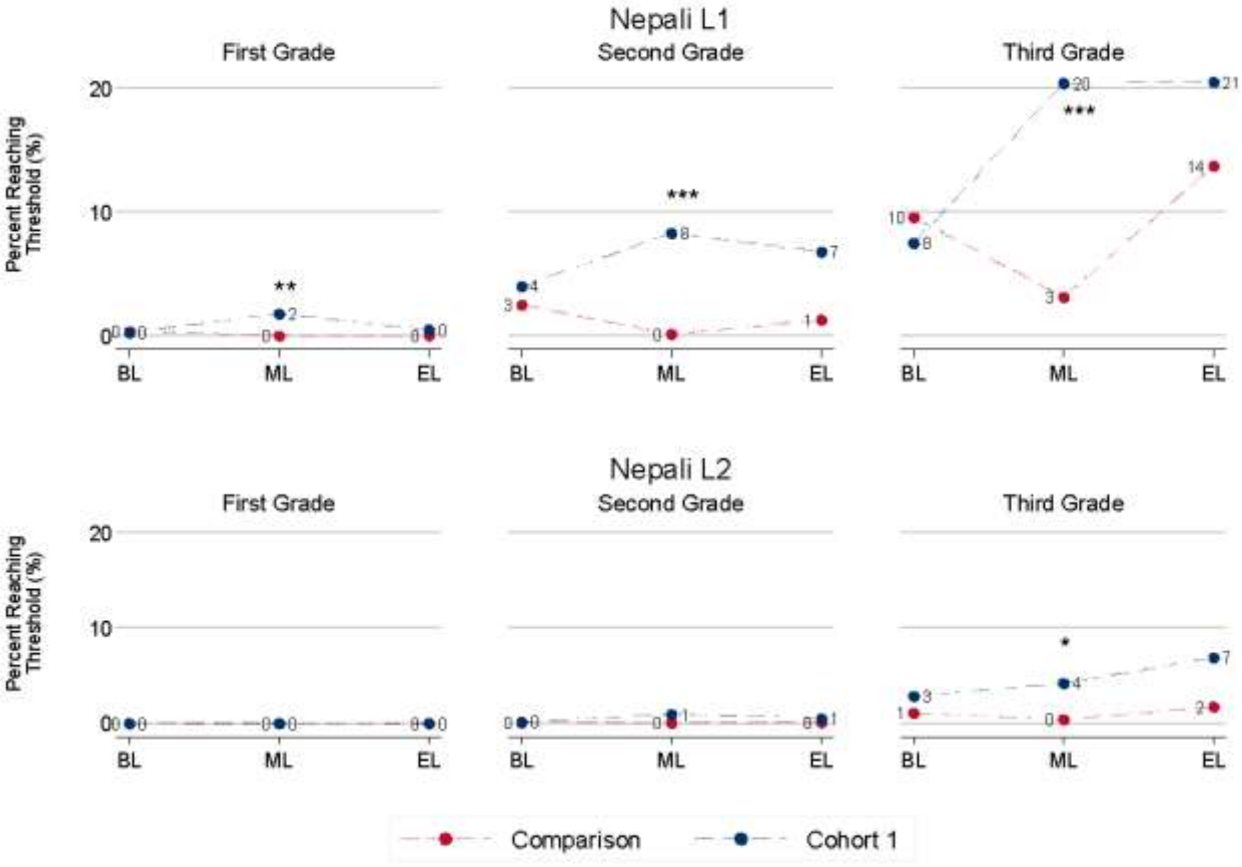
ANNEX IV: ADDITIONAL ANALYSES

Figure A4.1: Percentage of learners reaching Minimum Fluency Threshold (45 CWPM), cohort 1, by grade and learner language



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

Figure A4.2: Percentage of learners reaching Minimum Reading Comprehension Threshold (80%), cohort 1, by grade and learner language

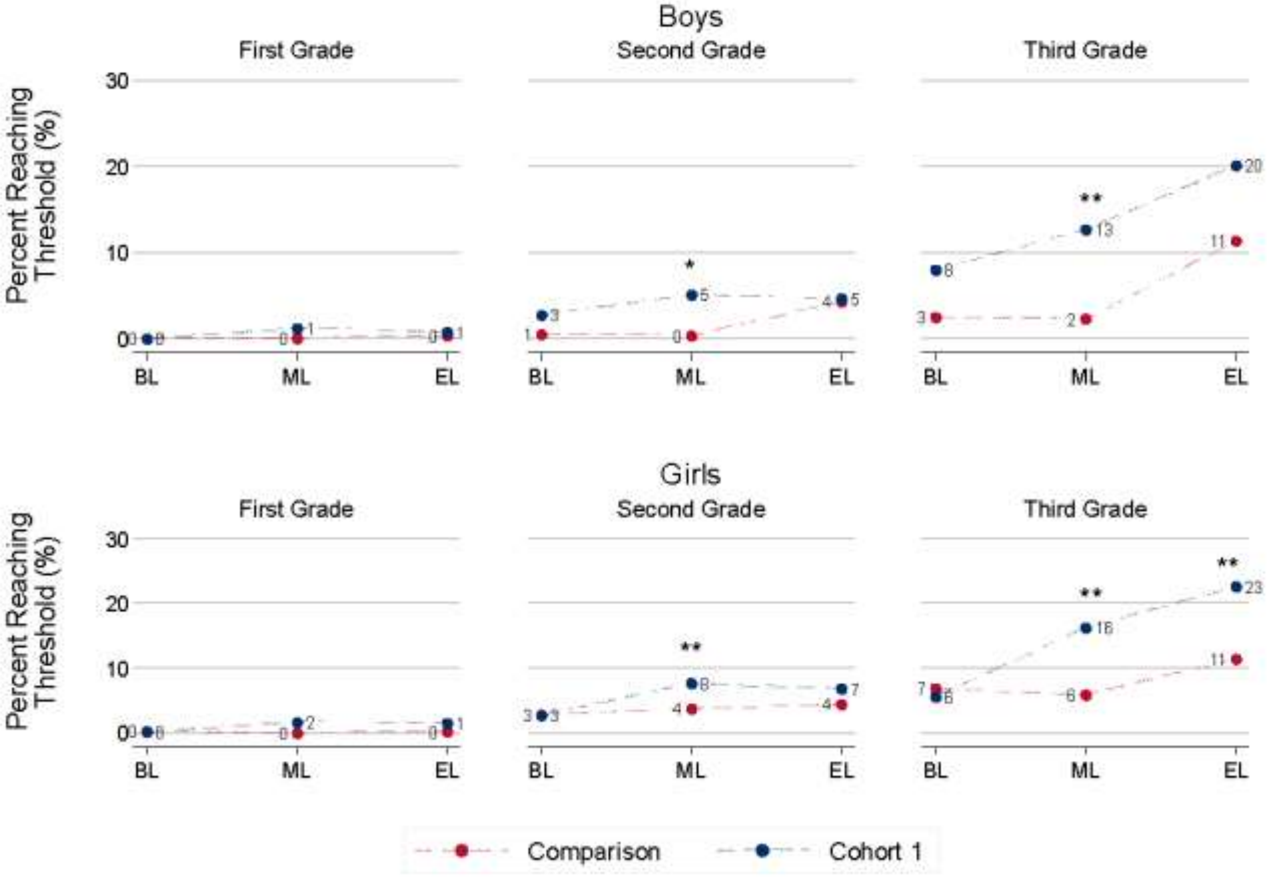


Note: Propensity score matching weights applied. DID compared to baseline *** p<0.01, ** p<0.05, * p<0.1

Table A4.1: NEGRP Effects for Boys and Girls and difference between the Effects (DIDID), Cohort 1

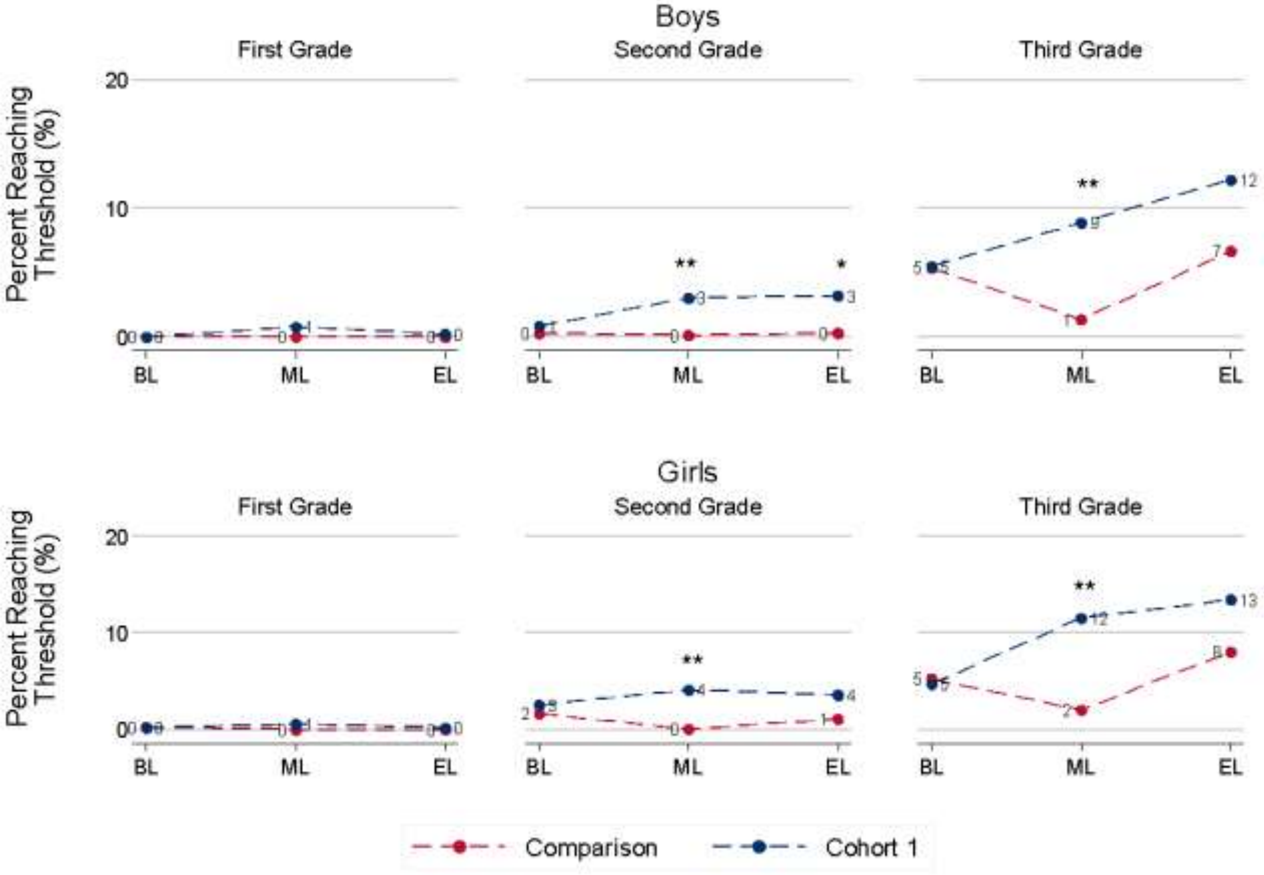
	Male Students			Female Students			Diff in Diff in Diff		
	Baseline Diff (1)	Midline Diff (2)	DiD (3=2 1)	Baseline Diff (4)	Midline Diff (5)	DiD (6=5 4)	DIDID (7=6 3)	Adjusted DIDID	Effect Size
Correct Sound of Letters Per Minute									
Grade 1	2.8	4.2	1.4	0	4.2	4.2*	2.8	4.3	0.33
Grade 2	2.1	5.2	3.1	-0.5	2.6	3.1	0.0	-0.7	-0.04
Grade 3	1.7	7.7	6.0	-2.6	5.8	8.4**	2.4	1.9	0.09
Correct Matra Per Minute									
Grade 1	1.1	2.9	1.8	0.4	2.8	2.4*	0.6	1.5	0.15
Grade 2	2.1	4.8	2.7	0.6	2.7	2.1	-0.6	-1.1	-0.06
Grade 3	1.9	7.3	5.4	-1.3	3.7	5.0	-0.4	-0.8	-0.03
Correct Invented Words Per Minute									
Grade 1	0.2	1.1	0.9**	0	1.2	1.2***	0.3	0.6	0.19
Grade 2	0.8	1.9	1.1	0.6	1.5	0.9	-0.2	-0.3	-0.05
Grade 3	1.2	3.5	2.3*	0.3	2.1	1.8	-0.5	-0.5	-0.05
Oral Reading Fluency									
Grade 1	0.6	1.9	1.3	0.1	2.1	2.0**	0.7	1.4	0.19
Grade 2	1.5	4	2.5	0	2.5	2.5	0.0	-0.4	-0.02
Grade 3	1.2	8	6.8	-1.4	5.6	7.0**	0.2	-0.1	0
Reading Comprehension Percentage of questions correct.									
Grade 1	1	2.6	1.6	0.3	3.6	3.3**	1.7	2.7	0.22
Grade 2	1.4	7.5	6.1	-0.4	5.2	5.6*	-0.5	-1.3	-0.06
Grade 3	-0.3	9.6	9.9	-3	4.3	7.3	-2.6	-2.7	-0.08
Listening Comprehension Percentage of questions correct.									
Grade 1	-1.8	10.6	12.4***	2.9	7.8	4.9	-7.5**	-6.9*	-0.29
Grade 2	-0.3	8.7	9.0**	1.7	4.3	2.6	-6.4	-7.6	-0.27
Grade 3	4.5	1.9	-2.6	-1.9	10.2	12.1***	14.7***	13.7***	0.41

Figure A4.3: Percentage of learners reaching Minimum Fluency Threshold (45 CWPM), cohort 1, by grade and gender



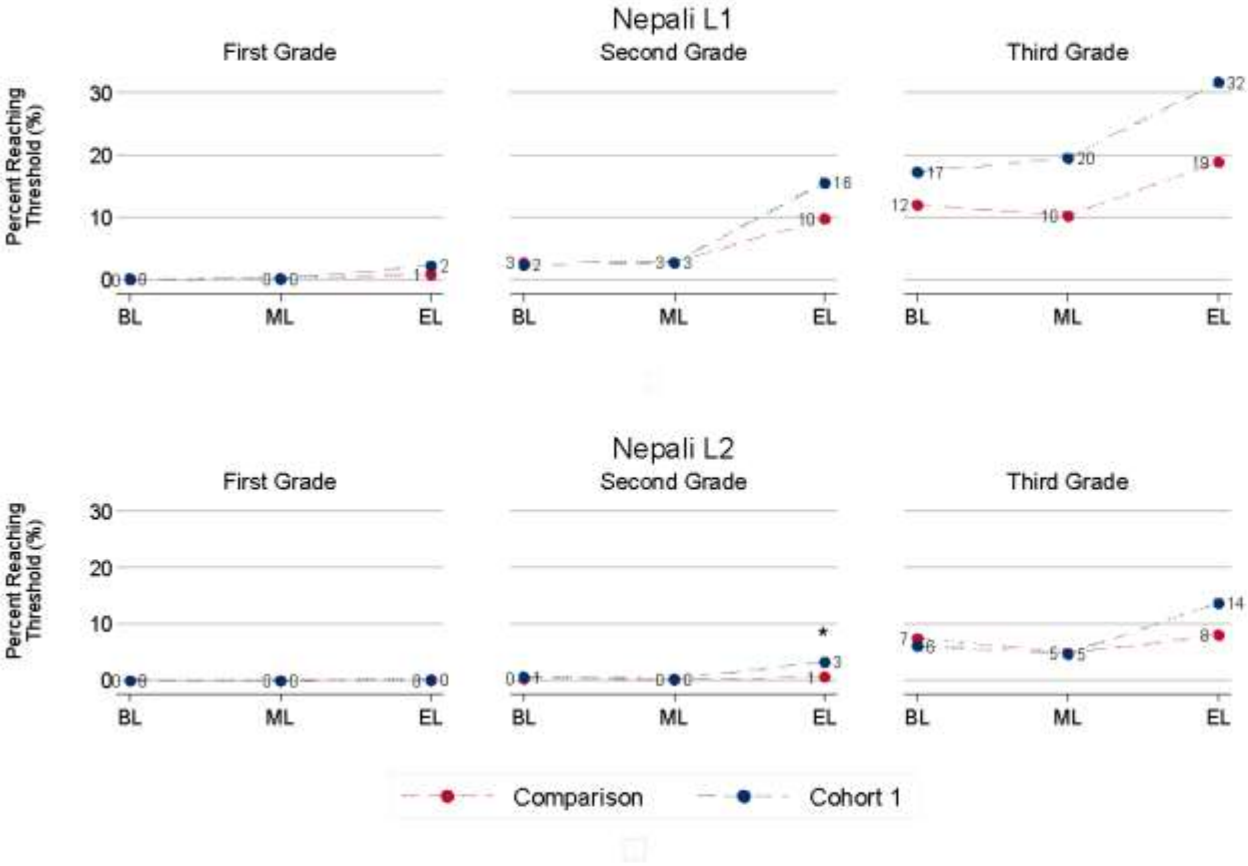
Note: Propensity score matching weights applied. DID compared to baseline *** p<0.01, ** p<0.05, * p<0.1

Figure A4.4: Percentage of learners reaching Minimum Reading Comprehension Threshold (80%), cohort 1, by grade and learner language



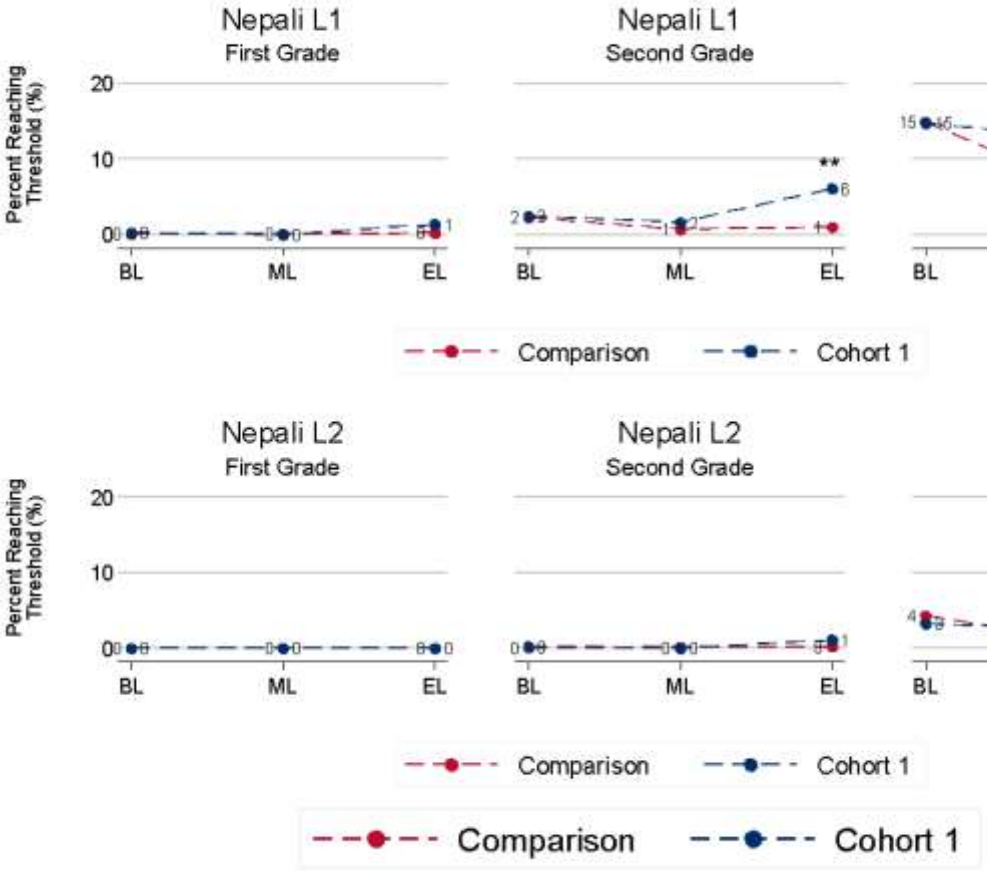
Note: Propensity score matching weights applied. DID compared to baseline *** p<0.01, ** p<0.05, * p<0.1

Figure A4.5: Percentage of learners reaching Minimum Fluency Threshold (45 CWPM), cohort 2, by grade and learner language



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

Figure A4.6: Percentage of learners reaching Minimum Reading Comprehension Threshold (80%), cohort 2, by grade and learner language

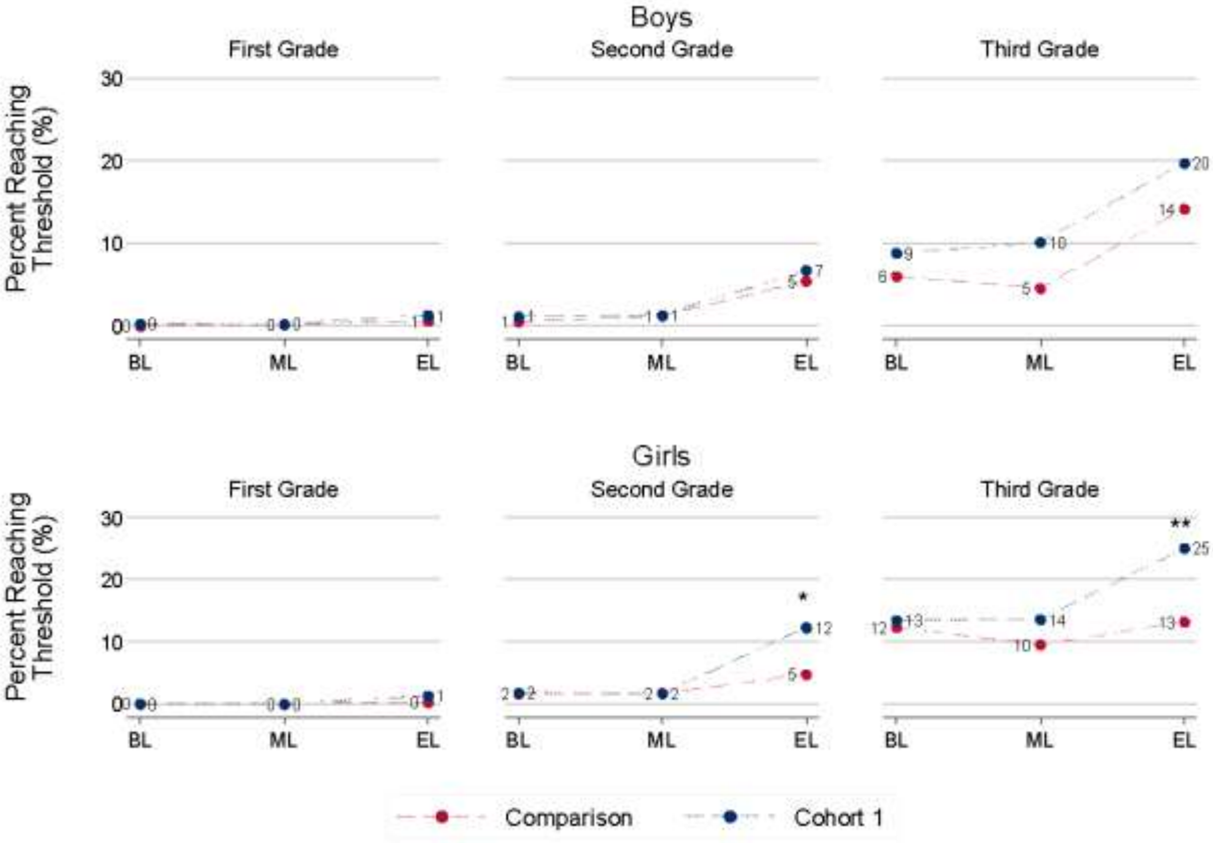


Note: Propensity score matching weights applied. DID compared to baseline *** p<0.01, ** p<0.05, * p<0.1

Table A4.2: NEGRP Effects for Boys and Girls and difference between the effects (DIDID), Cohort 2

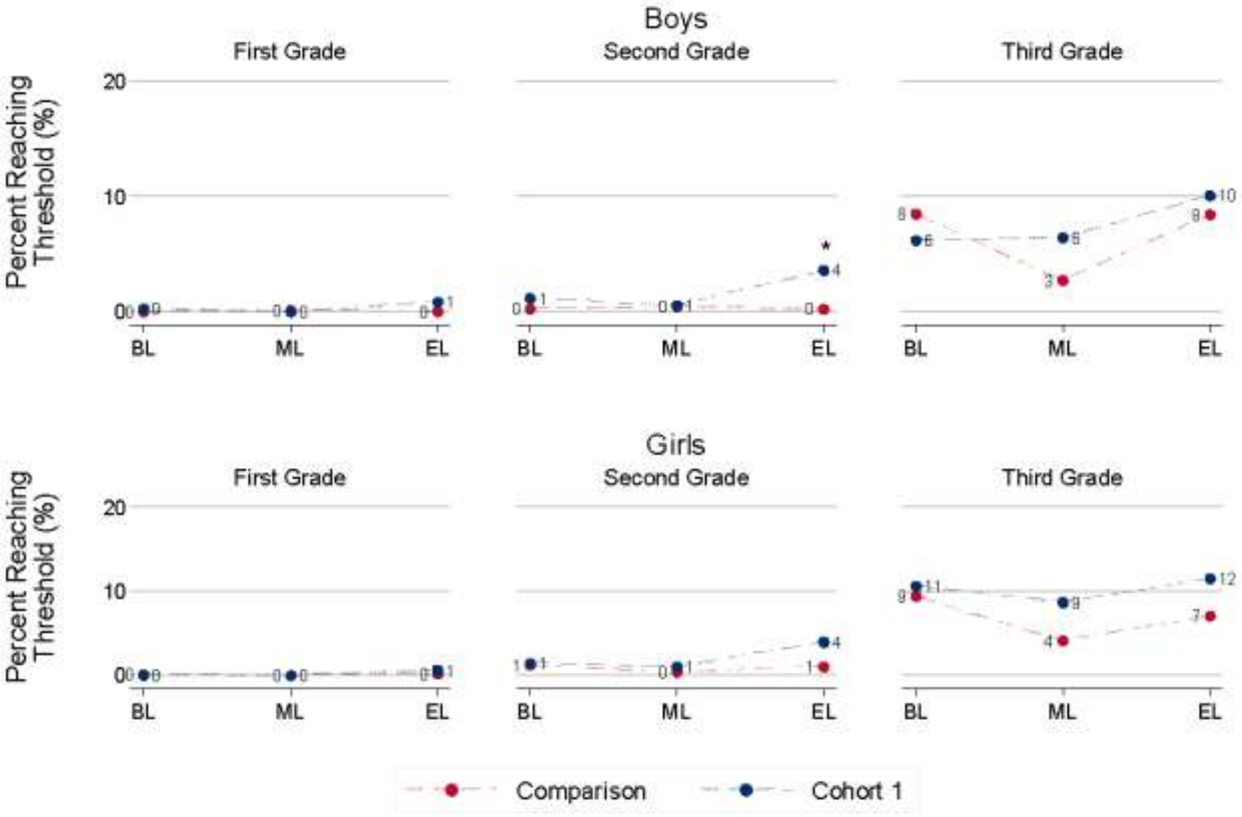
	Male Students			Female Students			Diff in Diff in Diff		
	Baseline Diff (1)	Midline Diff (2)	DiD (3=2 1)	Baseline Diff (4)	Midline Diff (5)	DiD (6=5 4)	DIDID (7=6 3)	Adjusted DIDID	Effect Size
Correct Sound of Letters Per Minute									
Grade 1	1.1	5.3	4.2*	0.9	6.4	5.5**	1.3	2.3	0.16
Grade 2	-0.9	6.4	7.3**	-0.8	5.8	6.6***	-0.7	-1.5	-0.08
Grade 3	-1.3	6.7	8.0*	-0.6	8.2	8.8***	0.8	0.5	0.02
Correct Matra Per Minute									
Grade 1	-0.3	4	4.3***	0.9	4.6	3.7***	-0.6	-0.5	-0.04
Grade 2	0.1	4.9	4.8*	1.5	7.2	5.7*	0.9	0.5	0.03
Grade 3	-0.1	6	6.1**	1.3	7.6	6.3***	0.2	0.1	0
Correct Invented Words Per Minute									
Grade 1	0	1.5	1.5***	0.1	1.8	1.7***	0.2	0.3	0.07
Grade 2	0.4	1.7	1.3	0.7	2.8	2.1*	0.8	0.7	0.1
Grade 3	0.4	2.1	1.7*	1.4	2.7	1.3	-0.4	-0.2	-0.02
Oral Reading Fluency									
Grade 1	-0.1	3	3.1***	0.4	3.6	3.2***	0.1	0.4	0.05
Grade 2	0.5	4.1	3.6	0.3	7.9	7.6***	4.0	3.8	0.21
Grade 3	-1.8	7.3	9.1***	0.6	9.6	9.0***	-0.1	-0.1	0
Reading Comprehension Percentage of questions correct.									
Grade 1	0.1	4.9	4.8***	0.2	5.6	5.4***	0.6	1.2	0.09
Grade 2	1.3	8.2	6.9	-0.6	10.6	11.2***	4.3	4.0	0.16
Grade 3	-4.7	9.2	13.9**	0.7	9.8	9.1***	-4.8	-4.5	-0.14
Listening Comprehension Percentage of questions correct.									
Grade 1	-2.4	4.7	7.1**	2	3.6	1.6	-5.5	-6.4*	-0.31
Grade 2	-4.3	3.8	8.1*	0.8	4.9	4.1	-4.0	-6.4	-0.22
Grade 3	0.3	-0.6	-0.9	0.9	6.2	5.3	6.2	6.0	0.18

Figure A4.7: Percentage of learners reaching Minimum Fluency Threshold (45 CWPM), cohort 2, by grade and gender



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

Figure A4.8: Percentage of learners reaching Minimum Reading Comprehension Threshold (80%), cohort 2, by grade and gender



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

Table A4.3: Materials available to Students in the Classrooms. Cohorts 1 and 2

Cohort 1	Baseline			Endline			DiD (7=6 3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Com p 4	Treat 5	Diff 6		
All or most students have mother language textbook	0	0.06	0.06	0.02	0.04	0.02	-0.04	-0.29
Reading materials (not textbook) easily accessible inside classroom	0.19	0.21	0.02	0.3	0.96	0.66	0.64***	1.34
Curriculum of related subject	0.29	0.21	-0.08	0.43	0.53	0.1	0.18	0.36
Teacher's Guidelines for Nepali Language	0.19	0.13	-0.06	0.47	0.6	0.13	0.19	0.38
Teacher's Guidelines for Local Language	0	0.01	0.01	0	0.06	0.06	0.05*	0.28
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	0.89	0.96	0.07	0.04	0.19
Chalk/Marker	0.88	0.82	-0.06	0.85	0.94	0.09	0.15	0.45
Pen/Pencil	0.55	0.58	0.03	0.81	0.65	-0.16	-0.19	-0.43
Notebook	0.06	0.17	0.11	0.22	0.21	-0.01	-0.12	-0.29

Cohort 2	Baseline			Endline			DiD (7=6 3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
All or most students have mother language textbook	0	0.02	0.02	0.05	0.05	0	-0.02	-0.09
Reading materials (not textbook) easily accessible inside classroom	0.28	0.13	-0.15	0.44	0.92	0.48	0.63***	1.35
Curriculum of related subject	0.25	0.34	0.09	0.51	0.3	-0.21	-0.30*	-0.59
Teacher's Guidelines for Nepali Language	0.19	0.27	0.08	0.46	0.78	0.32	0.24	0.51
Teacher's Guidelines for Local Language	0	0.01	0.01	0	0.01	0.01	0.00	0
Supplementary reading material avail.	0.97	0.95	-0.02	0.95	0.96	0.01	0.03	0.15
Blackboard/Whiteboard	0.9	0.94	0.04	0.91	0.95	0.04	0.00	0
Chalk/Marker	0.43	0.53	0.1	0.81	0.92	0.11	0.01	0.03
Pen/Pencil	0.05	0.12	0.07	0.22	0.43	0.21	0.14	0.3
Notebook	0.97	0.95	-0.02	0.95	0.96	0.01	0.03	0.15

Table A4.4: Teacher Support, Reading Assignments and Attitudes about Reading. Cohorts 1 and 2

Cohort 1	Baseline			Endline			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	0.24	0.04	-0.2	-0.25***	-0.72
Individualized remedial support inside class	0.51	0.49	-0.02	0.41	0.32	-0.09	-0.07	-0.12
Additional practice time inside class	0.49	0.54	0.05	0.13	0.45	0.32	0.27**	0.59
Peer pairing or small group work	0.25	0.38	0.13	0.04	0.18	0.14	0.01	0.00
Whole class revision	0.17	0.23	0.06	0.06	0.18	0.12	0.06	0.15
Additional reading materials or project work outside class	0.03	0.12	0.09	0.01	0.04	0.03	-0.06	-0.33
Additional Support: Parent-teacher conference or communication	0.29	0.14	-0.15	0.17	0.15	-0.02	0.13	0.38
Conducts at least 1 formal meeting w/ parents per term	0.44	0.28	-0.16	0.56	0.45	-0.11	0.05	0.10
Sends at least 1 student progress report to parents per term	0.39	0.35	-0.04	0.61	0.46	-0.15	-0.11	-0.20
Reading Assignments								
Gives daily reading assignment to complete outside school	0.79	0.68	-0.11	0.36	0.83	0.47	0.58***	1.20
Attitudes								
All learners can learn to read.	0.53	0.67	0.14	0.66	0.65	-0.01	-0.15	-0.32
All learners can learn to write.	0.59	0.76	0.17	0.66	0.78	0.12	-0.05	-0.11
Children acquire reading skills by exposure, without being taught to read.	0.41	0.4	-0.01	0.14	0.22	0.08	0.09	0.23
Give learners time each day to read freely materials of their own choice.	0.98	0.94	-0.04	0.97	1.00	0.03	0.07**	0.60
Learners must be able to recite a text before they can read it.	0.4	0.42	0.02	0.3	0.12	-0.18	-0.20	-0.52
Better to teach R&W separately	0.68	0.87	0.19	0.59	0.71	0.12	-0.07	-0.17
Learners cannot write an original passage until at least grade 3 or 4.	0.67	0.69	0.02	0.66	0.49	-0.17	-0.19	-0.38
Important to give learner time each day to write on topics of own choice.	0.99	0.92	-0.07	0.99	0.99	0.00	0.07*	0.61
It is important to correct ALL the errors in sentences learners produce.	0.96	0.9	-0.06	0.91	0.94	0.03	0.09	0.30
Reading stories to learners helps them develop their reading skills	0.66	0.78	0.12	0.8	0.83	0.03	-0.09	-0.23
Young learners must memorize a text before they can understand it.	0.29	0.23	-0.06	0.46	0.31	-0.15	-0.09	-0.20

Cohort 1	Baseline			Endline			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	0.72	0.77	0.05	-0.03	-0.07
A learner writes "well" is not make any grammatical or spelling mistake.	0.67	0.66	-0.01	0.78	0.67	-0.11	-0.10	-0.25
Some students learn to read more slowly as not understand language well.	0.94	0.86	-0.08	0.76	0.63	-0.13	-0.05	-0.11
If a student can read quickly, that means he/she is a good reader.	0.58	0.71	0.13	0.85	0.63	-0.22	-0.35**	-0.77

Cohort 2	Baseline			Endline			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	0.26	0.18	-0.08	-0.12	-0.29
Individualized remedial support inside class	0.51	0.64	0.13	0.38	0.28	-0.1	-0.23*	-0.49
Additional practice time inside class	0.47	0.53	0.06	0.24	0.28	0.04	-0.02	-0.05
Peer pairing or small group work	0.25	0.36	0.11	0.07	0.18	0.11	0.00	-0.03
Whole class revision	0.09	0.19	0.1	0.06	0.09	0.03	-0.07	-0.31
Additional reading materials or project work outside class	0.03	0.06	0.03	0.01	0.12	0.11	0.08*	0.32
Additional Support: Parent-teacher conference or communication	0.18	0.22	0.04	0.2	0.12	-0.08	-0.12	-0.32
Conducts at least 1 formal meeting w/ parents per term	0.48	0.3	-0.18	0.65	0.52	-0.13	0.05	0.08
Sends at least 1 student progress report to parents per term	0.42	0.14	-0.28	0.56	0.43	-0.13	0.15	0.28
Reading Assignments								
Gives daily reading assignment to complete outside school	0.82	0.67	-0.15	0.43	0.65	0.22	0.37***	0.74
Attitudes								
All learners can learn to read.	0.55	0.69	0.14	0.63	0.54	-0.09	-0.23**	-0.47
All learners can learn to write.	0.55	0.66	0.11	0.60	0.70	0.10	-0.01	-0.02
Children acquire reading skills by exposure, without being taught to read.	0.35	0.36	0.01	0.15	0.17	0.02	0.01	0.05
Give learners time each day to read freely materials of their own choice.	0.98	0.94	-0.04	0.94	0.98	0.04	0.08	0.35
Learners must be able to recite a text before they can read it.	0.34	0.29	-0.05	0.11	0.07	-0.04	0.01	0.03

Cohort 2	Baseline			Endline			DiD (7=6 3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
Better to teach R&W separately	0.73	0.78	0.05	0.54	0.62	0.08	0.03	0.06
Learners cannot write an original passage until at least grade 3 or 4.	0.70	0.69	-0.01	0.63	0.59	-0.04	-0.03	-0.06
Important to give learner time each day to write on topics of own choice.	0.98	0.95	-0.03	0.99	0.99	0.00	0.03	0.21
It is important to correct ALL the errors in sentences learners produce.	0.92	0.86	-0.06	0.93	0.96	0.03	0.09	0.44
Reading stories to learners helps them develop their reading skills	0.75	0.81	0.06	0.8	0.84	0.04	-0.02	-0.03
Young learners must memorize a text before they can understand it.	0.28	0.34	0.06	0.35	0.26	-0.09	-0.15	-0.33
Silent reading should be avoided, as it can't check if learner reading.	0.57	0.72	0.15	0.77	0.77	0	-0.15	-0.38
A learner writes "well" is not make any grammatical or spelling mistake.	0.67	0.66	-0.01	0.82	0.66	-0.16	-0.15	-0.34
Some students learn to read more slowly as not understand language well.	0.95	0.75	-0.2	0.64	0.65	0.01	0.21*	0.44
If a student can read quickly, that means he/she is a good reader.	0.63	0.64	0.01	0.83	0.68	-0.15	-0.16	-0.37

Table A4.5: Teacher Practices Indexes. Cohorts 1 and 2

Cohort 1	Baseline			Endline			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	12.59	18.6	6.01	6.56***	1.16
Student-Centered Classroom Index	3.64	4.19	0.55	4.59	5.46	0.87	0.32	0.21

Cohort 2	Baseline			Endline			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	13.98	17.57	3.59	4.52***	0.85
Student-Centered Classroom Index	4.07	3.86	-0.21	4.85	5.46	0.61	0.82**	0.58

Table A4.6: School Management Committee. Cohorts 1 and 2

Cohort 1	Baseline			Endline			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	0.28	0.21	-0.07	-0.17	-0.37
Management Index	7.05	7.18	0.13	7.92	8.01	0.09	-0.04	-0.02
SMC met at least once per month in last year	0.68	0.52	-0.16	0.5	0.48	-0.02	0.14	0.26
Early Grade Literacy is high priority for SMC	0.71	0.47	-0.24	0.51	0.64	0.13	0.37*	0.73
PTA meets at least every two months this year	0.32	0.26	-0.06	0.07	0.2	0.13	0.19	0.56
School engages PTA/community for book drives & book donations	0.39	0.48	0.09	0.38	0.34	-0.04	-0.13	-0.27
School works with PTA to manage resources for EGR improvement programs	0.55	0.55	0	0.6	0.65	0.05	0.05	0.1
School provided guidance to parents to help children read	0.84	0.78	-0.06	0.87	0.79	-0.08	-0.02	-0.05
School asks parents to help with homework and read to children	0.95	0.9	-0.05	0.95	0.87	-0.08	-0.03	-0.07

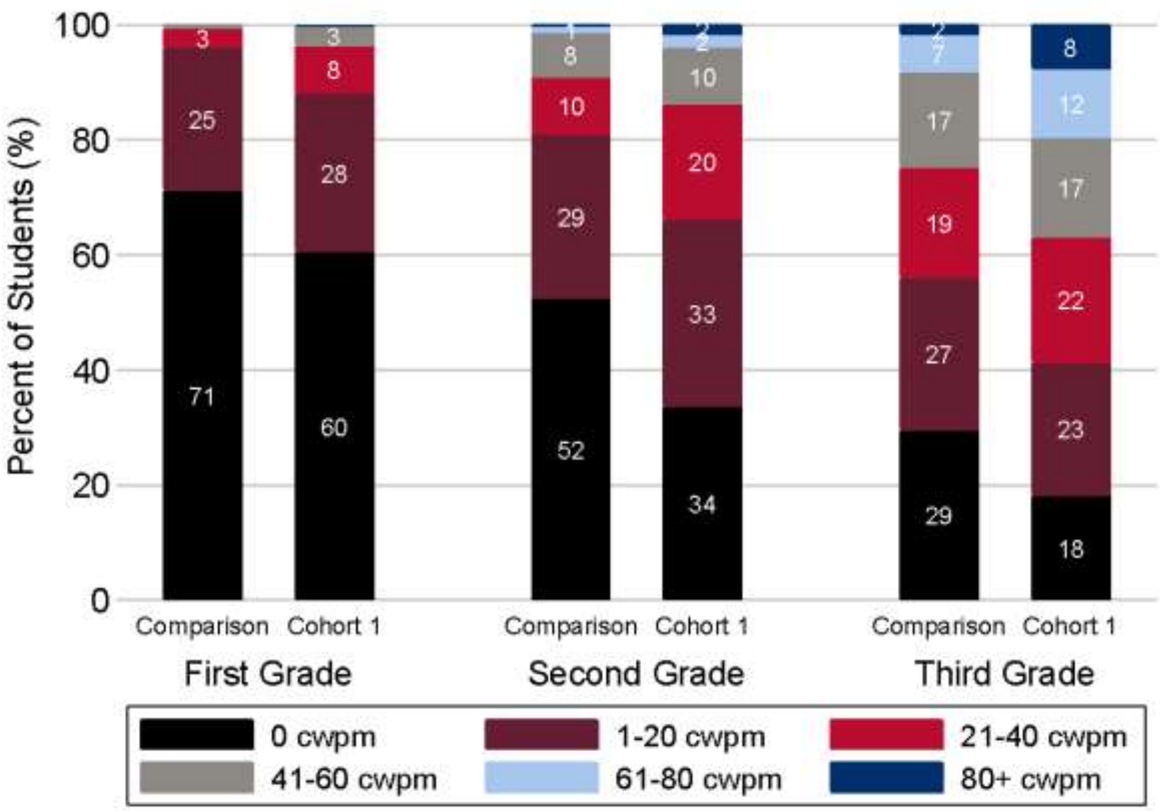
Cohort 2	Baseline			Endline			DiD (7=6 3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
Received school management capacity building training in last two years	0.43	0.37	-0.06	0.35	0.51	0.16	0.22	0.44
Management Index	7.8	7.1	-0.7	7.96	8.54	0.58	1.28**	0.54
SMC met at least once per month in last year	0.57	0.57	0	0.49	0.48	-0.01	-0.01	0
Early Grade Literacy is high priority for SMC	0.68	0.45	-0.23	0.52	0.76	0.24	0.47**	0.98

Cohort 2	Baseline			Endline			DiD (7=6 3)	Effect Size
Teachers	Comp 1	Treat 2	Diff 3	Comp 4	Treat 5	Diff 6		
PTA meets at least every two months this year	0.32	0.24	-0.08	0.07	0.3	0.23	0.31*	0.79
School engages PTA/community for book drives & book donations	0.31	0.41	0.1	0.5	0.53	0.03	-0.07	-0.12
School works with PTA to manage resources for EGR improvement programs	0.59	0.53	-0.06	0.75	0.88	0.13	0.19**	0.51
School provided guidance to parents to help children read	0.86	0.69	-0.17	0.86	0.83	-0.03	0.14*	0.41
School asks parents to help with homework and read to children	0.96	0.94	-0.02	0.93	0.98	0.05	0.07	0.33

Table A4.7: At Home Reading Activities. Cohorts 1 and 2

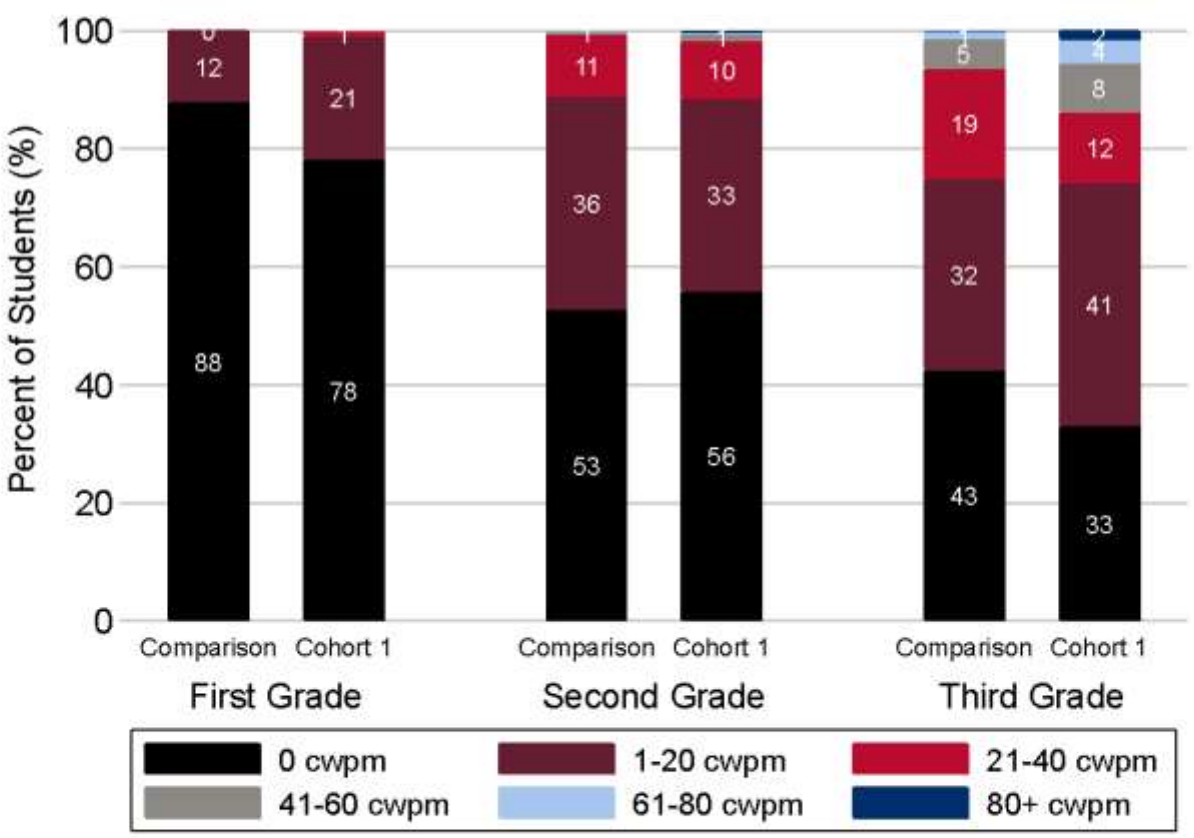
	Baseline			Endline			DiD (7=6 3)	Effect Size
	Comp	Treat	Diff	Comp	Treat	Diff		
Cohort 1								
Subscribe to children's magazines	0.02	0.09	0.07	0.03	0.05	0.02	-0.05	-0.25
You or someone in your household reads to your child at least once a week	1	0.71	-0.29	0.85	0.75	-0.1	0.19	0.44
Your child reads to you or someone in your household at least once a week	0	0.01	0.01	0.02	0.03	0.01	0.00	-0.06
Cohort 2								
Subscribe to children's magazines	0	0	0	0.02	0.05	0.03	0.03	0.17
You or someone in your household reads to your child at least once a week	0.62	0.59	-0.03	0.53	0.71	0.18	0.21	0.41
Your child reads to you or someone in your household at least once a week	0.58	0.52	-0.06	0.9	0.89	-0.01	0.05	0.16

Figure A4.9: Oral Reading Fluency Distributions at Endline, by Grade. Cohort 1, Nepali L1



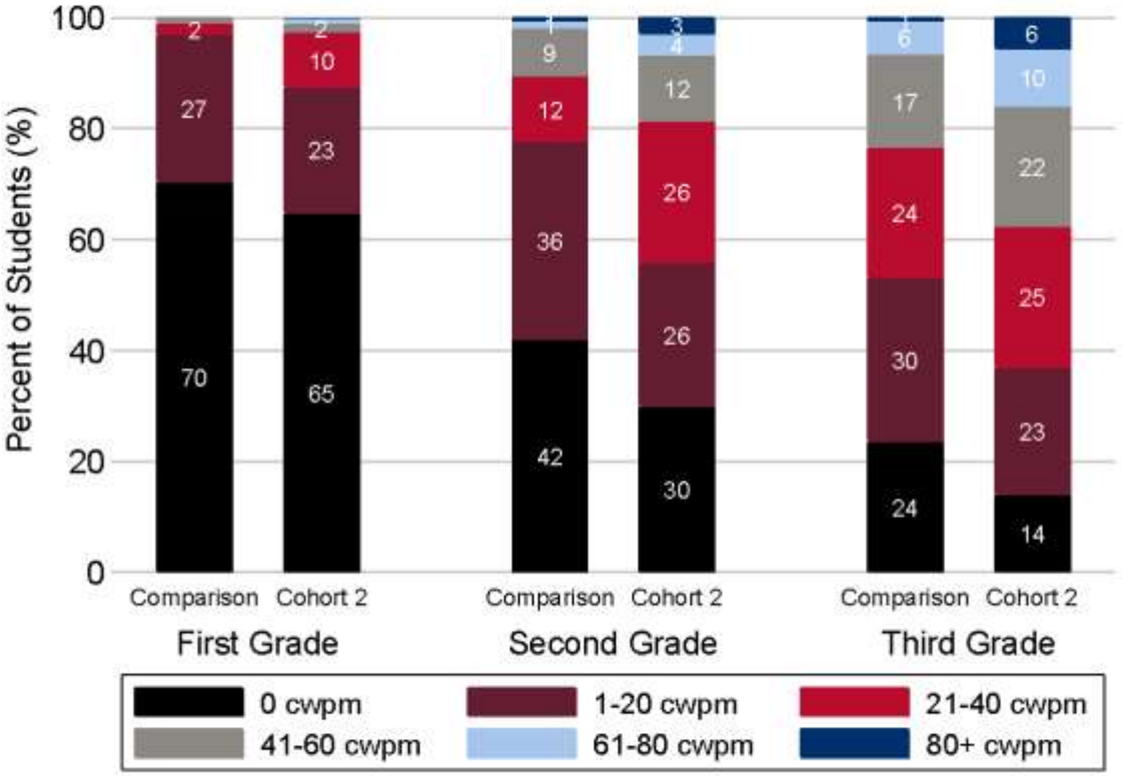
Note: Propensity score matching weights applied.

Figure A4.10: Oral Reading Fluency Distributions at Endline, by Grade. Cohort 1, Nepali L2



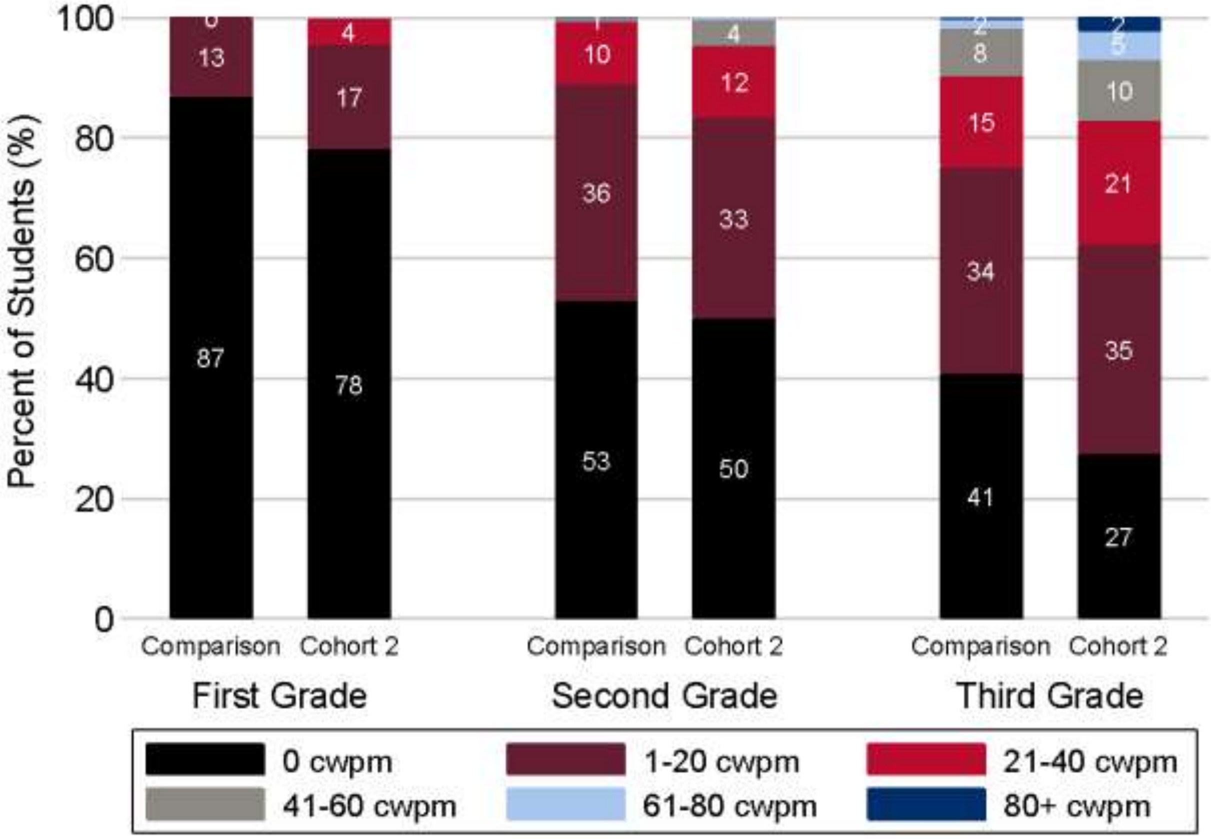
Note: Propensity score matching weights applied.

Figure A4.11: Oral Reading Fluency Distributions at Endline, by Grade. Cohort 2, Nepali L1



Note: Propensity score matching weights applied.

Figure A4.12: Oral Reading Fluency Distributions at Endline, by Grade. Cohort 2, Nepali L2



Note: Propensity score matching weights applied.

Table A4.8 Percentage of Non-readers (zero cwpm) by treatment group, language, grade and sex

	Grade 1			Grade 2			Grade 3		
ALL	Baseline	Midline	Endline	Baseline	Midline	Endline	Baseline	Midline	Endline
Treatment	85.8%	79.1%	73.1%	57.7%	52.6%	44.6%	38.5%	31.2%	23.8%
Cohort 1	84.9%	70.1%	69.8%	61.7%	46.8%	48.7%	44.5%	29.5%	27.0%
L1 Students	73.3%	44.2%	57.2%	48.1%	26.3%	31.4%	31.6%	18.4%	15.7%
L2 Students	90.3%	80.5%	75.5%	67.7%	56.0%	57.5%	51.3%	34.6%	32.7%
Cohort 2	86.3%	84.3%	74.9%	55.4%	56.0%	42.2%	35.3%	32.2%	22.0%
L1 Students	78.7%	77.9%	67.3%	43.4%	46.4%	30.0%	21.3%	17.7%	14.6%
L2 Students	91.7%	89.7%	80.6%	65.5%	65.0%	53.5%	45.9%	44.4%	27.9%
Comparison	89.5%	90.8%	88.0%	62.0%	68.4%	61.2%	41.8%	39.0%	41.1%
L1 Students	77.4%	82.5%	80.2%	46.0%	47.2%	50.8%	26.6%	27.6%	27.4%
L2 Students	92.4%	92.7%	90.4%	66.6%	75.7%	66.9%	47.1%	42.8%	47.5%
BOYS	Baseline	Midline	Endline	Baseline	Midline	Endline	Baseline	Midline	Endline
Treatment	86.0%	79.9%	71.8%	57.1%	53.1%	44.3%	36.0%	33.5%	23.7%
Cohort 1	86.2%	71.3%	65.6%	61.7%	48.2%	42.8%	42.0%	31.4%	26.6%
L1 Students	77.5%	46.1%	53.1%	53.4%	26.6%	29.0%	30.1%	17.9%	16.0%
L2 Students	90.6%	82.5%	71.7%	65.9%	60.0%	50.6%	49.9%	37.5%	33.3%
Cohort 2	86.0%	84.6%	75.5%	54.4%	55.8%	45.3%	33.3%	34.6%	22.2%
L1 Students	76.5%	81.0%	66.2%	44.8%	48.8%	33.4%	22.8%	19.6%	14.5%
L2 Students	93.0%	87.9%	82.8%	62.4%	63.1%	55.3%	41.5%	46.6%	28.2%
Comparison	87.8%	88.8%	87.0%	55.8%	67.8%	65.7%	36.4%	36.5%	45.4%
L1 Students	71.3%	79.2%	80.8%	42.7%	50.1%	61.4%	30.9%	32.9%	36.2%

	Grade 1			Grade 2			Grade 3		
L2 Students	90.6%	91.4%	89.0%	60.3%	73.8%	68.2%	38.2%	37.7%	50.2%
GIRLS	Baseline	Midline	Endline	Baseline	Midline	Endline	Baseline	Midline	Endline
Treatment	85.6%	78.4%	74.3%	58.1%	52.2%	45.1%	40.6%	29.5%	23.9%
Cohort 1	84.1%	68.1%	74.0%	61.4%	45.8%	53.2%	46.3%	28.1%	27.3%
L1 Students	70.1%	42.0%	61.5%	42.7%	26.0%	33.5%	33.0%	18.8%	15.3%
L2 Students	90.2%	79.0%	79.1%	68.9%	53.5%	62.3%	52.1%	32.5%	32.3%
Cohort 2	86.6%	84.2%	74.5%	56.1%	56.1%	40.3%	37.1%	30.4%	21.8%
L1 Students	80.6%	74.8%	68.3%	41.9%	44.0%	28.3%	20.0%	16.2%	14.7%
L2 Students	90.6%	91.3%	78.9%	68.2%	66.5%	52.0%	49.7%	42.6%	27.6%
Comparison	91.1%	92.3%	88.8%	67.0%	68.3%	58.1%	45.6%	40.9%	38.3%
L1 Students	81.2%	85.6%	79.7%	49.4%	42.5%	42.5%	23.9%	23.6%	20.7%
L2 Students	94.1%	93.7%	91.5%	71.3%	77.0%	66.0%	53.3%	46.8%	45.9%

Note: Sample weights applied to recover population representativeness

Table A4.9 Average Oral Reading Fluency (cwpm) by treatment group, language, grade and sex

	Grade 1			Grade 2			Grade 3		
ALL	Baseline	Midline	Endline	Baseline	Midline	Endline	Baseline	Midline	Endline
Treatment	1.5	2.3	3.6	7.1	8.4	12.6	15.7	17.5	23.2
Cohort 1	1.7	4.2	3.4	7.0	10.8	10.8	13.3	19.4	21.2
L1 Students	3.9	10.2	6.8	12.3	22.8	19.7	20.6	31.9	32.4
L2 Students	0.7	1.8	2.3	4.7	5.4	6.2	9.5	13.6	15.6
Cohort 2	1.5	1.2	3.5	7.1	7.0	13.7	17.0	16.5	24.4
L1 Students	2.3	1.9	5.4	10.2	10.1	20.0	22.8	24.8	30.8
L2 Students	0.8	0.7	2.1	4.5	4.0	7.8	12.5	9.5	19.3
Comparison	1.2	0.8	0.9	6.4	4.6	6.2	14.1	12.5	13.8
L1 Students	2.9	1.9	2.5	11.4	8.9	10.1	23.3	18.3	22.5

	Grade 1			Grade 2			Grade 3		
L2 Students	0.8	0.6	0.4	5.0	3.1	4.1	10.9	10.5	9.7
BOYS	Baseline	Midline	Endline	Baseline	Midline	Endline	Baseline	Midline	Endline
Treatment	1.7	2.2	3.8	6.6	8.0	11.5	15.7	16.3	22.3
Cohort 1	1.8	4.0	4.4	6.8	10.2	11.7	15.5	17.7	20.2
L1 Students	3.9	9.2	6.6	10.9	20.2	20.2	20.3	28.3	32.4
L2 Students	0.7	1.7	3.4	4.7	4.7	6.9	12.4	12.9	12.6
Cohort 2	1.6	1.3	3.4	6.5	6.7	11.4	15.8	15.5	23.3
L1 Students	2.6	1.8	5.6	8.6	9.2	15.8	20.1	23.3	29.0
L2 Students	0.8	0.8	1.7	4.7	4.2	7.7	12.5	9.3	18.9
Comparison	1.4	1.0	0.9	7.2	4.8	5.8	14.0	12.9	13.1
L1 Students	3.8	2.4	2.6	9.9	8.1	8.2	18.8	15.5	20.3
L2 Students	1.0	0.6	0.3	6.3	3.7	4.3	12.4	12.0	9.4
GIRLS	Baseline	Midline	Endline	Baseline	Midline	Endline	Baseline	Midline	Endline
Treatment	1.4	2.4	3.4	7.5	8.7	13.2	15.6	18.5	24.0
Cohort 1	1.6	4.7	3.0	7.3	11.2	10.1	11.8	20.6	21.8
L1 Students	4.0	11.1	7.1	13.8	25.2	19.4	20.8	34.3	32.4
L2 Students	0.6	1.9	1.3	4.7	5.8	5.8	7.8	14.1	17.4
Cohort 2	1.3	1.2	3.6	7.7	7.1	15.1	17.9	17.2	25.3
L1 Students	2.1	2.0	5.2	11.6	11.0	22.6	25.1	26.0	32.4
L2 Students	0.8	0.5	2.5	4.3	3.8	7.8	12.6	9.7	19.6
Comparison	1.0	0.7	0.9	5.7	4.5	6.6	14.1	12.2	14.2
L1 Students	2.2	1.5	2.5	13.0	10.1	11.6	26.1	20.4	24.2
L2 Students	0.6	0.6	0.4	4.0	2.7	4.1	9.8	9.3	10.0

Note: Sample weights applied to recover population representativeness

Table A4.10 Percentage of learners reaching reading benchmark (45 cwpm and 80% comprehension) by treatment group, language, grade and sex

	Grade 1			Grade 2			Grade 3		
ALL	Baseline	Midline	Endline	Baseline	Midline	Endline	Baseline	Midline	Endline
Treatment	0.1%	0.2%	0.6%	0.9%	2.0%	3.2%	5.9%	7.6%	9.4%
Cohort 1	0.2%	0.5%	0.8%	1.6%	3.4%	3.7%	5.1%	9.6%	10.7%
L1 Students	0.6%	1.6%	1.0%	5.2%	9.6%	9.1%	11.1%	24.1%	17.4%
L2 Students	0.0%	0.0%	0.7%	0.0%	0.6%	0.9%	2.0%	2.7%	7.4%

	Grade 1			Grade 2			Grade 3		
Cohort 2	0.1%	0.0%	0.4%	0.5%	1.2%	3.0%	6.3%	6.6%	8.6%
L1 Students	0.2%	0.0%	1.0%	1.0%	2.5%	4.9%	10.7%	11.8%	13.9%
L2 Students	0.0%	0.0%	0.0%	0.1%	0.0%	1.1%	2.9%	2.2%	4.3%
Comparison	0.0%	0.0%	0.0%	0.8%	0.2%	1.0%	5.3%	3.2%	4.3%
L1 Students	0.2%	0.0%	0.0%	2.1%	0.3%	1.4%	15.2%	6.2%	8.2%
L2 Students	0.0%	0.0%	0.0%	0.5%	0.2%	0.8%	1.8%	2.1%	2.5%
BOYS	Baseline	Midline	Endline	Baseline	Midline	Endline	Baseline	Midline	Endline
Treatment	0.1%	0.1%	0.8%	0.9%	1.4%	3.0%	5.2%	5.6%	9.0%
Cohort 1	0.0%	0.4%	1.2%	1.3%	3.2%	3.5%	6.5%	7.4%	9.9%
L1 Students	0.0%	1.3%	0.7%	3.8%	7.8%	7.5%	11.7%	17.1%	17.6%
L2 Students	0.0%	0.0%	1.5%	0.0%	0.6%	1.2%	3.1%	3.1%	5.1%
Cohort 2	0.2%	0.0%	0.6%	0.7%	0.5%	2.7%	4.5%	4.6%	8.5%
L1 Students	0.5%	0.0%	1.4%	1.3%	1.0%	4.0%	6.9%	8.3%	12.7%
L2 Students	0.0%	0.0%	0.0%	0.3%	0.0%	1.6%	2.7%	1.6%	5.3%
Comparison	0.0%	0.0%	0.0%	0.7%	0.5%	0.9%	3.2%	3.4%	5.8%
L1 Students	0.0%	0.1%	0.0%	0.8%	0.4%	1.7%	9.5%	5.2%	10.9%
L2 Students	0.0%	0.0%	0.0%	0.6%	0.5%	0.4%	1.1%	2.9%	3.1%
GIRLS	Baseline	Midline	Endline	Baseline	Midline	Endline	Baseline	Midline	Endline
Treatment	0.1%	0.2%	0.3%	0.9%	2.5%	3.4%	6.4%	9.2%	9.7%
Cohort 1	0.3%	0.5%	0.4%	1.9%	3.6%	3.9%	4.1%	11.0%	11.2%
L1 Students	1.0%	1.8%	1.4%	6.5%	11.2%	10.5%	10.6%	28.9%	17.3%
L2 Students	0.0%	0.0%	0.0%	0.0%	0.6%	0.8%	1.3%	2.5%	8.7%
Cohort 2	0.0%	0.0%	0.3%	0.4%	1.8%	3.2%	7.8%	8.2%	8.7%
L1 Students	0.0%	0.0%	0.7%	0.8%	4.0%	5.7%	14.2%	14.5%	15.0%
L2 Students	0.0%	0.0%	0.0%	0.0%	0.0%	0.7%	3.1%	2.7%	3.4%
Comparison	0.1%	0.0%	0.0%	0.9%	0.1%	1.1%	6.7%	3.0%	3.4%
L1 Students	0.4%	0.0%	0.1%	3.4%	0.3%	1.3%	18.8%	7.0%	6.2%
L2 Students	0.0%	0.0%	0.0%	0.3%	0.0%	1.0%	2.3%	1.6%	2.1%

Note: Sample weights applied to recover population representativeness

ANNEX V: SAMPLE

Before NORC was asked to conduct the IE of NEGRP, 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 a 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 NEGRP model, not impact.

RESEARCH QUESTIONS

- NEGRP improved the reading outcomes of pupils who speak Nepali as a first language (L1 learners) in cohorts 1 & 2
- NEGRP 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 NEGRP.

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 intracluster 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 NEGRP

	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	

	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 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 A5.2: Sample Design NEGRP*** 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 NEGRP

	Percentage of Nepali Speakers in School	Pupils (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 will 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 1 (created by NORC to complement RTI treatment sample design)

Given that all schools in treatment districts will receive NEGRP, 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 1 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 NEGRP 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 NEGRP- 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 A5.3: 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

	% of Nepali speakers in school	Total pupils (grade 1 3)	% L1	% L2	Proportion of pupils	Schools to be sampled proportionally	Adjustment	Schools to be sampled	L1 Pupils Sampled	L2 Pupils Sampled
Cohort 1	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%

ANNEX VI: CONSTRUCTION OF INDEXES

A. Teacher Reading Instruction Practices Indexes

We created two indexes to measure teachers' reading instruction practices in the classroom. The first index –Index I- 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?

Practices Index II- using calculation guidelines from USAID. This index includes a subset of questions used in Index I, but requires specific combinations of teaching practices that reflect categories such as phonemic awareness instruction, fluency modeling, reading comprehension exercises, etc.. Index II takes values ranging from 0 to 13, giving one point for each of 13 practices, calculated as follows:

Teaching Reading Instruction Practices Index II Calculation

Requirements	Category
Did teacher show how to pronounce sounds/letters/words/syllables correctly? AND Did students pronounce sounds/letters/words correctly?	Phonemic Awareness
Did students practice reading/pronouncing sounds/letters/words separating? OR Did students practice reading/pronouncing sounds/letters/words put together?	Graph phonemic awareness
Did teacher read text with proper sound/pattern/rhythm for students to listen?	Fluency modeling
Did students have opportunity to read alone/in pairs with proper sound/pattern/rhythm?	Students read aloud
Did teacher introduce new vocab words or discuss meaning of vocab words? OR Did teacher ask students to use vocab words in sentence/activity oral/writing?	Vocabulary
Did teach have students answer questions before/while reading/listening to text? OR Did teach ask students questions about read/listening text after text finished?	Reading Comprehension
Did students have opportunity to practice writing accuracy? OR Did students have opportunity to do any original writing?	Writing
Overall, did the teacher call on all students in the classroom? AND Overall, did the teacher call on, and respond to, boys and girls equally?	Equity
Did the teacher use at least two different kinds of grouping?	Grouping
During lesson, were most of students primarily doing what teacher asked? OR During lesson, did more than half of children volunteer to answer questions? OR If children reading, are majority of children eyes on text as they read?	Student Participation
If student responded correctly, did teacher gave them positive feedback? OR If student responded incorrectly, did teacher gave constructive feedback?	Feedback
During lesson, did teach move around to monitor students work individually/in groups? OR Did see examples of teach modeling, guiding & letting students practice (I DO, WE DO, YOU DO)? OR During lesson, did teacher do (in)formal check of students understanding/performance?	Monitoring
Are there posters / charts / pictures or paintings on the wall? OR Is student work displayed on the walls?	Print-Rich Environment

B. School Management Index

USAID/Nepal, the EGRP team and local stakeholders defined a School Leadership and Management Index. The index includes 14 items related to the school priorities, actions devoted to promote reading, parental involvement, student reading performance monitoring, etc. We provide the complete list of items below:

- 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.

ANNEX VI: DISCLOSURE OF CONFLICT OF INTEREST FOR USAID EVALUATION TEAM MEMBERS

Name	Alicia Menendez
Title	Principal Investigator – Evaluator Specialist
Organization	NORC at the University of Chicago
Evaluation Position?	X Team Leader ☐Team member
Evaluation Award Number (contract or other instrument)	GS-10F-0033M/AID-OAA-M-13-00013
USAID Project(s) Evaluated *(Include project name(s), implementer name(s) and award number(s), if applicable)*	Reading and Access – NEPAL Early Grade Reading Program Impact Evaluation
I have real or potential conflicts of interest to disclose.	No
If yes answered above, I disclose the following facts: *Real or potential conflicts of interest may include, but are not limited to:* *1. Close family member who is an employee of the USAID operating unit managing the project(s) being evaluated or the implementing organization(s) whose project(s) are being evaluated.* *2. Financial interest that is direct, or is significant though indirect, in the implementing organization(s) whose projects are being evaluated or in the outcome of the evaluation.* *3. Current or previous direct or significant though indirect experience with the project(s) being evaluated, including involvement in the project design or previous iterations of the project.* *4. Current or previous work experience or seeking employment with the USAID operating unit managing the evaluation or the implementing organization(s) whose project(s) are being evaluated.* *5. Current or previous work experience with an organization that may be seen as an industry competitor with the implementing organization(s) whose project(s) are being evaluated.* *6. Preconceived ideas toward individuals, groups, organizations, or objectives of the particular projects and organizations being evaluated that could bias the evaluation.*	

I certify (1) that I have completed this disclosure form fully and to the best of my ability and (2) that I will update this disclosure form promptly if relevant circumstances change. If I gain access to proprietary information of other companies, then I agree to protect their information from unauthorized use or disclosure for as long as it remains proprietary and refrain from using the information for any purpose other than that for which it was furnished.

Signature	
Date	June 23, 2020

Name	**Gregory Haugan**
Title	Principal Research Analyst
Organization	NORC at the University of Chicago
Evaluation Position?	□ eam Leader ☒ Team member
Evaluation Award Number *(contract or other instrument)*	AID-OAA-M-13-00010
USAID Project(s) Evaluated *(Include project name(s), implementer name(s) and award number(s), if applicable)*	USAID/NEPAL Impact Evaluation of the Early Grade Reading Program (EGRP) in Nepal
I have real or potential conflicts of interest to disclose.	□ Yes ☒ No
If yes answered above, I disclose the following facts: *Real or potential conflicts of interest may include, but are not limited to:* 1. *Close family member who is an employee of the USAID operating unit managing the project(s) being evaluated or the implementing organization(s) whose project(s) are being evaluated.* 2. *Financial interest that is direct, or is significant though indirect, in the implementing organization(s) whose projects are being evaluated or in the outcome of the evaluation.* 3. *Current or previous direct or significant though indirect experience with the project(s) being evaluated, including involvement in the project design or previous iterations of the project.* 4. *Current or previous work experience or seeking employment with the USAID operating unit managing the evaluation or the implementing organization(s) whose project(s) are being evaluated.* 5. *Current or previous work experience with an organization that may be seen as an industry competitor with the implementing organization(s) whose project(s) are being evaluated.* 6. *Preconceived ideas toward individuals, groups, organizations, or objectives of the particular projects and organizations being evaluated that could bias the evaluation.*	

I certify (1) that I have completed this disclosure form fully and to the best of my ability and (2) that I will update this disclosure form promptly if relevant circumstances change. If I gain access to proprietary information of other companies, then I agree to protect their information from unauthorized use or disclosure for as long as it remains proprietary and refrain from using the information for any purpose other than that for which it was furnished.

Signature	X Gregory L. Haugan
Date	07/17/2020

U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT

1300 Pennsylvania Avenue, NW
Washington, DC 20523