USAID READING FOR ETHIOPIA'S ACHIEVEMENT DEVELOPED MONITORING AND EVALUATION (READ M&E)

EARLY GRADE READING ASSESSMENT (EGRA) 2018 ENDLINE REPORT

September 2018
This publication was produced for review by the United States Agency for International Development. It was prepared by American Institutes for Research.

USAID READING FOR ETHIOPIA'S ACHIEVEMENT DEVELOPED MONITORING AND EVALUATION (READ M&E)

Early Grade Reading Assessment (EGRA) 2018 Endline Report

October 2018

Contract No. AID-663-C-15-00001

This study is made possible by the generous support of the American people through the United States Agency for International Development (USAID). The contents are the responsibility of the American Institutes for Research and do not necessarily reflect the views of USAID or the United States Government.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS LIST

EGRA	Early Grade Reading Assessment
FWS	Familiar Words Reading
ILS	Initial Letter Sound
IQPEP	Improving Quality in Primary Education Program
IWR	Invented Words Reading
LNR	Letter Name Recognition
MoE	Ministry of Education (Ethiopia)
ORF	Oral Reading Fluency
READ M&E	Reading for Ethiopia's Achievement Developed Monitoring and Evaluation
RSEB	Regional State Education Bureau
RTI	RTI International
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

1. INTRODUCTION

The Reading for Ethiopia's Achievement Developed Monitoring and Evaluation (READ M&E) project, supported by the United States Agency for International Development, is designed to monitor and evaluate the three READ implementation programs. As part of this mandate, READ M&E carried out the endline administration of the Early Grade Reading Assessment (EGRA) in June 2018 to evaluate the reading fluency and comprehension levels of students in grades 2 and 3 in seven languages within five regions of Ethiopia: Afaan Oromo, Aff Somali, Amharic, Haddiysa, Sidamu Affo, Wolayttatto, and Tigrigna. EGRA's primary purpose is to mediate important policy and pedagogical discussions about where, when, and how to improve the level of reading among early grade students in Ethiopia.

The EGRA tool developed in Ethiopia contains four timed subtasks (letter name recognition, familiar words reading, invented words reading, and oral reading fluency) and three untimed subtasks (phonemic awareness, reading comprehension, and listening comprehension).

In June 2018, READ M&E collected data from 459 schools and assessed 17,879 students in grades 2 and 3with the purpose of serving multiple projects' evaluation needs. The results presented in this report are based on the 355 schools that were sampled for the 2018 EGRA endline administration, as a follow-up to the 2014 baseline and 2016 midline administrations. This report provides extensive information about the study design, data collection procedure, methods of analysis, and 2018 EGRA results, along with the comparative summaries from the previous two EGRA administrations conducted in 2014 and 2016. The report concludes with a discussion of the study's main findings and recommendations for policy makers.

2. 2018 EGRA DESIGN

Development of Pilot Forms

Considering that the 2018 endline EGRA targeted the same schools that received the 2016 midline administration, the READ M&E team enhanced test security by developing and piloting two new forms of four subtasks that used printed stimuli: Familiar Words Reading, Invented Words Reading, Oral Reading Fluency, and Reading Comprehension.

The development of new EGRA forms was a joint collaboration effort between READ M&E staff, MOE staff, and local mother tongue language experts who had worked on the development of the 2016 EGRA tool. These experts developed stimulus material for two new EGRA forms (A and B) for each of the four selected subtasks.

Piloting

READ M&E used Nexus 7 Tablets loaded with Tangerine software to complete both the pilot data collection in November 2017 and operational data collection in April - May 2018. Both administrations were carried out by trained EGRA assessors who were selected from colleges of teacher education, universities, the MoE, RSEBs, the National Education Assessment and Examinations Agency, zone education departments, and preparatory schools.

The pilot served a dual purpose: to evaluate the quality of the newly developed stimulus material and to ensure comparability between the 2016 midline and 2018 endline administrations by using a common-persons equating design. Based on the results of pilot data analysis, READ M&E selected one of the two piloted forms to be used for the operational administration. The pilot data analysis also enabled computation of the equating relationship to establish comparability between newly developed forms and the forms used in the 2016 EGRA.

Sample

To enhance accuracy of comparisons between midline and endline administrations, READ M&E decided that, to the degree possible, the same schools assessed in the 2016 EGRA should be also selected for the 2018 EGRA administration. READ M&E assessed a total of 198 schools in both the 2016 midline and 2018 endline administrations. These schools constitute 56% of the 2018 sample of 355 schools.

In both the 2016 EGRA and the 2018 EGRA, READ M&E selected zones according to their accessibility, security level, and the extent to which they were affected by floods and droughts. Consequently, a fully stratified random sampling procedure was not possible. Within the selected zones, the schools were sampled randomly to the degree possible (50 schools per language). At each school, 20 students were selected from grade 2 and 20 students were selected from grade 3. An equal number of girls and boys were randomly selected from each grade.

Operational Data Collection

To ensure high quality data, READ M&E conducted an intensive enumerator training prior to deployment to schools. The 2018 EGRA enumerator training and data collection were conducted in two phases, organized by language groups. The first group (Amharic, Afan Oromo, and Sidamu Afoo) was trained on April 25-27 and data were collected between April 30 and May 11. The second group (Tigrigna, Wolayttatto, Hadiyissa, and Aff Somali) was trained May 16-18 and data were collected May 21 through June 1.

During the 2018 operational data collection efforts, each team of test administrators consisted of four enumerators and one supervisor. The supervisor’s primary responsibility was to oversee data collection implementation, monitor enumerators’ performance, and conduct principal and teacher interviews. Data collection was also supervised by READ M&E and USAID representatives.

Data Analysis

The scores for timed tasks were calculated as the number of letters or words correctly read per minute, whereas the scores for untimed tasks were calculated as the percentage of correct responses out of the total number of questions in the subtask.

The READ M&E team analyzed EGRA data and presented the results using two major reporting frameworks:

1. Subtask scores: Correct words (letters) per minute for timed tasks and percentage of correct answers for untimed tasks.

2. Benchmark levels: Percentage of students reaching each EGRA benchmark level.

Through data analysis, READ M&E seeks to address the following research questions:

- What is the overall reading performance of Ethiopian students in early grades, and is it aligned with expectations? What is the percentage of students who attained the desired benchmark levels? What is the percentage of non-readers?

- What is the difference between student performance in reading comprehension and listening comprehension subtasks?

- What is the difference between reading performance at different grade levels?

- Is there a difference in EGRA performance between boys and girls?

- What are the differences among baseline, midline, and endline results and do they follow a systematic trend?

- How is reading performance associated with background factors assessed through directors', teachers', and students' questionnaires?

All the analyses were carried out with the sampling weights based on the school size, which is determined by the number of students enrolled in corresponding grades.

3. RESULTS OF THE 2018 EGRA ENDLINE

Overall Reading Performance on the 2018 EGRA

The 2018 EGRA results show that 6.2% of Ethiopian students across all languages achieved the targeted reading benchmark—*reading fluently with full or almost full comprehension*. This benchmark ranged from 11.5% among students in Amharic to 2.0% among students in Tigrigna. Given that the next benchmark level (*reading with increasing fluency and comprehension*) reflects relatively functional reading proficiency, these two benchmark levels can be combined for monitoring and evaluation analysis. Thus, looking at the top two benchmark levels combined, the overall percentage of Ethiopian students who exhibit relatively functional reading proficiency is 32.4%. This benchmark ranges from 50% among students in Amharic to 16.0% among students in Haddiysa.

Comparisons Between Reading Comprehension and Listening Comprehension

Students performed much higher on listening comprehension than on reading comprehension in all seven languages. The overall percent-correct average for reading comprehension questions is 20%, whereas for listening comprehension it is 69%. This difference is a result of the different amount of information that the students received through these two perception modes. In other words, children's inability to read a given text restricts the amount of acquired information; as a result, there is less information for the children to process and comprehend. In contrast, in the listening mode, children receive the entire information contained in the given passage.

Grade-level Comparisons

In each language, the mean differences between the two grade levels in all EGRA subtasks were statistically and practically significant in favor of grade 3. This means that in all languages, grade 3 students were able to read substantially better than students in grade 2, which indicates positive gain across grades. The average size of grade differences for Oral Reading Fluency (ORF) across all languages as measured by Cohen's d[1] is 0.55, which corresponds to the average gain of 18 words per minute. For reading comprehension, the grade difference size by Cohen's d was 0.51, corresponding to the average gain of 12 percent-correct points. These differences fall deep into the category of educationally significant effect sizes, indicating that substantial reading gains are occurring between grades 2 and 3.

Gender Comparisons

EGRA results from the 2018 administration illustrate gender inequality in Ethiopia. Boys performed significantly higher than girls in 6 languages (with an average Cohen's d of 0.29, which reflects a significant educational effect). Only in Amharic did girls outperform boys with negligible size of difference (0.12).

Relationship Between ORF and Reading Comprehension Scores

In each language, a strong relationship exists between ORF and reading comprehension scores (correlation coefficients over 0.80), which, again, is attributable to the amount of information received. The more words a student can read, the more information is available to process and comprehend. This emphasizes the importance of ORF as a condition for learning.

4. COMPARISON OF EGRA PERFORMANCE ACROSS YEARS[2]

There has been little change in overall reading performance at the aggregated national level across the three EGRA administrations. The EGRA test results do not provide evidence that could be considered satisfactory in the context of the country's ongoing developmental activities. The percentage of students performing at the upper two benchmark levels was 31.3% in 2014. This percentage increased slightly to 34.3% in 2016 and then slipped back to 32.4% in 2018, with the differences being too small to be considered practically significant.

[1] Cohen's d is widely used as a measure of effect size. It expresses the size of the observed difference as a fraction of the pooled standard deviation. Cohen (1977) defined effect sizes as "small, *D* = 0.2," "medium, *D* = 0.5," and "large, *D* = 0.8." Wolf (1986) offered an educationally referenced interpretation: 0.25 = educationally significant (something was learned), 0.50 = practically/clinically significant (something really changed).

[2] Because reliable data could not be obtained for Aff Somali in 2014 and Wolayttatto in 2016, these observational points are not included in comparisons and computations of aggregated results.

While there were small changes in student reading performance over time at the national level, there were mixed results at the individual language levels. The 2016 and 2018 ORF scores demonstrated a practically significant gain in Aff Somali and a substantial drop in Sidamu Affo. In other languages, the differences between the 2016 EGRA and the 2018 EGRA were relatively small, with negligible or marginal practical significance. The comparison between 2014 and 2016 ORF scores reveals substantial gains in Sidamu Affo and Amharic. Relatively negligible differences were observed in the other languages.

A similar pattern was observed in reading comprehension scores. The 2016 and 2018 reading comprehension scores show a practically significant gain in Haddiysa and a substantial drop in Sidamu Affo, whereas negligible differences were observed in the other languages. Changes in reading comprehension scores between 2014 and 2016 showed a substantial gain in Amharic and a notable drop in Haddiysa; the other differences were practically insignificant.

5. FACTORS ASSOCIATED WITH STUDENT PERFORMANCE

READ M&E evaluated the association between student reading performance on the ORF subtask and background factors, which were determined based on responses to directors', teachers', and students' questionnaires. The analyses of background factors revealed a large amount of pedagogically relevant and actionable information, which is summarized in this section, but extensively presented in Chapter 5, and discussed in reference to policy deliberations in Chapter 6.

Directors' Questionnaire

The directors' questionnaire revealed factors that significantly influence student reading performance in their schools, which include the existence of a person responsible for reviewing the mother tongue lesson plans, the frequency of the review of the mother tongue lesson plans, and the existence of a person responsible for observing teachers teach the mother tongue. We also used the directors' questionnaire to assess the availability of school resources for reading and the results yielded significant associations of certain school resources with student reading performance. For example, the availability of student textbooks and teachers' guides for mother tongue teachers, the number of mother tongue teachers at the school, teachers' educational qualifications, the availability of supplementary reading materials, and grade 2 and grade 3 students making use of the school library all had a positive association with an increase in ORF scores.

Teachers' Questionnaire

Data collected through teacher questionnaires showed that certain characteristics and behavior patterns were significantly related to their students' reading performance. For example, teacher gender; length, frequency and duration of training received; years of service; using student textbooks and the teacher guide every time they teach; providing remedial classes for students who are lagging; discussing with parents when a student is lagging; and using different methods to monitor student's reading progress. READ M&E developed policy recommendations stemming from these findings, which are presented in Chapter 6.

Students' Questionnaire

Multiple student background variables were significantly associated with their ORF performance, such as: gender, availability of a mother tongue textbook, bringing the mother tongue textbook to class every day, reading books in languages other than the mother tongue, and borrowing supplementary reading materials. Significant home resources include having books at home, having literate family members, receiving help with reading, and having enough time to read at home. Students who reported that their schools have certain resources, such as having a school library and presence of a reading corner at the school, also showed higher performance in ORF.

6. DISCUSSION AND POLICY CONSIDERATIONS

Considering that the goal of reading is to gain information and construct meaning, the findings of the 2018 EGRA show that the number of students able to respond correctly to the reading comprehension subtask questions is disconcertingly low, at 32% overall. A high percentage of Ethiopian students cannot read enough words within one minute to develop an understanding of what they read. This is not an issue of students not understanding the language, as scores are higher on the listening comprehension subtask. We can conclude that students' lack of success in reading comprehension is related more to the lack of appropriate decoding skills than to their inability to comprehend the information. Therefore, students are not sufficiently mastering preliminary decoding and other tasks to acquire and comprehend enough information in the desired time. Where must reading interventions focus on to turn this pattern around?

The methodological limitations that existed between baseline and midline EGRA (traditional vs. tablet-based administration) were not an issue when comparing endline with midline EGRA performance. The 2018 EGRA did not demonstrate that reading scores improved over time, as it could be expected based on the ongoing reading interventions in the past five years. On the other hand, relatively large gains in student performance between grades 2 and 3 suggest that these increases should be interpreted as the strong effects of reading instruction.

The 2018 EGRA demonstrated a higher reading performance among boys in all languages except Amharic, which reveals that gender inequality is still a prevalent issue in Ethiopia. Policymakers need to address this issue, as do civil society members, teachers, and parents.

Finally, the 2018 EGRA included a comprehensive and valuable analysis of factors associated with reading performance. The analysis provides important actionable information that policymakers, administrators, teacher educators and teachers can and must use to improve delivery of reading instruction with support from the broader community.

On the way forward, READ M&E proposes three major lines of action:

1. Increase the breadth and depth of current interventions. The analysis of background factors revealed many associations that could be positively affected by the reading intervention currently in progress. These activities primarily focus on two major resources—human (teacher training at both pre-service and in-service levels) and materials (supplies for teaching reading, books, libraries, reading rooms, etc.).

2. Develop and implement the instruments and procedures for evaluating two major intervention outcomes: *teacher competencies* and *school competencies.* These two tools can provide valuable insights into the strengths and weaknesses of the primary recipients of intervention (teachers and schools) and inform adjustments of the intervention where necessary.

3. Conceptualize, develop, and implement a formative assessment system that empowers teachers to monitor and promote student reading proficiency in early grades. This comprehensive system should include both classroom-based assessments embedded in instruction and periodic EGRA-like assessments that teachers can use for monitoring and promoting their students' reading performance.

All three of the strategies mentioned above should be thoroughly discussed and customized to the diverse needs of regions and unique characteristics of the MT languages in Ethiopia.

1. INTRODUCTION

Since 2007, the United States Agency for International Development (USAID) has supported the development and administration of an Early Grade Reading Assessment (EGRA) in over 120 languages in more than 70 countries throughout the world. The purpose of the EGRA is to help USAID partner countries systematically measure how well primary school children are acquiring key reading skills. The EGRA measures the prereading and basic reading skills a child must acquire to read fluently and with comprehension—the ultimate goal of reading.

The design of the EGRA measurement tool is based on reading research regarding the types of skills that are necessary for reading acquisition. The EGRA tool developed in Ethiopia has four timed subtasks (letter name recognition, familiar words reading, invented words reading, and passage reading) and three untimed subtasks (phonemic awareness, reading comprehension, and listening comprehension). A single test administrator gives the test orally to an individual student. The 2018 EGRA measured the emerging reading skills of approximately 18,000 children in seven mother tongues in five regions of Ethiopia.

EGRA results are intended to help education policymakers, administrators, teacher educators, teachers, parents, and donors in establishing priorities to improve interventions that will in turn increase foundational reading skills. However, the EGRA should not be used as a high-stakes accountability tool for evaluation of children and teachers, nor for direct, cross-language comparisons of reading achievement. As administered in Ethiopia, EGRA is also not suitable for providing information about an individual child's progress toward learning to read; rather, it is an overall measure of the early reading performances in the education system in the regions in which the assessment is administered.

This report is the endline in the series following the 2016 midline and 2014 baseline reports. It presents findings from the systematic investigation of early grade reading outcomes in Ethiopia in 2018 conducted by USAID's Reading for Ethiopia's Achievement Developed Monitoring and Evaluation (READ M&E). This report provides extensive information about the study design, data collection procedure, methods of analysis, and 2018 EGRA results, along with the comparative summaries from the previous two EGRA administrations conducted in 2014 and 2016. The report concludes with a discussion of the study's main findings and recommendations for policy makers.

1.1 HISTORY AND PURPOSE OF EGRA IN ETHIOPIA

In May and June 2010, RTI International (RTI), Improving Quality in Primary Education Program (IQPEP), and the Ethiopia Ministry of Education (MoE) collaboratively carried out the first administration of EGRA in Ethiopia. It was conducted in eight regions, in six languages: Tigrigna, Afaan Oromo, Amharic, Aff Somali, Sidamu Affo, and Hararigna. Approximately 90% of the population speaks at least one of these languages. After receiving the results of the 2010 EGRA, USAID, the MoE, and other development partners that support education in Ethiopia agreed to focus on improving early grade reading and writing. Thus, the current READ programs funded by USAID seek to improve the quality of mother tongue reading and writing education for children in the early grades, with the purpose of enabling greater learning in upper grades.

IQPEP, in collaboration with the MoE and regional state education bureaus (RSEBs), conducted a second EGRA in May 2013 to formatively assess the impact of the interventions on students' reading abilities. The findings showed some improvements in both reading fluency and comprehension when compared with the 2010 EGRA.

In May 2014, IQPEP conducted the EGRA in a representative sample of the 2,615 IQPEP-supported schools. The results for the final EGRA under the IQPEP program showed that students were making progress in acquiring prereading skills in Ethiopia, though the progress was slow. There was also significant variation depending on the language and region. In June 2014, RTI conducted a baseline EGRA for the Haddiysa and Wolayttatto languages. The results revealed that some Haddiysa- and Wolayttatto-speaking students were only beginning to learn to read in their respective language by grade 3. Table 1 shows the history of the EGRA in Ethiopia.

Table 1. EGRA In Ethiopia

Year	Conducted by	Languages	Sample Size	Data Collection Period
2010	RTI, IQPEP and MoE	6 (Amharic, Afaan Oromo, Tigrigna, Sidamu Affo, Hararigna, Aff Somali)	8 regions, 90 woredas, 338 schools, 13,079 grades 2 and 3 students	May 10, 2010–June 16, 2010
2013	FHI 360/ IQPEP	5 (Amharic, Afaan Oromo, Tigrigna, Sidamu Affo, Aff Somali)	8 regions, 53 woredas, 240 (120 control) schools, 9406 (4,699 control) students	May 2013
2014	FHI 360/ IQPEP	6 (Amharic, Afaan Oromo, Tigrigna, Sidamu Affo, Hararigna, Aff Somali)	8 regions, 53 WEO, 240 (120 control) schools, 9,406 (4699 control) students	May 2014
2014	RTI	2 (Haddiysa and Wolayttatto)	2 zones (Hadiya and Wolayta) 2,000 students	June 2014
2016	AIR/READ M&E	7 (Amharic, Afaan Oromo, Aff Somali, Tigrigna, Sidamu Affo, Haddiysa, and Wolayttatto)	5 regions, 7 languages, 13,475 grades 2 and 3 students	May 23, 2010–June 12, 2016
2018	AIR/READ M&E	7 (Amharic, Afaan Oromo, Aff Somali, Tigrigna, Sidamu Affo, Haddiysa, and Wolayttatto)	5 regions, 7 languages, 17,879 grades 2 and 3 students from 459 schools	June 2018

Notes. RTI is RTI International; IQPEP is Improving Quality in Primary Education Program; MoE is Ministry of Education; AIR is American Institutes for Research; READ M&E is Reading for Ethiopia's Achievement Developed Monitoring and Evaluation.

1.2 EGRA LIMITATIONS

The EGRA in Ethiopia is a set of subtasks that measure foundational skills identified in the research as predictive of later reading success. It is not intended to be an accountability measure that determines student grade promotion or evaluates individual teachers. Instead, EGRA is designed to complement, rather than replace, existing curriculum-based, pencil-and-paper assessments. However, because of the constraints imposed by children's limited attention span and stamina, neither EGRA nor any other single instrument can measure all skills required for students to read with comprehension. EGRA is not intended to be an instructional program but, rather, to inform policy makers and many other stakeholders of aggregate progress in a large sample of students with respect to reading outcomes. As such, it can also mask progress in subsets or in particular schools. EGRA cannot fully determine background or literacy behaviors that could influence a student's ability to read.

1.3 SEVEN SUBTASKS OF THE EGRA

EGRA measures skills in: phonological awareness, decoding, reading fluency, reading comprehension, and listening comprehension. Each of the seven EGRA component is described below.

The ***letter name recognition*** subtask assesses students' knowledge of the alphabetic principle, the foundation of learning to read. The alphabetic principle is the understanding that words are composed of letters (i.e., graphemes) that represent sounds. When children understand that sounds correspond to letters, they can begin to learn to decode words (McBride-Chang & Kail, 2002; 2004; McBride-Chang & Ho, 2000).

Research in other languages has suggested that comprehension can occur only after 80% of letters (fidels) are mastered (Seymour et al., 2003). EGRA measures the ability to read and name the letters of the alphabet naturally and without hesitation. This timed test (1 minute) assesses automaticity and fluency in recognizing letter names. It contains 100 randomly arranged letters in both lower- and uppercase form.

The ***initial letter sound*** subtask is an assessment of students' phonological awareness skill. A phoneme is the smallest, linguistically distinctive unit of sound allowing for differentiation of two words in a language. The 2000 National Reading Panel meta-analysis of the literacy research (conducted only on literacy in the English language) determined that skills in phoneme identification and phonological awareness are strongly associated with good reading comprehension. Phonemic awareness is the foundation for learning phonological awareness, a domain that includes skills in hearing and manipulating onsets, rhymes, and syllables (Snow et al., 1998; NIHCD, 2006).

For the initial letter sound subtask, the student stimulus includes a list of the 10 randomly arranged words that begin with frequently used letters. The frequency of letters in everyday use is determined during development of the subtask by text analysis and calculations of word count frequencies. The administrator reads each word two times and then asks the students to make the first sound of the word. If a student does not answer within 3 seconds, a "no answer" response is recorded. The maximum score for this section is 10 points, with 1 point assigned for each correct answer.

The ***familiar words reading*** subtask assesses the student's ability to recognize and read high-frequency words. Frequency of words is determined through a word count analysis of the most commonly used words in textbooks of appropriate level. The list of words is derived from the fifty most frequently used words in the grade 2 textbooks. For this task, EGRA assessors can attain a measure of decontextualized decoding skill that is a distinct skill from reading comprehension from text (Gove, 2009). Unlike *Oral Reading Fluency*, this subtask presents a list of unrelated words that are not presented as a story or complete text: Fifty words are randomly arranged in the student stimulus sheet. The familiar words reading task is scored on a *words-per-minute* calculation that calls for the administrator to determine how many words the student attempts, how many the student reads correctly, and in what time over the course of 60 seconds.

The ***invented words reading*** subtask assesses the ability of students to decode one- and two-syllable non-words that could plausibly exist in the language in question. The invented words reading subtask provides a measure of decoding related to that of the familiar words reading task but has the advantage of not allowing respondents to *sight-read* words. To achieve fluency in reading, students need to acquire both sight-reading and decoding skills. According to Hirsch (2003), there is significant evidence that overreliance on sight-word vocabulary often leads to regression in reading development by age 9 or 10.

Fifty non-words are randomly arranged in a list in the student booklets, and students are asked to read as many of the non-words as they can. The invented words reading task is graded on a *words-per-minute* calculation that calls for the administrator to determine how many words a student attempts, how many are read correctly, and the amount of time in which they were read on this 1-minute subtask.

The ***oral reading fluency*** (ORF) subtask assesses the ability of a student to read passage texts with speed, accuracy, and proper expression. The purpose of the timed ORF subtask is to examine whether pupils in grades 2 and 3 can read a passage with speed and accuracy with grade-appropriate words (familiar words) as presented in the pupil booklets. The *Oral Reading Fluency* task is "oral" in that pupils read the passage aloud. Oral reading is assessed because empirical studies in many contexts have demonstrated that there is a strong correlation between oral fluency and reading comprehension (Fuchs, Fuchs, Hosp, & Jenkins, 2001). Although *Oral Reading Fluency* is considered a precursor to reading comprehension, and is just statistically correlated with reading comprehension, it represents an important foundational skill.

In the 2016 and 2018 EGRAs, the ORF subtask included paragraphs with about 60 words. In subtask design, test developers conducted textbook reviews to determine which words could be considered grade appropriate. The stories created were appropriate for specific regions and targeted at grade 2. The subtask was scored on a *words-per-minute* calculation that called for the administrator to determine how many words were attempted and how many were read correctly in 1 minute.

The **reading comprehension** subtask, which relies on questions about the text read in the ORF subtask, determines students' understanding of the text and their ability to answer factual questions and make inferences based on what they read. After a student completes the ORF subtask, the administrator moves to the reading comprehension subtask, which includes a series of questions about the passage that the student had just read.

Research indicates that a student's ability to correctly understand and interpret oral stimuli (linguistic comprehension) and make meaning from what he/she hears is a core skill related to reading comprehension (Hoover & Gough, 1986; Kamhi & Catts, 1991). In this EGRA subtask, the child demonstrates ***listening comprehension*** skill by answering several questions from a simple oral story (series of sentences) read aloud by the administrator (an interactive situation). According to O'Maggio (1986), the core dimensions of listening include retaining parts of language stored in short-term memory, discriminating among distinctive sounds, detecting key ideas, and guessing meaning from context.

The listening comprehension subtask includes a paragraph of approximately 40 words. The test administrator reads the passage aloud only once at a pace of about one word per second. When the administrator completes reading of the text aloud, he or she then asks students five comprehension questions.

2. 2018 EGRA DESIGN

This section details how the 2018 EGRA tools were developed, how they were piloted, and which methods were used for equating, i.e., for making the 2018 and 2016 data comparable. Given that this was the second time in Ethiopia that the EGRA data were collected using tablets, this section presents substantial discussion on how this new method of data collection was performed. This section also covers the sample size, selection of schools, and geographical coverage, taking into account accessibility issues.

2.1 OBJECTIVES OF THE 2018 ETHIOPIAN EGRA

The 2018 EGRA measured the emerging reading skills of approximately 18,000 children in seven mother tongues in five regions of Ethiopia. The 2018 EGRA tool was equated to the 2016 EGRA tool to enable a comparison of the results from the two administrations (midline and endline). In Chapter 4, we discuss the comparative review of EGRA results across the 2014 baseline, 2016 midline, and 2018 endline administrations. The comparison between the 2014 EGRA results and the 2016 EGRA results was presented in the 2016 EGRA report, where we noted that comparison across those years had limitations, such as different EGRA administration modes (traditional paper vs. tablet based), different project ownership (IQPEP vs. READ M&E), and a different sampling plan.

We advise caution in comparing results across the seven languages as each language has its own unique span of time toward fluency. However, one can make comparisons using benchmarks specific to each language. USAID and MoE conducted a benchmarking workshop in 2015 when experts from all 7 languages met and established benchmarks that share common general meaning. Cut scores delineating benchmark levels were customized to specific characteristics of each language. Chapter 3 describes more about benchmark data.

Considering the slow implementation of the READ suite of interventions, this 2018 EGRA should not be taken as a measurement of impact, but it can provide valuable pedagogical information for making policy decisions. For this reason, the EGRA 2018 sample also included the schools that were a part of the READ Technical Assistance (TA) performance study, when multiple implementation indicators were assessed. This data will enable READ M&E to evaluate how implementation indicators might be related to the EGRA performance in corresponding schools. The implementation indicators provide highly relevant monitoring and evaluation data that can inform policy related decisions for improving reading interventions.

2.2 DEVELOPMENT AND PILOTING OF THE 2018 EGRA TOOL

Taking into consideration the security of testing materials used for administration of the 2016 EGRA (potential exposure risks), and the need to provide ample material for the construction of operational test forms used in the 2018 endline administration, the READ M&E team decided to develop and pilot two sets of new stimulus material for four selected subtasks that used printed stimuli:

- Familiar Words Reading (contains 50 stimuli; developed 100 new stimuli)

- Invented Words Reading (contains 50 stimuli; developed 100 new stimuli)
- Oral Reading Fluency (contains one passage; developed two new passages)
- Reading Comprehension (contains five questions; developed five new questions for each new passage)

For the subtasks using printed stimulus material, there is a possibility of uncontrolled distribution of these materials. However, for the subtasks not using printed stimulus materials (phonological awareness and listening comprehension), it may be reasoned to assume that the same testing material can be used because the security risk is low or minimal. Likewise, for the letter naming subtask, it can be reasoned that the same material is reusable because it contains a large number of stimuli (100) that cover the entire alphabet and that are not easily recallable.

2.2.1 DEVELOPMENT OF PILOT FORMS

To develop new stimulus material, AIR hired local mother tongue language experts, who had worked on the development of the 2016 EGRA tool. These experts developed stimulus material for two new EGRA forms (A and B) for each of the 4 selected subtasks. Language experts and the MoE reviewed each item of the new EGRA forms until they reached consensus for each language's tool. The participants in the workshop compared items with the 2016 midline tool (used as a reference form) and checked alignment with the newly developed mother tongue curriculum.

The workshop participants took into consideration the characteristics of each language. For example, Amharic and Tigrigna are written with symbols called *fidels*, which are represented as syllables (consonant and vowel) rather than as phonemes, as in alphabetic languages such as English. However, there is direct fidel-sound correspondence, and children must learn the fidels and their corresponding sounds to learn to read. Thus, it is important that the EGRA in Ethiopian languages test for phonemic awareness as well as syllabic awareness. Therefore, the revised 2018 EGRA measures phonemic awareness, syllabic awareness, letter naming fluency, familiar words reading fluency, invented words reading fluency, passage reading fluency, reading comprehension, and listening comprehension.

During the EGRA development and revision process, READ M&E and the workshop participants systematically reviewed the level of difficulty of each stimulus in the two new forms, position and distribution of the words within the test in both forms, and the nature of the comprehension passage in terms of number of words and grade-level suitability.

In addition to evaluating the quality of the newly developed stimulus material, the purpose of piloting was to enable comparative inferences between the 2016 midline and 2018 endline administrations. It was decided that the comparability would be established using a common-persons design (explained later in the text), which is considered the most suitable approach to equating the EGRA subtasks.

2.2.2 USING NEXUS 7 TABLETS AND TANGERINE SOFTWARE

READ M&E used Nexus 7 Tablets loaded with Tangerine software to complete the 2018 EGRA pilot data collection, as well as to carry out the operational data collection 6 months later. Tablets were programmed with the seven EGRAs in

November 2017, and the programming was slightly modified after the pilot administration. The modifications included fixing typographical errors, glitches in counting protocols, and issues with timing. These issues were satisfactorily resolved before the training of EGRA assessors in May 2018.

AIR and the MoE selected EGRA assessors from Colleges of Teacher Education, Universities, the MoE, Regional State Education Bureaus, National Education Assessment and Examinations Agency, Zone Education Departments, and preparatory schools. The minimum education qualification was a M.Ed. or MA.

The READ M&E team conducted a training workshop in November 2017. The purpose of the workshop was to provide the assessors with techniques on how to administer EGRA 2018 orally on a one-on-one basis using tablets. Since a large proportion of assessors had participated in EGRA 2016 data collection, not many assessors experienced problems with using tablet technology. Figure 1 shows an assessor using an EGRA tablet.

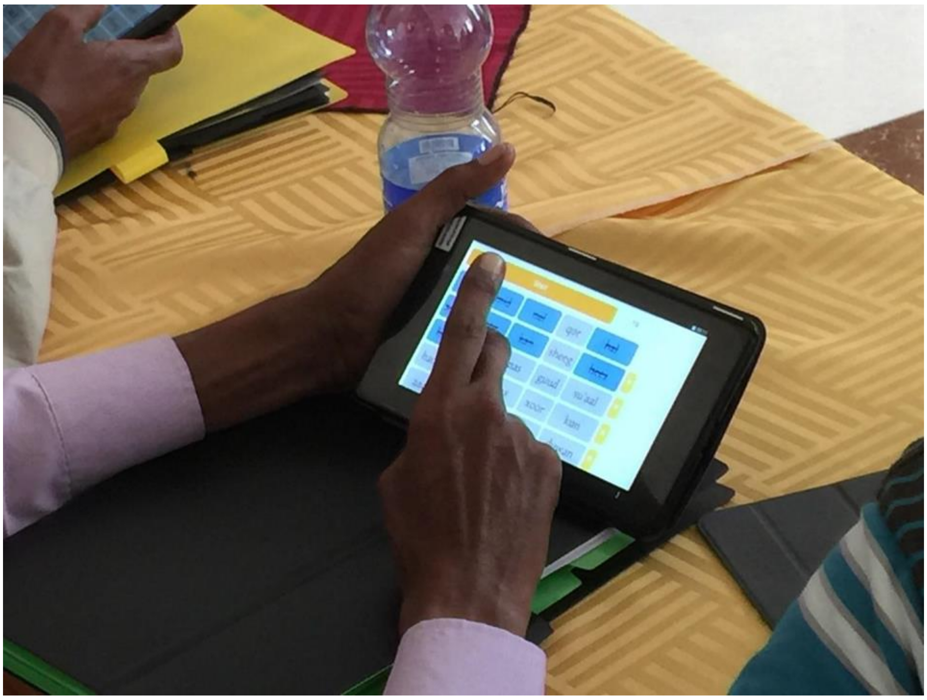


Figure 1. An assessor uses an EGRA tablet.

2.2.3 SUMMARY OF PILOT PROCEDURES

To equate 2016 EGRA and 2018 piloted forms through the common-persons design method, READ M&E sampled eight schools, selecting 40 pupils per school in each language (total 2,240 pupils). Twenty grade 2 students and 20 grade 3 students represented by an equal number of boys and girls were selected from the total grade 2 and grade 3 students in attendance. The selected children were given four piloted subtasks from the 2016 EGRA and from one of the piloted forms of the 2018 EGRA. To control for a possible effect of practicing, the two instruments were administered using the counterbalanced order. This method eliminated any bias from practicing and the child being familiar with the test. It is also important that the same assessor administered both forms to the same students to rule out possible influence of a person who was administering the test.

The detailed descriptions of the pilot data collection design, as well as the results of item analysis and equating analysis, are presented in Appendix 1A.

2.3 2018 EGRA ENDLINE ADMINISTRATION

As in the 2016 EGRA, the implementation of the 2018 EGRA in seven languages in five regions was a logistical challenge due to such issues as a large sample size, accessibility of schools, security, and the effects of floods and droughts. However, the READ M&E staff was experienced and trained for the necessary procedures and arrangements. In this section, we present a description of the sampling design, assessor training, administration of EGRA in schools, and quality control procedures. This section ends with a discussion of the administrative challenges that READ M&E encountered.

2.3.1 SAMPLING DESIGN RATIONALE FOR THE 2018 EGRA

We present here the rationale for the sampling design used for the 2018 EGRA endline administration. Because one of the major objectives of educational studies is to evaluate changes in individual or institutional proficiencies over time, educational researchers collect data that typically span a period of many years. Two basic sampling designs can be used to acquire data at multiple points over a longer period:

- ***Repeated cross-sectional design***, which gathers information using a different sample of observational units (persons or institutions) at each point across the study timeline.

- ***Longitudinal design***, which collects information from the same sample of observational units (persons or institutions) at predefined intervals over time.

Both data collection strategies have certain advantages and disadvantages, and selection of one or another approach should depend on study objectives. Based on review of ample body of literature, Almond & Sinharay (2012) concluded that, for answering questions about persons (or institutional) growth, a longitudinal study is preferable to repeated cross-sectional samples. Researchers have argued that repeated cross-sectional studies conflate several sources of variability (differences in the initial status of observational units, differences in the growth curves, and observational unit-by-measurement-occasion differences) in ways that are not easily separated.

In their elaborate review of the research on school effectiveness, Thomas, Salim, and Jung Peng (2013) state that datasets generated by the Southern and East African Consortium for Monitoring Education Quality (SACMEQ) in 2000–02 provided useful information on the influence of different student intake and school factors on student attainment outcomes. However, the authors point out that SACMEQ datasets are limited due to the cross-sectional nature of data collection design, which means that the progress at the individual or institutional level cannot be examined; nor can the key student, classroom, or school factors—which may explain differences in achievement progress—be evaluated. On the other hand, longitudinal datasets allow examination of value-added factors that assume that schools add value to their students' achievements. The authors provide examples from two low-income contexts (China and Zanzibar) that illustrate the need for longitudinal design and improved educational evaluation methods to inform and support school improvement initiatives.

It has also been reasoned that longitudinal studies provide information that enables understanding the patterns of changes over time. The cross-sectional designs can provide information only about the growth of averages, whereas longitudinal designs provide better understanding of patterns of individual/institutional changes and insights into contextual information associated with causes of these changes that can bear relevance for policy decisions. Furthermore, longitudinal designs are very effective in doing research on developmental trends. They are also more powerful than cross-sectional studies because they are based on the same observational units, which means that there is less random error involved in drawing inferences about changes over time.

In the context of the core differences between longitudinal and cross-sectional designs presented above, and considering READ M&E's mandate, it is certain that the longitudinal design will provide more useful information for serving the ultimate evaluation objectives of the project and provide the 'value added' information relevant for policy decisions and planning of improvement activities. Thus, the READ M&E technical team decided that the EGRA 2018 endline data collection will be conducted on the same sample of schools utilized for EGRA 2016 midline study (to the degree possible). In this case, schools are treated as sampling units and the design is considered as longitudinal because repeated observations are made on the same units at different points in time.

Another consideration in planning the sample for EGRA 2018 was the importance of analyzing associations between intervention fidelity and outcome indicators. Fidelity of intervention indicators were collected through the ***READ TA performance evaluation study*** against its implementation plan. We collected a lot of information that targeted the four intermediate results of the READ TA: 1) Were the reading and writing materials appropriate for primary classrooms developed? Were the preservice and in-service teacher trainings developed? 2) Did teachers use and apply the language-specific teaching and learning methodologies that focus on helping students learn to read and write effectively? 3) Were the appropriate technologies and teacher aids used to support language teaching and learning? and 4) Was technical support provided to RSEBs and MoE for the READ Institutional Improvement (READ II) project?

READ M&E assessed the outcome indicators of reading and comprehension skills of students in grades 2 and 3 as measured by EGRA instruments in seven mother tongue languages. Both types of evidence, fidelity of intervention indicators and outcome indicators, need to be ***jointly analyzed*** to make inferences about the effectiveness and potential improvements in design and delivery the reading intervention. Fidelity of intervention indicators are essential in providing information about the degree to which the recipients (teachers, and through them students) were actually exposed to the intervention, whereas outcome indicators are gauging the degree to which the intervention may have contributed to the targeted change, in this case improvement in student reading and comprehension.

During the year 2017, READ M&E completed a process evaluation of factors associated with READ TA's implementation performance, using a comprehensive set of quantitative and qualitative indicators. However, to enable some inferences about the effects of the intervention, it is necessary to consider the implementation indicators along with the student (and preferably teacher) outcome indicators. In other words, it will be essential to assess the degree to which students measured by EGRA were exposed to the intervention – how often, how much, and how long they were educated under the provisions of designed intervention. For this reason, READ M&E included 34 schools covered by the READ TA performance evaluation study into the EGRA 2018 sample.

Finally, READ M&E agreed to support the World Bank baseline study by including additional schools according to the World Bank (WB) criteria (full primary with O-classes) in the 2018 EGRA sample. A total of 187 schools represent the sample for the WB baseline study, 77 of which were added to the 110 schools that were already included in READ M&E and READ TA portions of the sample.

2.3.2 CHARACTERISTICS OF THE 2018 EGRA SAMPLE

The structure of the 2018 EGRA sample is presented in Figure 2, which shows the number of schools that belong to each of the sample components (READ M&E, READ TA, and WB), including their overlapping numbers.

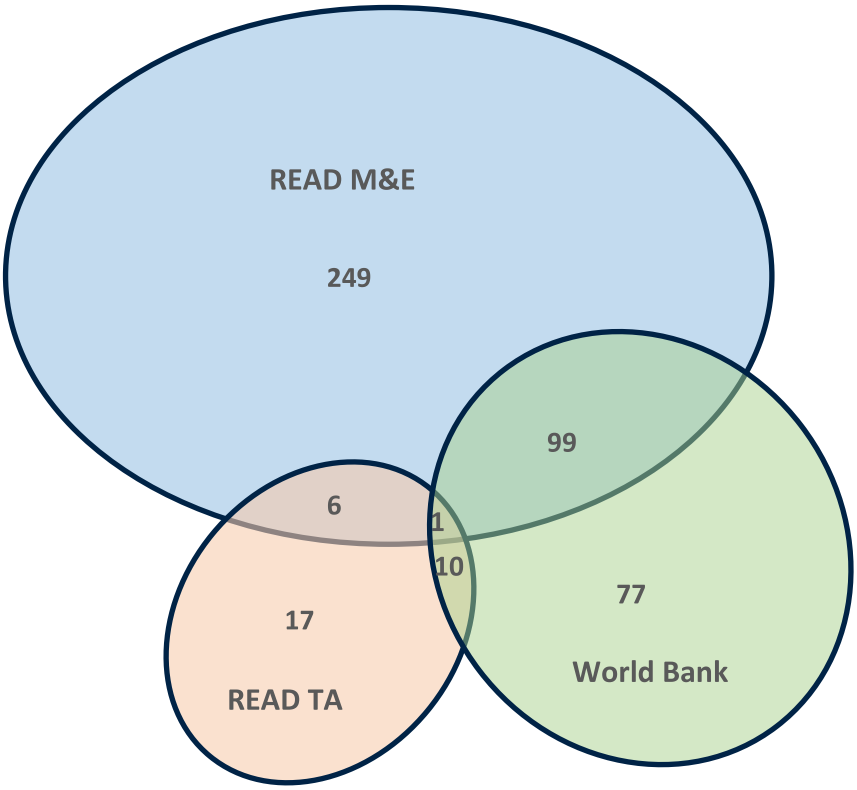


Figure 2. 2018 EGRA: Number of Schools by Sample Components

As Figure 2 illustrates, the portion of the 2018 EGRA sample that was intended to serve for READ M&E study consisted of 355 schools. An additional 27 schools were included from the READ TA study, and an additional 77 schools were included to support the WB baseline study, making a total of 459 schools assessed in the 2018 EGRA administration.

As explained earlier, to enable longitudinal design, the same schools assessed in the 2016 EGRA would be assessed for the 2018 EGRA study (to the degree possible). A total of 198 schools were assessed in both 2016 midline and 2018 endline administrations, which constitutes 56% of the READ M&E 2018 sample. For the EGRA midline schools that could not be accessed in the 2018 endline administration, the replacements schools were selected using the rules for identification of replacement schools (similar location, size, and demographics).

The 2016 EGRA sampling plan, which is described in the READ M&E 2016 midline report, can be summarized as follows:

- Power analysis determined that 300 schools were needed to enable the desired power of statistical analysis, which comes out to roughly 42 schools per language. Therefore, the study met this criterion.

- To conduct the common-persons design study for comparability between the baseline (EGRA 2014) and the midline (2016 EGRA), five schools per language were added, resulting in 47 schools per language.

- The resulting total of 329 schools was increased by three schools per language, bringing the total to 350 schools.

- The 350 schools were sampled from all five regions and seven languages (50 schools per language). At each school, 20 students were selected from grade 2 and 20 from grade 3. An equal number of girls and boys were randomly selected from each grade.

Zones and Schools: In both the 2016 EGRA and the 2018 EGRA, READ M&E selected zones according to their accessibility, security level, and their susceptibility to extreme weather patterns. A full random sampling procedure for the 2016 EGRA was not possible because of security and safety issues, along with the impact of the drought. READ M&E avoided areas listed as priority zones by the Emergency Education Cluster report.

In the 2018 EGRA, READ M&E faced similar issues, so not all schools assessed in the 2016 EGRA could have been revisited in 2018. We replaced schools that were inaccessible due to conditions such as difficult weather (rain and flooding) or security and safety issues. Inaccessibility was defined as: schools requiring that assessors walk for over an hour to reach them; the school was closed; flooding made the school not reachable; or schools that required boats or motorcycles to reach, which assessors did not have. The RSEBs confirmed availability of the sample schools.

The list of zones from which READ M&E drew the 2018 EGRA random sample of schools, along with the number of schools in which data were successfully collected, and the assessed number of students, are presented in Table 2. A map displaying the reading performance of students in assessed zones is provided in Section 2.3.

Table 2. Zones Included in the 2018 EGRA, Along with Number of Schools and Students

Region	Language	Zone	No. of Schools	No. of Students
Amhara	Amharic	Agew Awi	1	39
		Bahir Dar City	1	40
		Debub Gonder	9	370
		Debub Wollo	20	772
		Dessie City Ad.	2	80
		Misrak Gojjam	17	670
		Semen Showa	9	336
		Semen Wollo	6	238
		W. Gojjam	19	738
		Amhara TOTAL	84	3283
Oromia	Afaan Oromo	Arsii	30	1203
		Baale	15	580
		Buunnoo Bedellee	5	189
		E. Harargee	9	335
		East Wollega	3	120
		Gujii	13	522
		I Abbaa Boora	3	120
		Naqamtee	2	78
		Shaw Bahaa	5	193
		W. Harargee	9	261
		Wallagaa Lixaa	4	160
		Oromia TOTAL	98	3761
SNNPR	Haddiysa	Hadiya	50	2002
	Sidamu Affo	Sidama	54	2147
	Wolayttotto	Wolayta	49	1959
	SNNPR TOTAL		153	6108
Somali	Aff Somali	Faafan	45	1615
		Siti	9	332
		Somali TOTAL	54	1947
Tigray	Tigrigna	Central	20	796
		Eastern	3	120
		Mekelle	3	120
		North West	24	957
		South east	7	279
		Southern	13	508
		Tigray TOTAL	70	2780
Overall TOTAL			459	17879

Note. SNNPR is Southern Nations, Nationalities, and People’s Region.

Student Selection: Students from grades 2 and 3 were selected using a random lottery method. If there were 20 or fewer children in a given class, all children in that class were assessed. In each of the classes, an equal number of girls and boys was selected (20 boys and 20 girls).

There were circumstances when it was necessary to replace some of the pupils in the already selected sample, such as pupils with an auditory of visual impairment. Such replacement was done after sample selection by the assessor in consultation with the supervisor.

There had been some concern about teachers and or principals swapping out lower performing students for higher performers, but assessors have reassured us that this swapping did not happen. Concern about students leaving the testing site or misbehaving also proved not to be true. Children in general were eager to have their turn with the assessor and enjoyed waiting. Most teams gave children numbers to keep them in order and to double check that they had not been 'swapped.'

2.3.2 ENSURING HIGH-QUALITY ASSESSORS

The 2018 EGRA 2018 assessor training and data collection were conducted in two phases, according to language groups. Table 3 below presents the schedule for the 2018 EGRA operational administration.

Table 3. Schedule of Training and Data Collection for the 2018 EGRA

Round One Languages: 250 Schools	Round Two Languages: 237 Schools
• Amhara • Oromia • Sidama	• Tigray • Wolayta • Hadiya • Somali
Training day: April 25–27 (Wed–Fri) Deployment: April 28–29 (Sat–Sun) Data collection days: April 30–May 11 Data submission days: May 14–15 (Mon-Tue)	Training day: May 16–18 (Wed–Fri) Deployment: May 19–20 (Sat–Sun) Data collection days: May 21–June 1 Data submission days: June 4–5 (Mon–Tue)

The training for the second phase of the 2018 EGRA was conducted from May 16–18, 2018, in Bishoftu. Both this training and subsequent enumeration were observed by a home office team member and USAID. In the section that follows, we present the details of the expected outcomes, processes, and challenges of the enumerator training and EGRA administration.

PREPARATION AND TRAINING LOGISTICS

Preparation for the April and May training(s) began months in advance with the piloting and equating of the EGRA subtasks, planning team composition (balancing new and returning enumerators in each team), and contacting education officials to provide notice of dates and times. In the weeks leading up to the event, the READ M&E team prepared materials, including EGRA enumerator manuals, subtask scripts and prompts, and materials for roles plays and practice. They also uploaded the necessary Tangerine software and subtasks onto the tablets; prepared the training budget; developed a process for getting individual employment contracts signed; and planned for the efficient distribution of materials, money, forms, subtask prompts, and allocation of vehicles to each team.

The efficient implementation of the training in Bishoftu indicated that the preparation work was done well. Having AIR registered in Ethiopia also made a positive difference in the efficiency of the training (compared to 2016), because the READ M&E team did not need to rely on a cumbersome third-party process to pay the enumerators and transportation companies. Feedback from participants was positive regarding the training and facilitation quality.

OVERVIEW AND TRAINING OUTCOMES

The purpose of this 3-day training event was to prepare 114 EGRA enumerators to successfully collect endline EGRA data in four of the seven READ M&E target regions (Tigray, Somali, Hadiya, and Wolayta). READ M&E chief of party Dr. Solomon Areaya and deputy chief of party Belen Mekonnen provided an overview of the agenda and expected learning outcomes. Addis Yigzaw (USAID) and Todd Drummond (AIR) welcomed and encouraged the 2018 trainee enumerators. By the end of the training, enumerators were expected to be able to accomplish the following:

- Explain the purpose of the EGRA and demonstrate understanding of its component parts (why EGRA, how to select the student sample, timed and untimed subtasks, how to administer the subtasks, surveys and principal interviews);
- Demonstrate skill in using the electronic tablets to collect and upload data;
- Articulate enumerator roles and responsibilities (supervisors vs. regular team members), e.g., supervisors are required to complete supervision reports for each site;
- Serve as effective liaison with schools, woreda education offices, and Regional State Education Bureaus (RSEBs), as necessary;
- Articulate the appropriate procedures for communication with the Addis Ababa office, manage logistics and teams, and troubleshoot problems (e.g., vehicle use, when to call Dr. Solomon for support, how to conduct debriefings with teams to share lessons learned, etc.).

ABOUT THE TRAINEES (EGRA ENUMERATORS)

During this second EGRA enumerator training, 114 potential enumerators from the four regions (males = 106, females = 8) participated. Enumerators with 2016 experience were recruited for the 2018 training. Each of the four regional groups was broken into smaller teams consisting of four enumerators and one supervisor per team (*N* = 5). The breakdown of the teams was as follows: Hadiya (five teams), Wolayta (five teams), Aff Somali (six teams), and Tigrigna (seven teams).[3] Most of the trainees currently work in various administrative positions in the education system. In many cases, enumerators were employees of the RSEBs, which facilitated access to schools. Other enumerators were from higher education or research institutions, and a few were university students. About 65% of the

[3] During round 1 training (Amhara, Sidama, Oromia) there were 135 total participants, male = 120, female = 15.

enumerators had previous EGRA experience. Veteran enumerators were intentionally assigned to teams with newcomers to ensure maximum leverage of their knowledge and experience (more details are provided in the Training Methods section).

POINTS OF EMPHASIS

In addition to preparing enumerators to effectively conduct the EGRA and resolve logistical challenges, the team emphasized that trainees pay specific attention to:

- Understanding the *intent* of the research. As the data were part of an endline study, EGRA administration is a critical endeavor for READ M&E. It was made very clear to enumerators that the data will be analyzed to improve the quality of education and teaching of reading in each of the regions. (i.e. you are only cheating yourselves if you knowingly or unknowingly get careless and fail to strictly follow data collection protocols).
- Assuming the intended role as teams are to visit schools not as “inspectors/evaluators” but rather as data collectors. We placed emphasis on the importance of being perceived as “neutral,” to discourage teachers from interfering in the process to attain high scores.
- The importance of “clean data” (i.e. correct sampling is critical and ensuring that schools did not include ineligible students in the study).

TRAINING DESIGN AND FACILITATION METHODS

After the introductions and training overview, facilitators made brief presentations on the following topics:

- Why EGRA? (current low performance, need to improve)
- What is the EGRA? (subtasks, how they are connected to the Simple View of Reading)
- How piloting was used to ensure comparability of inferences across EGRA years
- How to conduct the school-level student selection (sample) in grades 2 and 3
- Overview of basic Tangerine software and tablet functions

The above-listed activities were covered in the morning of the first day. Because the intent of the training design was to provide as many opportunities to practice administering the EGRA as possible, the presentations were brief. Throughout the training, each language group was observed by READ M&E and USAID team members, who moved among groups to better understand the focus and effectiveness of the activities.

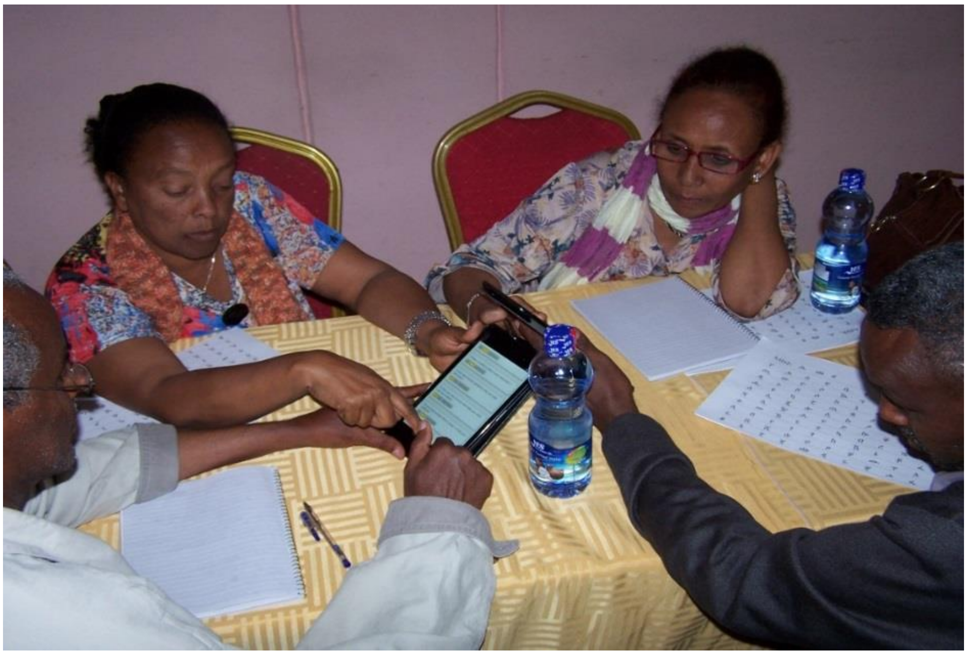


Figure 3. Enumerator trainees from the Tigray Region discuss the tablet functions.

Within each language group, two or three veteran enumerators with the necessary language skills facilitated learning activities. Care was taken to ensure that each of the five-person smaller teams had at least one (usually more) experienced EGRA enumerator who would be able to share experience and skills with newcomers on the team.

Training Methods. Methods of training the enumerators included the following:

- "Fishbowl" activities in which two facilitators modeled the one-to-one (enumerator and student) administration as trainees encircled them and followed along with their tablets, scoring the assessment as if they were conducting an actual assessment;
- Role-play pair work in which a model (practice) EGRA was scored;
- Debriefs and discussions of model administrations to facilitate peer learning; and
- Slide presentations of enumerator "uploaded" data to show what kinds of mistakes are frequently made during uploading to Tangerine, with an emphasis on how such mistakes impact overall data quality.

Training activities were scaffolded to move trainees from easy tasks to ones of increasing complexity across the 3 days. For example, on initial practice tasks, facilitators made more "obvious" mistakes that could be easily caught by inexperienced enumerators. As trainee skills improved, role modeling included less obvious mistakes (e.g., using a pronunciation that was "close to correct," but not quite correct, or skipping many items on timed sections, which required enumerators to adjust quickly). In some languages, issues of pronunciation, dialect, and even vocabulary (and, thus, debate about what constituted a "correct answer") evoked considerable discussion.

During the various activities, READ M&E core team members moved throughout the rooms to observe how well trainees navigated the tablets and marked answers correctly. Measures were taken to assess both individual trainee accuracy and inter-rater reliability across enumerators. This was done in several ways. First, because "incorrect" answers marked by enumerators showed up clearly as "blue" in the answer keys on the tablet screens, it was easy for observers to see consistency across enumerators or identify outlier performance for a given subtask, even when observers lacked knowledge of the language in question. Second, timed subtasks have a matrix, and patterns of marking can be compared quickly by standing behind a row of enumerators and observing the answers they mark as they practice. Outliers in performance were identified quickly and targeted support was provided.

RATER RELIABILITY

Some language groups (e.g., Tigrigna) assessed inter-rater reliability by setting up a role play scenario in which a volunteer played the role of a "child" making set "mistakes". Team members were asked to score the child role player individually. Then, supervisors collected enumerator marking data from their teams to analyze individual enumerators' accuracy and the extent of enumerator consistency (inter-rater reliability).

Any recurring differences in scoring (outliers) were discussed as a group and with individuals as necessary. Inter-rater reliability coefficients were not estimated because a more efficient approach to evaluating performance was needed. Instead, trainees simply discussed the marks of each enumerator to improve the quality and consistency of their work. By the end of the training, observers noted a marked increase in enumerator proficiency with the tablets as well as increased inter-rater reliability across their groups.

Before the training, the team set performance criteria for determining whether a trainee would be eligible for deployment. Attendance at all sessions and active participation in all activities were important criteria; one trainee was dismissed from the training for not meeting these expectations.

2.3.3 EGRA ADMINISTRATION IN THE REGIONS

A typical day of EGRA data collection started with the assessors leaving their hotel very early in the morning. Having made arrangements with the principal and woreda officials earlier in the week, they might have stopped to pick up a woreda official to guide them to a remote school. Sometimes the walk to the school was difficult; it involved crossing streams and walking for an hour or more.

Upon arrival at the school, the team leader introduces herself to the principal and explains the purpose and protocol of the EGRA. Usually, the team arrives before the morning assembly and can inform the grade 2 and 3 teachers to hold children in their lines for the assessors to select the children. The team leader counts the number of children in attendance and then calculates the interval needed to arrive at 10 boys and 10 girls for each grade. The team counts off the children and directs them to separate lines. The team writes down the children's names and gives each child a different number. The selected children move to a comfortable, visible area to wait their turn.

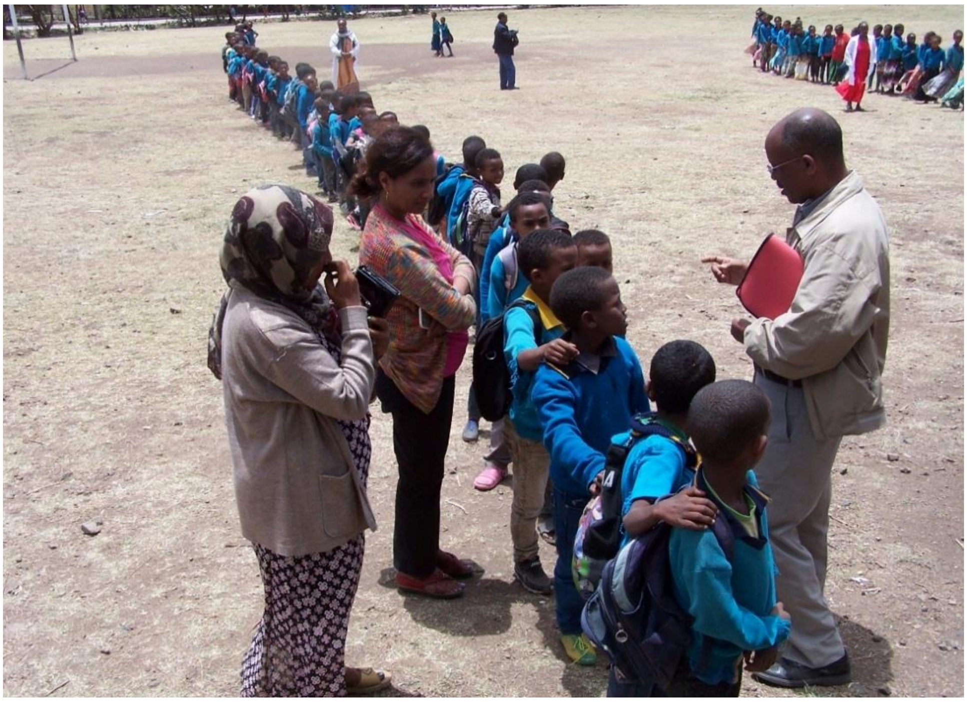


Figure 4. Students line up for sample selection in Maychew, Tigray Province.

The assessors set up areas with tables and chairs in private spaces around the school. They are usually visible to one another, but with significant space between to allow them to hear the child clearly. The waiting children are in view of the assessors, and sometimes the driver or a teacher on break will keep them calm. Usually, there is no problem because the children are happy to have this free time to play with their classmates.

The team leader sets up the other three assessors with their first group of children. When called, the child repeats his or her name, and the assessor verifies the name and number on the master list. After the EGRA is finished, the child receives a pencil as a token of appreciation and returns to the classroom.

While the assessors work with the children, the team leader conducts interviews with principals and teachers. If there is a clear sky and open space, the team leader tries to get the GPS coordinates of the school. Getting the GPS coordinates helps the enumerators identify individual schools. When the team leader finishes their interviews, they assist the team with assessing the remaining children. When all the children are assessed, the team thanks the principal and teachers. Before departure, teams have a brief meeting to upload the data and discuss the day’s events.

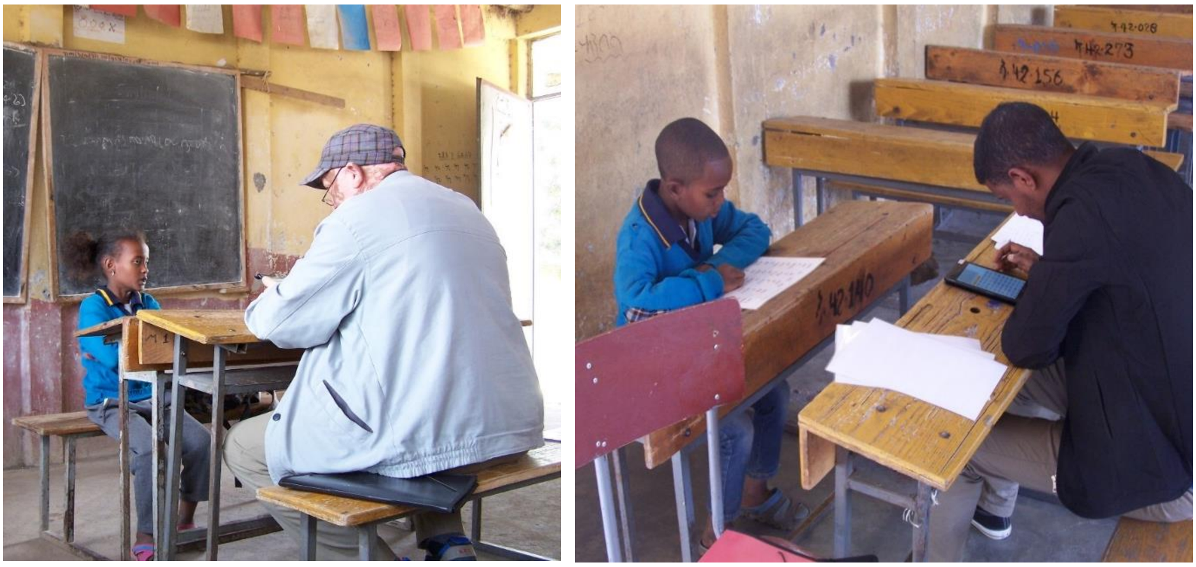


Figure 5. The EGRA is administered in Maychew, Tigray Province.

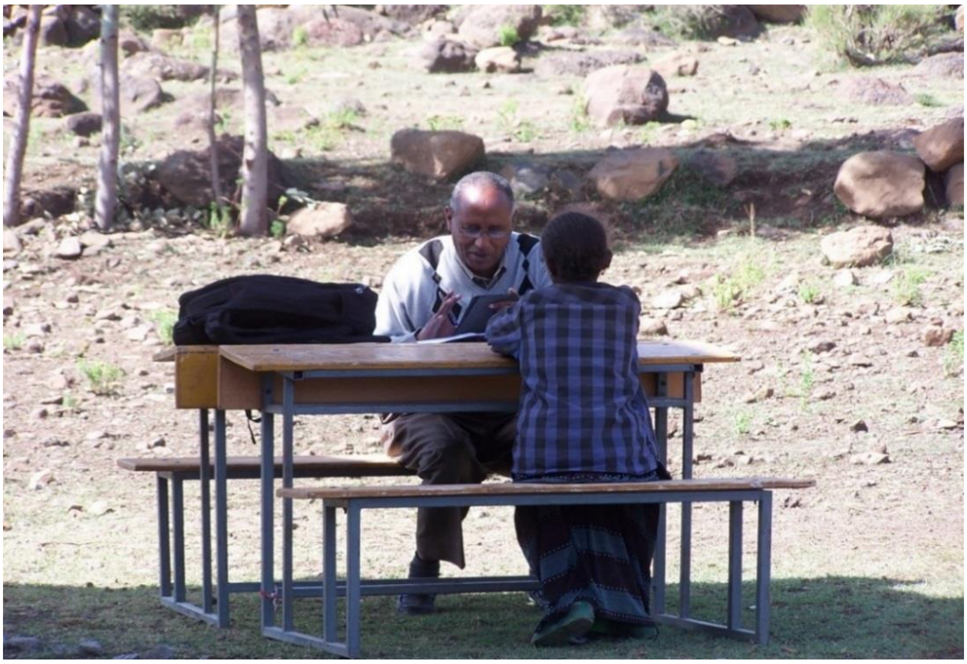
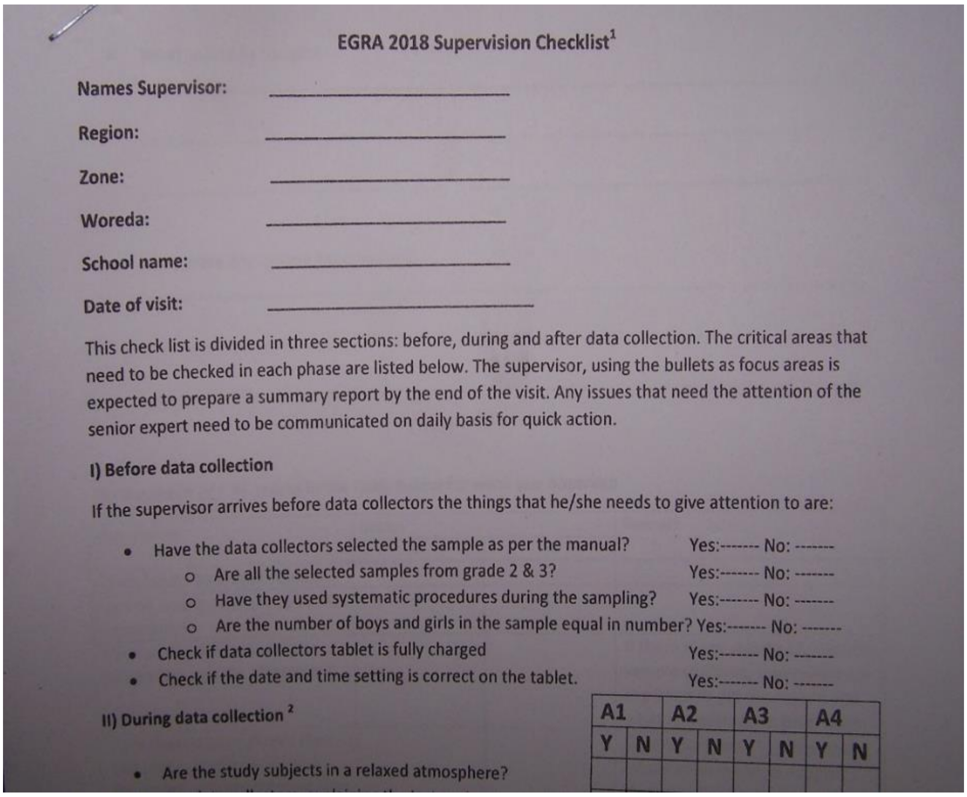


Figure 6. The EGRA is administered one-on-one at Alelibat school, Tigray Province.

2.3.4 SUPERVISION OF EARLY GRADE READING ASSESSMENT DATA COLLECTION

One READ M&E core team member monitored each region, patrolling the assigned area, visiting schools and authorities as necessary to troubleshoot complex issues as they arose. As necessary, these patrols reported issues to the chief of party, Dr. Solomon, and he took action to resolve the problems. Unlike 2016, in 2018 each team consisted of four enumerators and one supervisor whose primary responsibility was to oversee implementation, monitor enumerators' performance, and conduct the principal and teacher interviews. Each supervisor filled out a two-page checklist like the one below for every school visited (see Figure 7 below).



EGRA 2018 Supervision Checklist[1]

Names Supervisor: ____________________

Region: ____________________

Zone: ____________________

Woreda: ____________________

School name: ____________________

Date of visit: ____________________

This check list is divided in three sections: before, during and after data collection. The critical areas that need to be checked in each phase are listed below. The supervisor, using the bullets as focus areas is expected to prepare a summary report by the end of the visit. Any issues that need the attention of the senior expert need to be communicated on daily basis for quick action.

I) Before data collection

If the supervisor arrives before data collectors the things that he/she needs to give attention to are:

- Have the data collectors selected the sample as per the manual? Yes:------- No: -------
    - Are all the selected samples from grade 2 & 3? Yes:------- No: -------
    - Have they used systematic procedures during the sampling? Yes:------- No: -------
    - Are the number of boys and girls in the sample equal in number? Yes:------- No: -------
- Check if data collectors tablet is fully charged Yes:------- No: -------
- Check if the date and time setting is correct on the tablet. Yes:------- No: -------

II) During data collection [2]

	A1		A2		A3		A4	
	Y	N	Y	N	Y	N	Y	N
• Are the study subjects in a relaxed atmosphere?								

Figure 7. An excerpt from the 2018 EGRA Supervisor Checklist.

A READ M&E home office team member and USAID representative Addis Yigzaw traveled to the Tigray Province to observe data collection. They observed a total of eight enumerators and two team supervisors in two schools on May 21 (Maychew) and May 22 (Alelibat). Dr. Solomon and several other team members continued to observe EGRA administration until the completion of data collection. Overall, during the 2 days of visits, the observation team found the school principals, authorities, teachers, and students cooperative in the data collection. The READ M&E trained teams were organized, communicative, and professional with adults and students alike. No irregularities in student sampling or enumerator administration were observed. The enumerators performed well and demonstrated sound knowledge and skill in their endeavors. On May 23, Dr. Solomon and a home office representative visited the Tigray Province RSEB and met with the deputy director to thank him for his support and to answer questions about the EGRA study.

2.3.5 ADMINISTRATION CHALLENGES

Although the EGRA assessment went well and no significant issues disturbed the success of the data collection, there were multiple challenges. READ M&E deployed 249 data collectors plus its field office project staff to supervise the data collection. The mobilization required many vehicles to be deployed simultaneously.

Administering the EGRA via tablets requires that the tablets be in good condition and that an internet connection be available to upload the data on a regular basis. The worst-case scenario would be that the assessor did not have regular access to the internet and then the tablet failed, leading to a loss of all data. However, this did not happen, and data collectors quickly learned to upload data daily using their personal cell phones as internet hotspots, and all data were successfully uploaded.

Figure 8. USAID and READ M&E representatives meet with the Alelibat school principal.

THE ROLE OF EDUCATION OFFICIALS IN DATA COLLECTION

The training team did an excellent job of emphasizing the need for enumerator officials to “not wear their official hats” when collecting EGRA data. In spirit, there was agreement by enumerators that they would do their best. However, the question of data collectors’ actual “independence” needs to be considered more carefully, given that principals may still perceive EGRA as a high-stakes test for them and their students when they see familiar education officials show up at their schools.

EIGHTH GRADE EXAMINATIONS IN THE WOLAYTA REGION

In the Wolayta region, teams reported (after data collection commenced) that in their regional schools were closed for three days due to 8th grade examinations. This was surprising, as the enumerators from this region came from institutions where this information should have been known. Yet, it was not reported to READ M&E in advance. The enumerators stated that they would solve this problem by recalling the necessary grade 2 and grade 3 students to school even though schools would remain closed. However, the READ M&E team was concerned that there was a possibility that only select students would be recalled to schools and the data collected would be biased. Therefore, enumerators were instructed to move forward only with sampling if they had 75% of the class roster in attendance.

GENDER BALANCE OF ENUMERATORS

Most of the enumerators were male. In some respects, this reflects the gender balance in the education administration units throughout the country. However, to increase the number of female leaders serving as enumerators, READ M&E decided that going forward, READ M&E leadership will more vigorously request greater gender balance when RSEBs, zones, and woredas, and nominate candidates to participate in various READ M&E data collection activities. Greater female participation is an important USAID cross-cutting goal in all READ M&E activities.

RAMADAN AND DATA COLLECTION

The Somali group began data collection at the beginning of Ramadan. Because school attendance is already a challenging issue in the Somali region, the Somali group perhaps should have been scheduled to participate in the first cycle for the process to be complete before the start of Ramadan.

2.4 ENDLINE 2018 DATA ANALYSIS

This section describes approaches to data analysis for the 2018 EGRA endline administration. This entails computation of EGRA scores on timed and untimed tasks, analysis of student performance for each language presented in both subtask scores and percentages of students meeting benchmarks, comparative analysis of student performance on midline and endline administrations, and analysis of contextual factors associated with student performance.

2.4.1 COMPUTATION OF EGRA SCORES

The timed tasks in EGRA are letter name recognition, familiar words reading, invented words reading, and ORF. The scores for these tasks were calculated as the number of letters or words correctly read per minute. Three data points are needed to calculate the total score with the following formula applied:

$$WPM = \frac{NC}{(60 - TR)/60}$$

Where: WPM is words (letters) per minute, NC is number of correctly read words (letters), and TR is time remaining

EGRA implements an early-stop rule whereby if the learner does not provide a correct response for the first 10 items for letter name recognition, or five items for familiar words reading, invented words reading, and ORF, the subtask is discontinued, and the child gets a zero score on the task. Untimed tasks in EGRA are initial letter sound, reading comprehension, and listening comprehension. The scores for these tasks are calculated as the percentage of correct responses out of the total number of questions in the subtask.

$$PCT = \frac{NC}{TOT} * 100$$

Where: PCT is the percent-correct score, NC is the number of correct answers, and TOT is the total number of questions.

2.4.2 ANALYSIS OF STUDENT PERFORMANCE ON EGRA

The READ M&E team analyzed EGRA data at the subtask level to evaluate student performance in each language separately, generated evidence of student performance in relation to EGRA benchmarks, examined differences by gender and region, and evaluated differences between midline and endline results. The percentage of zero-scores (discontinuity rate) for each timed subtask is presented. In many areas of Ethiopia, there was a relatively high number of students who could not perform on the subtasks; thus, these students were characterized as non-readers.

All the analyses, including both subtask score and benchmark metrics, were carried out with sampling weights applied. The sampling weights were based on school size (enrollment in corresponding grades), with higher weights accorded to larger schools. When comparing performance by groups (e.g., gender, region), we provide the results of tests of the statistical significance of mean score differences in the appendices.

The purpose of testing for the statistical significance of differences in outcomes is necessary to determine whether the differences were an effect of chance due to sampling error or a result of some systematic factor present in the population. The null hypothesis states that there is no difference between the compared groups in the population. It is a common experience that the independent samples t-test and analysis of variance are reasonably robust to departures from normal distributions and the outcomes are typically the same as those produced by nonparametric tests. Preliminary data checks carried out by both parametric and nonparametric methods found similar results, so parametric methods were used throughout this study.

Because datasets with large *N*s frequently produce statistically significant results, it is important to estimate an effect size to obtain a fuller understanding of the practical effect of any differences in mean scores. We used an estimate of Cohen's d, which means that the effect-size coefficients are expressed in terms of standard deviations. For example, an effect size of 0.5 indicates that the difference between mean scores is one half of a standard deviation.

In Chapter 3: Results of the 2018 EGRA Endline for Each Language, we present the results of the 2018 EGRA endline administration expressed in two types of metrics: (a) the subtask scores (words per minute for timed tasks or percent-correct scores for untimed tasks), and (b) percentage of students falling within different performance levels according to established benchmarks.

The comparative analysis of the 2018 EGRA endline, the 2016 EGRA midline, and the EGRA 2014 baseline results is presented in Chapter 4: Comparison of EGRA Performance Across Years. In both chapters 3 and 4, an elaborate analysis is carried out using the READ M&E sample of schools, whereas the corresponding analysis for all schools assessed by the 2018 EGRA (including READ M&E, READ TA, and WB samples) is provided as a separate annex to this report.

In Chapter 5: Factors Associated with Student Performance, we present the results of analyses of factors associated with student performance based on the survey data. One limitation to analyzing these data with complex parametric models (e.g., regression analysis) is an assumption that the outcome variable (EGRA scores) must be normally distributed. However, the data for most languages in both grades indicate a considerable positive skew because of the large number of "zero" and low scores on the subtasks. As a result, emphasis in our analyses was placed on simple bivariate analyses proven as robust to departures from normal distribution, to test the relationship between student performance on EGRA subtasks and the categorical background variables of interest. Analysis of factors associated with student performance was done across all languages and grades. All statistical analyses were conducted using SPSS software.

3. RESULTS OF THE 2018 EGRA ENDLINE FOR EACH LANGUAGE

In this chapter, we discuss the major results of the 2018 EGRA endline administration, which are presented using two types of reporting frameworks. The first framework uses units as obtained by administering the EGRA subtasks (letters or words per minute for timed tasks, and percent-correct scores for untimed tasks). Although these units are straightforward, transparent, and easy to understand, they don't convey sufficient information about the value of results, nor can they be compared across languages without limitations imposed by language differences. Thus, the second reporting framework is based on reading performance standards, which resolves the limitations described previously by providing a means through which to place student achievement into meaningful and readily interpretable levels. These levels are comparable across languages because they are designed by each language separately, taking into consideration language specificities. The measure used in this reporting framework is percentage of students falling within different performance levels according to established benchmarks (reading with full, increasing, or limited comprehension).

3.1 MEAN SCORES OF EGRA SUBTASKS BY GRADE AND GENDER

Student achievement on the 2018 EGRA endline administration is presented in terms of subtask mean scores for the entire READ M&E sample and disaggregated by grade and gender for each of the seven languages.

3.1.1 TIMED SUBTASKS BY GRADE

This section presents the 2018 EGRA mean fluency scores for the timed subtasks: letter name recognition, familiar words reading, invented words reading, and oral reading fluency, by grade levels for each of the seven languages.

The mean fluency scores of the timed tasks for the seven languages by grade are presented in Table 4 and Notes. LNR is letter name recognition; FWR is familiar words reading; IWR is invented words reading; ORF is oral reading fluency.

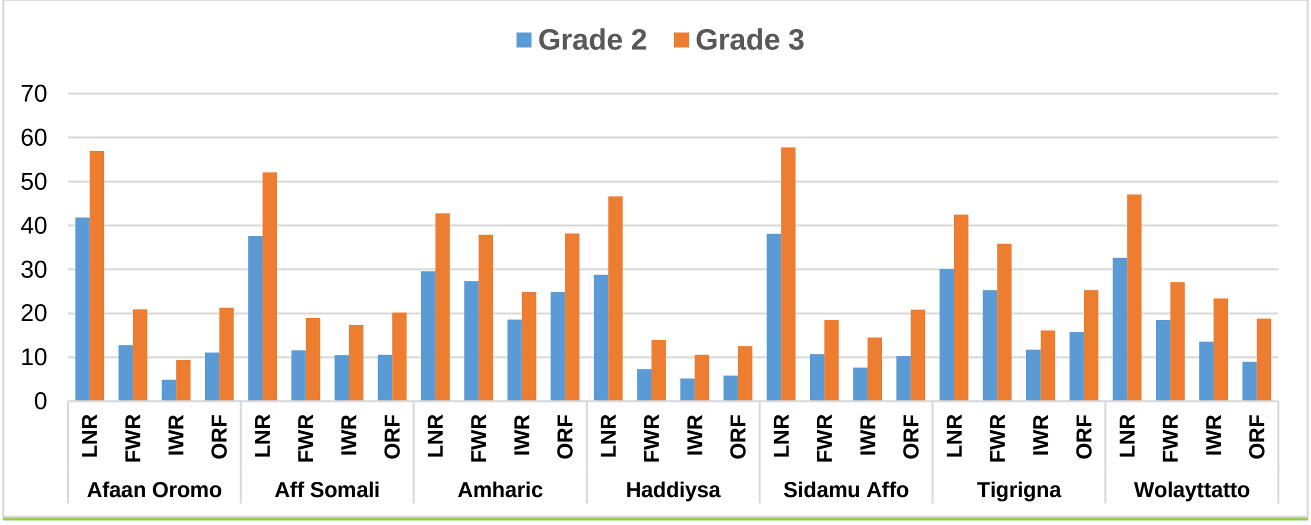
. Additionally, results of testing for statistically significant differences between grades (by independent sample t-tests) and practical differences between grades (by Cohen's d) are provided in Appendix 2.

In each language, the mean differences between the two grade levels in all EGRA subtasks are statistically and practically significant in favor of grade 3. This means that in all languages, grade 3 students were able to read substantially better than grade 2 students, which indicates positive grade gain. The sizes of grade differences for ORF across all languages measured by Cohen's d range from 0.46 for Haddiysa to 0.68 for Amharic, all categorized as educationally significant effect sizes, indicating that substantial reading gains are occurring between grades 2 and 3 (Wolf, 1986). This finding can be attributed to various factors, one of them is natural maturation, but it is highly plausible to attribute this gain to the effect of education.

Table 4. 2018 EGRA Mean Fluency Scores, by Grade

Language	Grade	Letter Name Recognition	Familiar Words Reading	Invented Words Reading	Oral Reading Fluency	Cohen's d for ORF
Afaan Oromo	Two	41.8	12.8	4.9	11.1	0.58**
	Three	56.9	20.9	9.4	21.3	
Aff Somali	Two	37.6	11.6	10.5	10.6	0.50**
	Three	52.1	19.0	17.4	20.2	
Amharic	Two	29.6	27.3	18.6	24.9	0.68**
	Three	42.8	37.9	24.9	38.1	
Haddiysa	Two	28.8	7.3	5.2	5.9	0.46*
	Three	46.7	13.9	10.6	12.5	
Sidamu Affo	Two	38.2	10.7	7.7	10.3	0.60**
	Three	57.6	18.5	14.5	20.8	
Tigrigna	Two	30.1	25.3	11.7	15.8	0.57**
	Three	42.5	35.9	16.1	25.3	
Wolayttatto	Two	32.6	18.5	13.5	8.9	0.49*
	Three	47.1	27.1	23.5	18.9	

Notes. LNR is letter name recognition; FWR is familiar words reading; IWR is invented words reading; ORF is oral reading fluency.
* Cohen's d of 0.25 and above indicates the size of difference is educationally significant; something was learned.
** Cohen's d of 0.50 and above indicates a strong educational effect; something substantially changed (Wolf, 1986).



Notes. LNR is letter name recognition; FWR is familiar words reading; IWR is invented words reading; ORF is oral reading fluency.

Figure 9. Mean Fluency Scores, by Grade and Language

3.1.2 TIMED SUBTASKS BY GENDER

Table 5 and Notes. LNR is letter name recognition; FWR is familiar words reading; IWR is invented words reading; ORF is oral reading fluency.

present the mean scores of the fluency tasks by gender in grades 2 and 3 for all languages. Statistical tests of gender differences (using t-tests), as well as measures of the corresponding sizes of differences (by Cohen's d), for all grades and languages, are presented in Appendix 3.

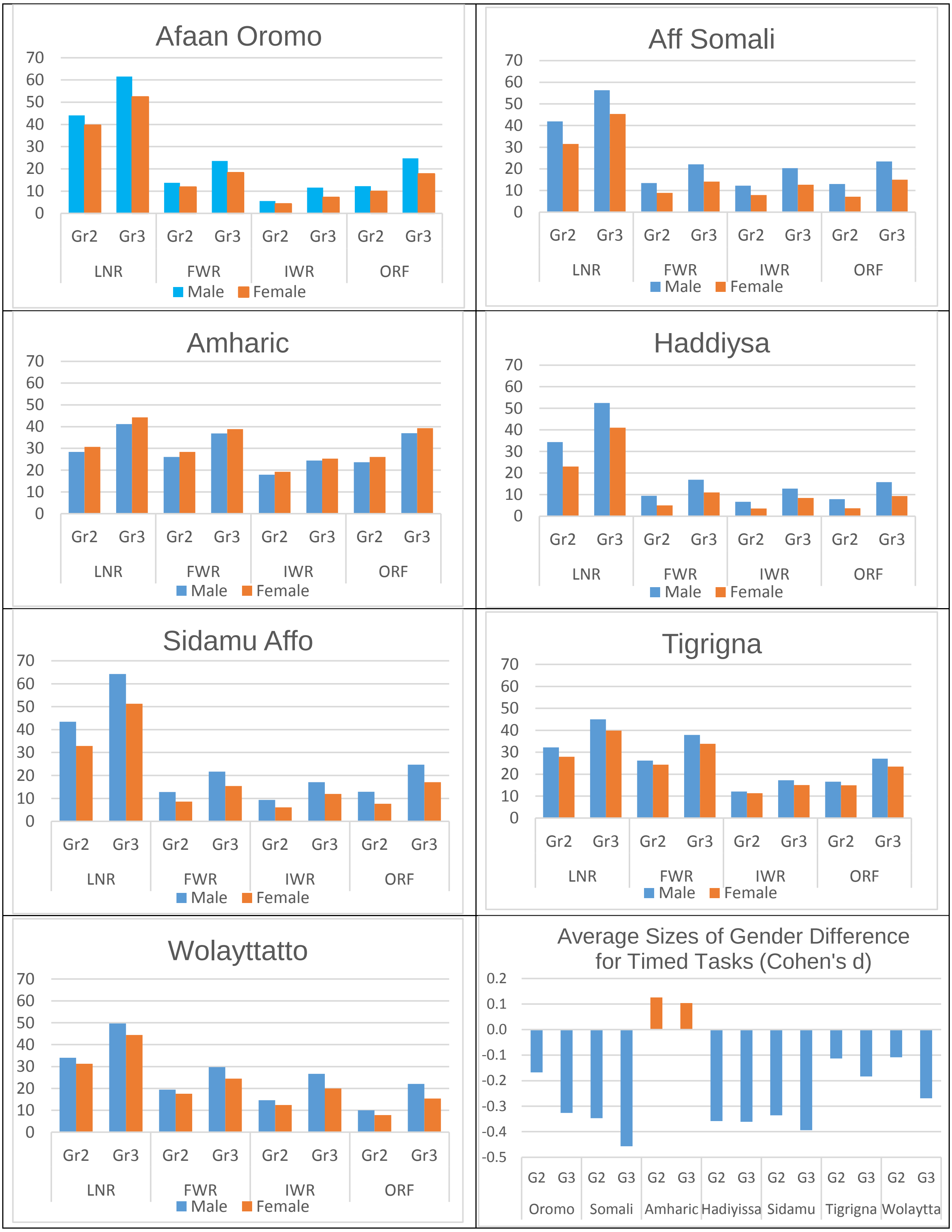
Unlike the 2016 EGRA, which showed stronger performance among boys in some languages and stronger performance among girls in others, the results of the 2018 EGRA clearly speak to gender inequality. The 2018 EGRA results show boys outperforming girls in reading skills in primary grades of Ethiopia.

Looking at the ORF (passage words) scores in grade 2, boys performed better than girls in most languages. Substantial differences are observed In Aff Somali (5.8 words per minute), in Haddiysa (4.3 wpm), and in Sidamu Affo (5.3 wpm). Boys also performed better in Afaan Oromo, Tigrigna, and Wolayttatto, but those differences, although statistically significant, were of smaller sizes that can be interpreted as practically insignificant. Girls outperformed boys only in Amharic (2.4 wpm), but the size of the difference is too small to be considered as practically significant.

Table 5. 2018 EGRA Mean Fluency Scores, by Gender (With Cohen's d for Oral Reading Fluency)

		Letter Name Recognition		Familiar Words Reading		Invented Words Reading		Oral Reading Fluency		d for ORF	
Language	Gender	Gr2	Gr3	Gr2	Gr3	Gr2	Gr3	Gr2	Gr3	Gr2	Gr3
Afaan Oromo	Male	44.1	61.5	13.8	23.5	5.6	11.6	12.3	24.8	-0.17	-0.33*
	Female	39.6	52.3	11.8	18.2	4.2	7.1	9.9	17.8		
Aff Somali	Male	41.9	56.3	13.5	22.1	12.3	20.3	13.0	23.4	-0.33*	-0.41*
	Female	31.5	45.4	8.9	14.1	7.9	12.7	7.2	15.1		
Amharic	Male	28.4	41.2	26.1	36.9	17.9	24.4	23.6	37.0	0.14	0.11
	Female	30.8	44.3	28.4	38.9	19.2	25.3	26.0	39.3		
Haddiysa	Male	34.3	52.5	9.5	16.9	6.7	12.8	8.0	15.8	-0.36*	-0.40*
	Female	23.0	41.0	5.0	11.0	3.6	8.4	3.7	9.3		
Sidamu Affo	Male	43.4	64.3	12.8	21.7	9.3	17.1	12.9	24.7	-0.35*	-0.40*
	Female	32.8	51.3	8.6	15.4	6.1	11.9	7.6	17.1		
Tigrigna	Male	32.2	45.0	26.2	37.9	12.1	17.2	16.5	27.1	-0.11	-0.20
	Female	28.0	39.9	24.3	33.8	11.4	15.0	15.0	23.4		
Wolayttatto	Male	34.0	49.7	19.5	29.7	14.6	26.7	10.0	22.1	-0.12	-0.30*
	Female	31.3	44.4	17.5	24.4	12.5	20.0	7.9	15.4		

Notes. LNR is letter name recognition; FWR is familiar words reading; IWR is invented words reading; ORF is oral reading fluency.
(-) Negative sign of Cohen's d indicates difference in favor of boys; blue shading denotes languages in which boys outperformed girls and orange shading vice versa.
* Cohen's d of 0.25 and above indicates that the size of the difference is educationally significant (Wolf, 1986).



Notes. LNR is letter name recognition; FWR is familiar words reading; IWR is invented words reading; ORF is oral reading fluency.

Figure 10. Gender Differences in Timed Tasks, by Grade (With Cohen's d)

The ORF results in grade 3 indicate that boys performed significantly better than girls in all languages except Amharic. Substantial differences in favor of boys are observed in Aff Somali (8.3 wpm), Haddiysa (6.5 wpm), Sidamu Affo (7.6 wpm), Afaan Oromo (7.0 wpm), and Wolayttatto (6.7 wpm), whereas in Tigrigna the difference is marginal (3.5 wpm). Girls performed slightly better only in Amharic, with the difference being negligible. The fact that gender equality is observed only in Amharic calls for further exploration to understand the reasons behind this finding. One of the plausible explanations could be that those who speak Amharic live in more affluent areas with better educated parents who value girls' education.

Notes. LNR is letter name recognition; FWR is familiar words reading; IWR is invented words reading; ORF is oral reading fluency.

above contains a series of graphs depicting gender performance in all timed subtasks for both grades and all seven languages. The graph in the lower left corner depicts the sizes of gender differences in terms of Cohen's d effect-size measures. This graph visualizes a strong dominance of boys in performance on the 2018 EGRA fluency subtasks (blue bars). Note that a negative sign of Cohen's d for boys simply reflects the direction of the difference based on how it was computed (the mean for boys was subtracted from the mean for girls). In most cases, Cohen's d measures of effect size are larger than 0.25 (in absolute value), which indicates that the differences by which boys outperform girls are educationally significant (Wolf, 1986).

3.1.3 UNTIMED SUBTASKS BY GRADE

This section shows the mean scores of the untimed tasks disaggregated by grade for each of the seven languages. The untimed subtasks are initial letter sound, as a measure of phonemic awareness; reading comprehension, and listening comprehension. More detailed results of statistical testing of differences between performance at grade levels (using t-tests) and corresponding sizes of differences (by Cohen's d) are presented in Appendix 4.

As Table 6 and Figure 11 illustrate, in all languages the mean differences between the two grade levels are statistically significant in favor of grade 3. Looking at reading comprehension as the most pertinent untimed task, the strongest grade gains are observed in the Afaan Oromo and Amharic languages (reading comprehension score increase by 15.5 and 17.1 percent-correct points, respectively), which is considered at a level of strong educational effect. Although reading comprehension grade gains vary among languages, from 7.5 in Haddiysa to 17.1 percent-correct points in Amharic, all of them fall into the category of moderate or strong educational effects. The increases of grade mean scores in phonemic awareness are smaller but more variable, ranging from 1.0% (Wolayttatto) to 14.0% (Afaan Oromo), which may be due to the different nature of the phonemic awareness subtask across languages. The observed grade gains in listening comprehension are also relatively small and variable (1.9% to 16.4%).

To summarize, considering that strong grade gains in reading comprehension are paralleled with strong gains in ORF, these findings encourage the claim that growth in students' reading comprehension, and ultimately in students' learning, is contingent on growth in reading fluency skills. Reading gains from grade 2 to grade 3 were apparent in 2014, but still smaller than the gains observed in 2016 and 2018.

The average Cohen's d across languages in 2014 was 0.51 for ORF and 0.43 for reading comprehension, whereas in 2016 it was 0.54 for ORF and 0.52 for reading comprehension. In 2018, the average Cohen's d was 0.55 for ORF and 0.51 for reading comprehension. This suggests that reading growth from grade 2 to grade 3 was getting steeper from 2014 to 2016, but then the growth rate from grade 2 to 3 did not change much from 2016 to 2018.

Table 6. Mean Scores of Untimed Tasks, by Grade and Language

Language	Grade	Initial Letter Sound	Listening Comprehension	Reading Comprehension	d for Reading Comprehension
Afaan Oromo	Two	47.2	60.4	13.0	0.62**
	Three	61.1	71.0	28.5	
Aff Somali	Two	83.2	77.1	10.1	0.50**
	Three	89.3	81.9	21.2	
Amharic	Two	81.7	66.2	24.8	0.59**
	Three	85.7	72.7	41.9	
Haddiysa	Two	85.1	73.5	12.9	0.43*
	Three	89.4	81.0	20.3	
Sidamu Affo	Two	90.5	82.9	9.2	0.51**
	Three	94.2	87.8	20.1	
Tigrigna	Two	72.6	46.3	17.6	0.40*
	Three	83.5	62.6	26.3	
Wolayttatto	Two	74.1	52.5	11.4	0.49*
	Three	75.1	54.4	21.4	

Note. * Cohen's d of 0.25 and above indicates the size of difference is educationally significant; something was learned. ** Cohen's d of 0.50 and above indicates a strong educational effect; something substantially changed (Wolf, 1986)

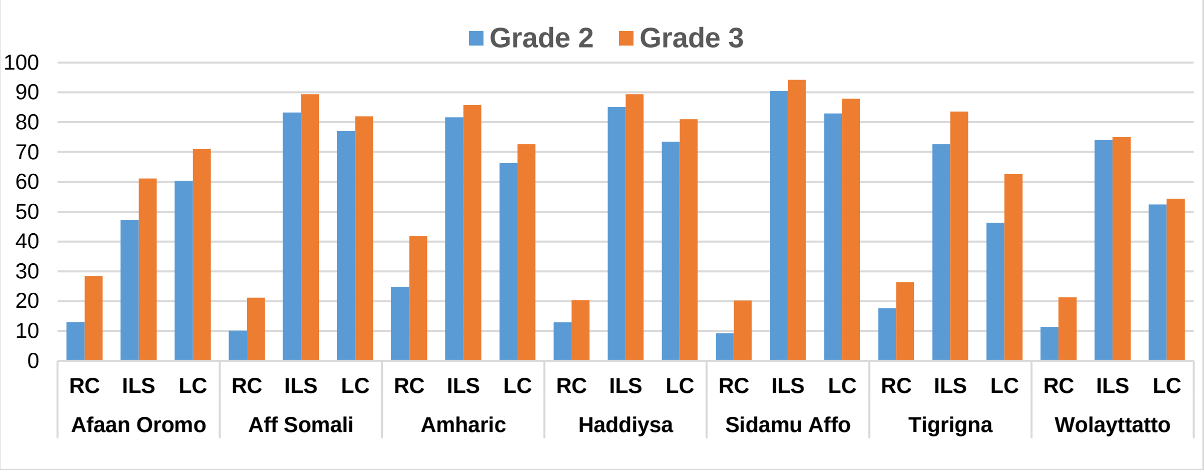


Figure 11. 2018 EGRA Mean Scores of Untimed Tasks, by Grade

3.1.4 UNTIMED SUBTASKS BY GENDER

This section presents the mean scores of the untimed tasks disaggregated by gender for each of the seven languages and two grades. The same untimed subtasks are reported here as in previous section: Initial Letter Sound (ILS), Reading Comprehension, and Listening Comprehension (LC). All the results of statistical significance testing between performance of boys and girls on untimed tasks (using t-test), along with corresponding sizes of difference (by Cohen's d) are presented in Appendix 5.

The results of the 2018 EGRA in untimed tasks in both grades 2 and 3 disaggregated by gender are presented in Table 7. Given that reading comprehension is the most pertinent task, the sizes of the differences in reading comprehension between boys and girls are presented in the last two columns (in terms of Cohen's d).

When looking at grade 2, the reading comprehension mean scores and corresponding effect sizes show that boys performed higher than girls in all languages but Amharic. Boys outperformed girls substantially in Sidamu Affo (by 4.9% points) and in Haddiysa (by 4.0% points), whereas in Aff Somali and Tigrigna, the differences were marginal, and in Affan Oromo and Wolayttatto, they were negligible. Girls showed better reading comprehension performance than boys only in Amharic, but the size of the difference is at the margins of practical significance.

Table 7. 2018 EGRA Mean Scores of Untimed Tasks, by Gender

Language	Gender	Initial Letter Sound		Listening Comprehension		Reading Comprehension		d for Reading Comprehension	
		Gr2	Gr3	Gr2	Gr3	Gr2	Gr3	Gr2	Gr3
Afaan Oromo	Male	50.9	66.8	60.3	72.1	13.8	32.9	-0.09	-0.29*
	Female	43.6	55.5	60.6	69.9	12.1	24.0		
Aff Somali	Male	84.0	90.3	79.0	85.6	11.7	24.8	-0.22	-0.35*
	Female	82.1	87.8	74.4	76.1	7.9	15.5		
Amharic	Male	81.1	85.2	64.7	72.7	22.6	39.8	0.17	0.13
	Female	82.2	86.2	67.7	72.6	26.9	44.1		
Haddiysa	Male	87.1	90.1	76.5	83.9	14.9	23.3	-0.30*	-0.29*
	Female	83.1	88.8	70.4	78.2	10.8	17.4		
Sidamu Affo	Male	91.5	93.6	83.0	86.0	11.7	23.7	-0.28*	-0.28*
	Female	89.4	94.7	82.9	89.7	6.7	16.7		
Tigrigna	Male	75.3	85.7	48.0	46.3	19.2	28.1	-0.19	-0.14
	Female	69.7	81.3	44.4	44.4	15.9	24.6		
Wolayttatto	Male	76.0	75.0	52.6	54.0	14.2	30.8	-0.09	-0.25*
	Female	72.0	75.0	52.3	54.8	11.9	22.7		

Note. (-) Negative sign of Cohen's d indicates difference in favor of boys
* … Cohen's d of 0.25 and above indicates that the size of difference is educationally significant (Wolf, 1986)

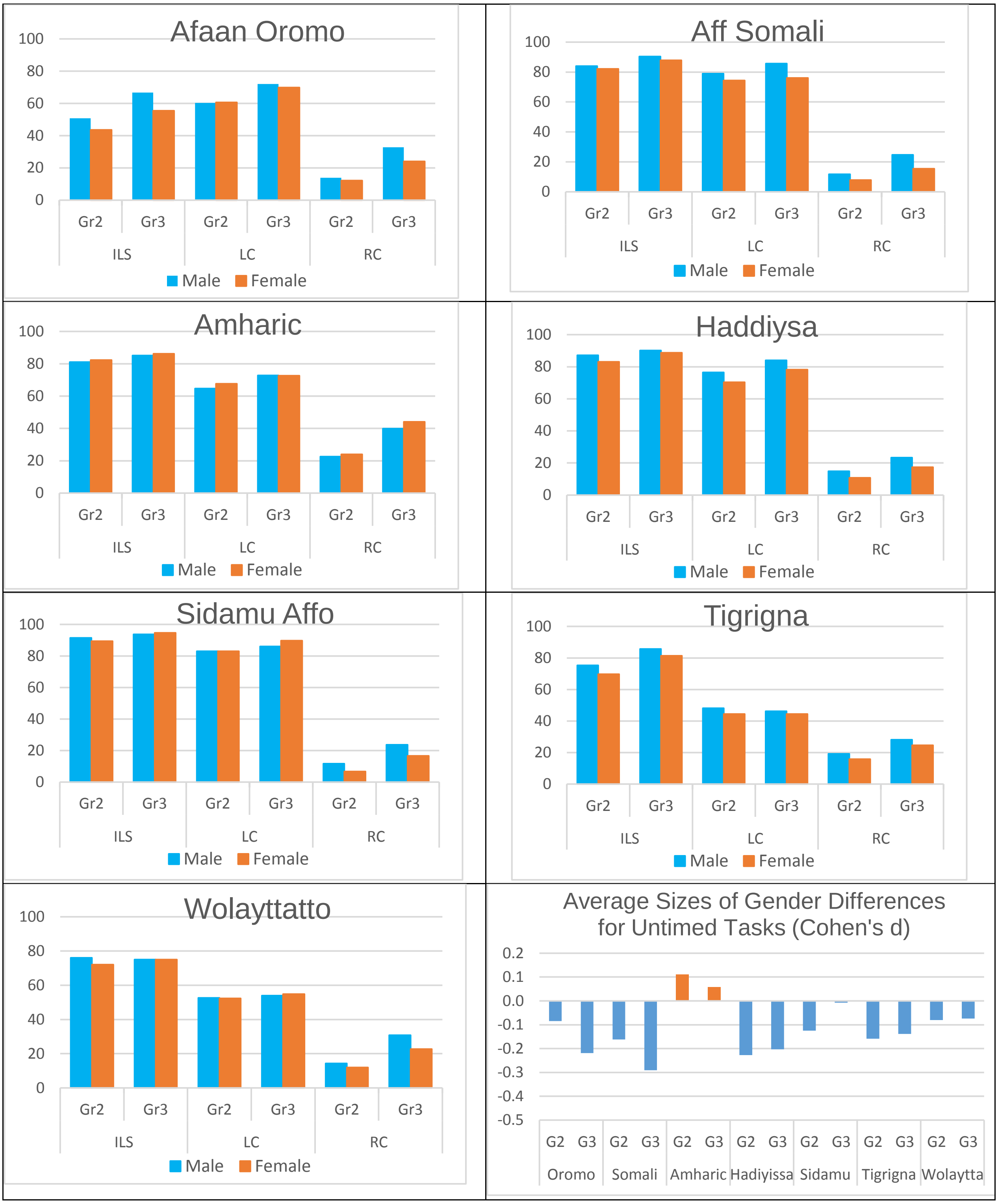


Figure 12. Gender Differences in Untimed Subtasks, by Grade (With Cohen’s d)

Regarding grade 3, gender differences in untimed tasks show similar patterns, but they are somewhat larger than in grade 2. Looking at the reading comprehension scores, boys substantially outperformed girls in five languages, and in one language the difference was below practical importance. Girls performed slightly better in reading comprehension only in Amharic; the size of the difference is negligible.

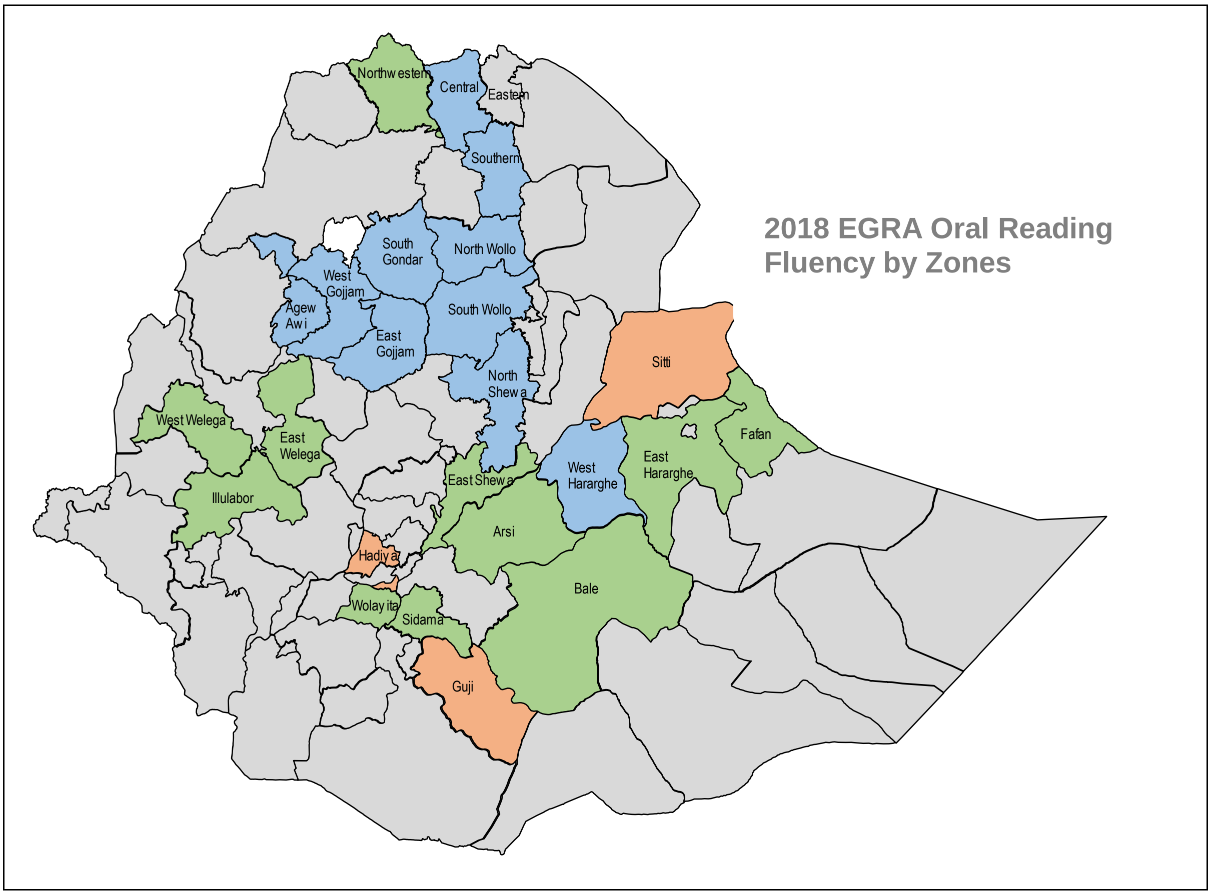
Figure 12 above contains graphs depicting gender performance for untimed tasks for grades 2 and 3 in all 7 languages. Based on the graph in the bottom right corner, which shows the average sizes of differences in terms of Cohen's d, boys demonstrated higher performance than girls in most languages. However, compared to gender differences in timed fluency tasks, the untimed portion of EGRA, dominated by reading comprehension, shows less gender inequality.

3.1.4 2018 EGRA RESULTS BY ZONES

The presentation of EGRA results by smaller geographical units (zones) may add value to policy and instructional support at the local level. In this section, we present the 2018 EGRA results in two major subtasks (ORF and reading comprehension) disaggregated by participatory zones in which the 2018 EGRA was administered. To enhance understanding of geographical distribution of EGRA performance, READ M&E presents the zonal results in the form of a map, in which the zones are divided into three categories: those performing around the national mean (color coded green), those that are substantially above the national mean (blue), and those substantially below the national mean (orange). The criterion for classifying zonal mean scores in categories substantially above or below the national mean is based on the effect size (Cohen's d) of 0.20. That is, if the zonal mean score was above the national average with an effect size of at least 0.20, it was classified as *substantially above*, or if the zonal mean score was below the national average for the effect size of at least 0.20, it was classified as *substantially below*.

It is important to note that, because of substantial differences between languages and the non-representative sampling of schools within zones, these results should be interpreted cautiously and used primarily for descriptive and formative purposes. In no case should these results be used for evaluative comparisons between languages, or for high-stakes evaluation of performance of local educational officials.

Table 8 and Table 9 present the results in participating zones on the ORF subtask, whereas Figure 13 and Figure 14 show the zonal results in the reading comprehension subtask. Appendices 6 and 7 present more detailed statistics for mean scores on the ORF and reading comprehension scores, respectively.

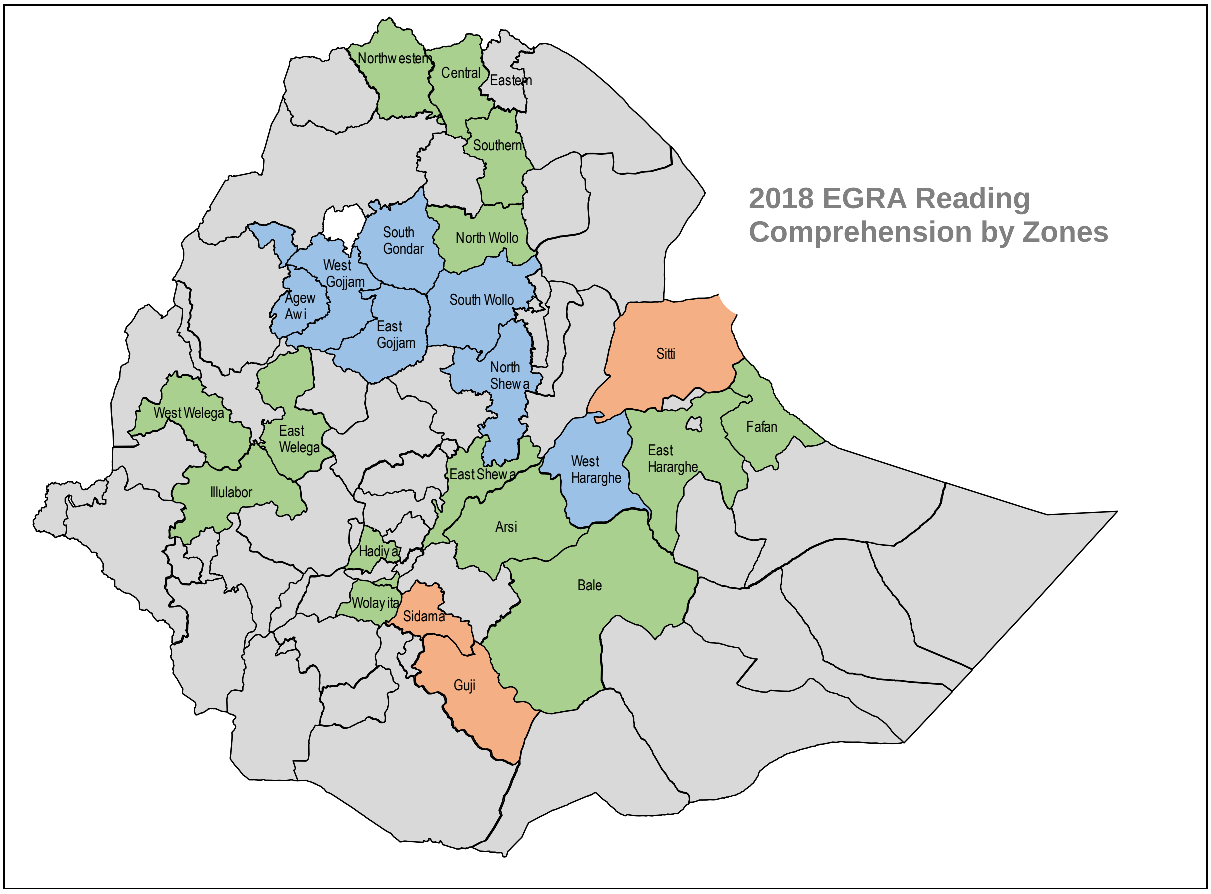


Note. Blue = above the national mean; green = around the national mean; orange = below the national mean. Categorization is based on effect size (Cohen's d) of 0.20.

Figure 13. 2018 EGRA Results for the Oral Reading Fluency Subtask, by Zones

Table 8. Zonal Mean Scores on the Oral Reading Fluency Subtask

Above the National Mean		Around the National Mean		Below the National Mean	
Zone	**ORF Mean**	**Zone**	**ORF Mean**	**Zone**	**ORF Mean**
Agew Awi	42.0	Northwestern	19.2	Sitti	12.5
North Shewa	38.7	Bale	17.8	Guji	11.9
West Gojjam	33.0	**National**	**17.5**	Hadiya	9.2
South Wollo	31.5	Arsi	17.5		
East Gojjam	31.1	West Welega	16.9		
South Gondar	26.9	Wolayita	16.3		
Central	22.3	Fafan	15.7		
North Wollo	22.2	East Welega	15.5		
Southern	21.9	East Hararghe	15.3		
West Haraghe	21.7	Sidama	15.3		
		East Shewa	14.4		
		Illubabor	14.3		



Note. Blue = above the national mean; green = around the national mean; orange = below the national mean. Categorization is based on effect size (Cohen's d) of 0.20.

Figure 14. 2018 EGRA Results for the Reading Comprehension Subtask, by Zone

Table 9. Zonal Mean Scores on the Reading Comprehension Subtask

Above the National Mean		Around the National Mean		Below the National Mean	
Zone	**Mean**	**Zone**	**Mean**	**Zone**	**Mean**
Agew Awi	51.4	Southern	23.9	Guji	15.1
North Shewa	40.3	Central	23.4	Sidama	14.4
West Gojjam	36.6	Arsi	22.7	Sitti	11.7
East Gojjam	33.5	West Welega	22.7		
South Wollo	33.2	Bale	21.9		
West Haraghe	30.6	Northwestern	20.4		
South Gondar	26.7	**National**	**20.3**		
		Wolayita	19.9		
		Illubabor	19.5		
		East Welega	19.0		
		North Wollo	18.3		

Above the National Mean		Around the National Mean		Below the National Mean	
Zone	Mean	Zone	Mean	Zone	Mean
		East Hararghe	18.3		
		East Shewa	18.1		
		Hadiya	16.6		
		Fafan	16.2		

3.2 2018 EGRA RESULTS BY PERFORMANCE STANDARDS

Performance standards are a necessary component of educational assessments to provide a criterion-referenced evaluation framework for interpreting student performance. The purpose of performance standards is both formative and summative. The major role of EGRA performance standards is to support the MoE in developing the capacity of teachers to monitor and boost students' learning progress, and to promote literacy acquisition during the first four years of schooling at the system level.

The benefits of establishing performance standards for early grade reading are as follows: (a) they increase the capacity of policymakers to better support the implementation of the reading curriculum, and (b) they create an evaluation framework for monitoring the progress of literacy acquisition.

3.2.1 WHAT ARE BENCHMARKS FOR READING PERFORMANCE?

Benchmarks for student performance in early grade reading are a product of conceptual and operational specifications of the ***reading skills students are expected to acquire*** through the early grades of primary school. They are typically used to set national targets for reading achievement by first defining the targeted competency level, and then specifying the percentage of students expected to reach that level in a given time interval.

For example, the most desired reading performance standard is typically formulated as follows: "By the end of grade 3, students will be expected to read a grade-appropriate text fluently with a level of comprehension of at least 80%." Then we could set the expectation or target, for example: "At least 55% of students will reach this level of proficiency by 2025."

Although the reading curriculum for early grades usually provides a detailed guideline for content and skills to be acquired, along with corresponding instructional methodology, it typically does not specify what the targeted rates of fluency and comprehension should be for each language.

Benchmarks for reading fluency and comprehension in each language enable the MoE and other stakeholders to monitor progress in achieving reading targets and, even more importantly, to provide guidance to policymakers and instructional support teams at the school, district, provincial, and national levels.

READ M&E uses the benchmarks developed in the January 2015 workshop held by USAID and the MoE and facilitated by RTI. Based on the relationship between reading fluency and reading comprehension (more detail is provided in the next section), the three different levels of reading performance are conceptualized and operationalized in terms of cut scores on the ORF subtask (USAID Ethiopia, 2015):

- ***Level 1: Reading fluently with full comprehension****—students achieving the level of reading fluency that the data indicate corresponds with full or almost full comprehension;*

- ***Level 2: Reading with increasing fluency and comprehension****—students who have some reading fluency but have not yet reached the above-mentioned level of fluency and comprehension; and*

- ***Level 3: Reading slowly and with limited comprehension****—students scoring above zero but at the lower end of the reading fluency score distribution.*

A fourth level of reading ability was also discussed and is important to monitor: students who are not yet reading. This level did not require participants to set a benchmark because it is defined as those students who score zero on the oral passage reading portion of EGRA.

Based on the cut scores in the ORF subtask, which delineate reading proficiency levels for each language that were established in the benchmarking workshop, we classified student performance on three corresponding levels. The fourth (lowest) performance level included students with zero scores, or non-readers. It should be noted that the application of benchmarks that were established on the traditional EGRA paper-based form onto the data collected by tablets was enabled through the comparability study carried out as a part of the 2016 midterm EGRA data collection. Further comparability between the 2018 and 2016 EGRA forms was established by the equating study carried out as a part of 2018 pilot administration.

3.2.1 RELATIONSHIP BETWEEN ORAL READING FLUENCY AND READING COMPREHENSION SCORES

The rationale for EGRA benchmarks relies on the relationship between reading comprehension and ORF. This relationship can be visualized by conditional boxplots showing how are the reading comprehension scores related to the ORF scores at grades 2 and 3. We present two examples of these plots, in Figure 15 for Amharic and in Figure 15. Association of Reading Comprehension with Oral Reading Fluency for Amharic (Correlation = 0.86)

for Tigrigna, as well as the description of boxplot elements in Figure 17. Association of Reading Comprehension with Oral Reading Fluency for Tigrigna (Correlation = 0.83)

. The boxplots in our graphs represent distributions of ORF scores (given in words per minute) that are conditional to each level of reading comprehension scores (given in number-correct 0 to 5 points, corresponding to percent-correct scores 0% to 100% in increments of 20%).

For interpretation of EGRA results expressed in percentage of students attaining benchmark levels, special attention should be paid to the statistical relationship between reading comprehension and ORF. To facilitate interpretation, we placed colored horizontal lines that demarcate benchmark levels established in the 2015 workshop (blue horizontal lines represent benchmark cut scores for grade 2, and green horizontal lines represent cut scores for grade 3). Thus, three reading proficiency levels (*full*, *increasing*, and *limited*) are identified based on these ORF cut scores.

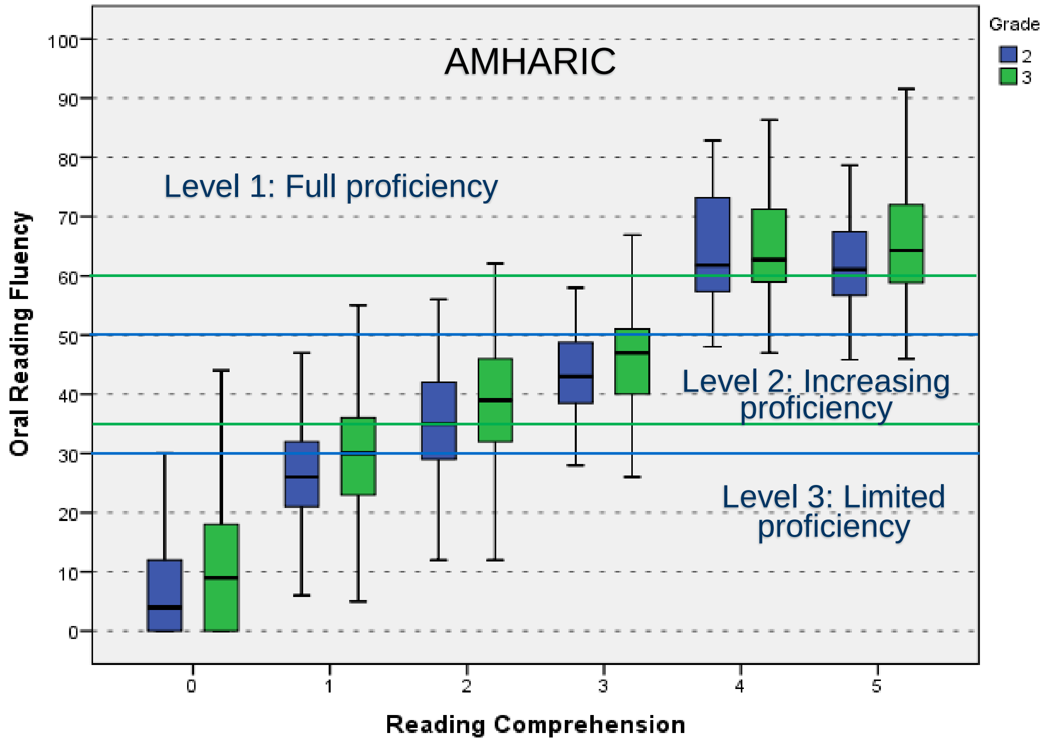
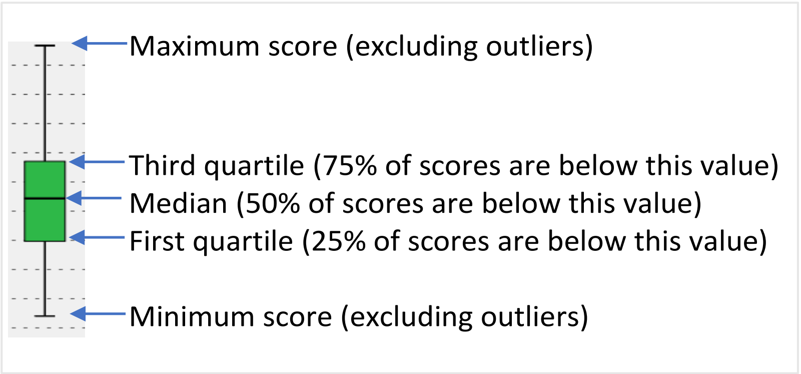


Figure 15. Association of Reading Comprehension with Oral Reading Fluency for Amharic (Correlation = 0.86)


Maximum score (excluding outliers)
Third quartile (75% of scores are below this value)
Median (50% of scores are below this value)
First quartile (25% of scores are below this value)
Minimum score (excluding outliers)

A boxplot is a graphical representation of frequency distribution depicting information about the minimum, first quartile, median, third quartile, and maximum scores.

Figure 16. Meaning of the Elements of a Boxplot

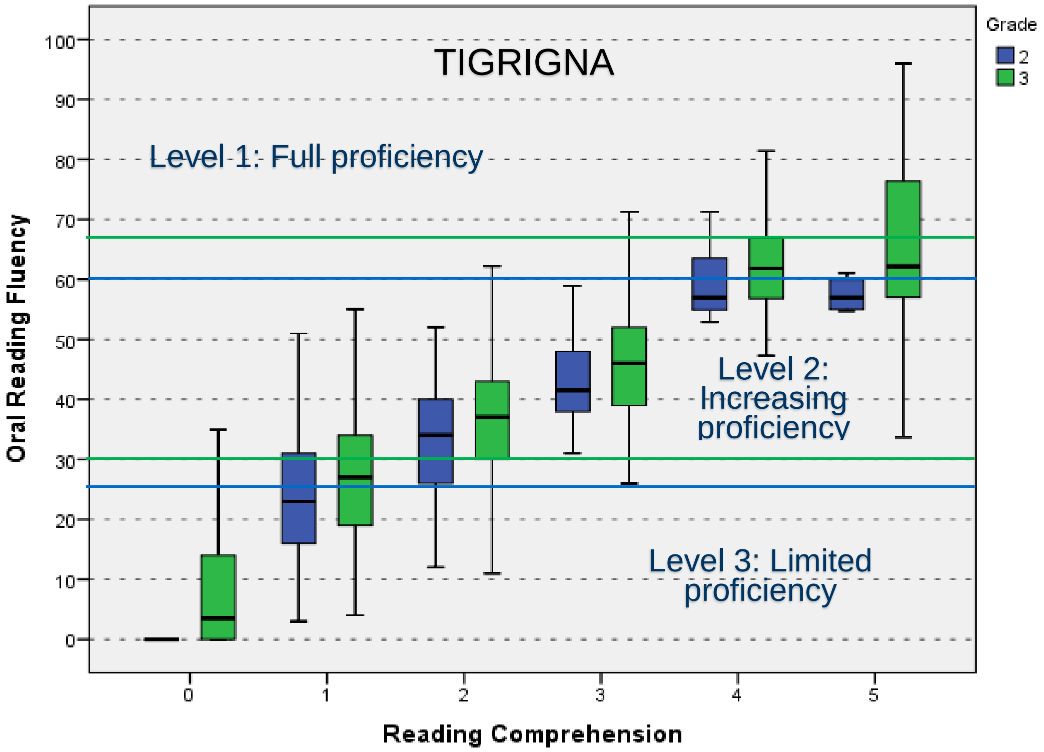


Figure 17. Association of Reading Comprehension with Oral Reading Fluency for Tigrigna (Correlation = 0.83)

Reading benchmarks are sufficiently defined by ORF because the ORF highly correlates with reading comprehension scores (0.87 across all languages), as well as with other EGRA measures of reading fluency (ORF correlates with familiar words reading, 0.92, and with invented words reading, 0.88). Further support to representativeness of ORF is found in the factor analysis studies that consistently determine that all these variables measure the same construct for which the ORF is the strongest marker.

More information about the relationship between the ORF and reading comprehension scores by grade for each language is presented in Appendix 8.

3.2.2 PERCENTAGE OF STUDENTS AT BENCHMARK LEVELS

Based on the proposed benchmarks from the study conducted in 2015, the percentages of students in the 2018 EGRA who reached each performance level were calculated for each language and grade separately. The results presented in

can be used as a basis for defining regional (and national) reading performance targets in future years. Similarly, this information can be used by the MoE to support the development of teachers' capacity to monitor and evaluate student learning progress, as well as by policymakers to better support the implementation of the reading curriculum.

When evaluated against the benchmarks set by Ethiopian experts, the 2018 EGRA results show that grades 2 and 3 students have relatively low reading performance.

A small percentage of students reached the highest benchmark (Level 1), which reflects “reading fluency that enables full or almost full comprehension.” When looking at specific languages and grades assessed in Ethiopia, the percentage of students reaching Level 1 reading proficiency in grade 2 vary between 0.4% for Tigrigna to 8.7% for Amharic. In grade 3, these percentages ranged from 3.6% for Tigrigna to 14.3% for Amharic. When aggregated across all languages assessed by EGRA in Ethiopia, on average 6.2% of students attained reading proficiency that enables full or almost full comprehension. Aggregation across languages using benchmark levels is justified because benchmarks are customized specifically for each language and grade.

However, a larger percentage of students in most languages attained Level 2, which refers to “reading with increasing fluency and comprehension.” The percentages of students attaining this “increasing proficiency” level vary across languages and grades from 6.1% in grade 2 Haddiysa to 48.6% in grade 3 Tigrigna. Overall, based on the 2018 EGRA data, an average of 26.3% (about one-quarter) of students in Ethiopia are attaining reading proficiency at the “increasing” level.

For monitoring purposes, the two levels referring to “full or almost full” and “increasing” reading proficiency may be combined to render a category of reading proficiency that could be tentatively qualified as acceptable or described as readers who can attain either full or partial reading comprehension. The overall percentage of students who fall in these two top levels across all assessed languages and grades in Ethiopia is 32.4%, or about one-third of the student population.

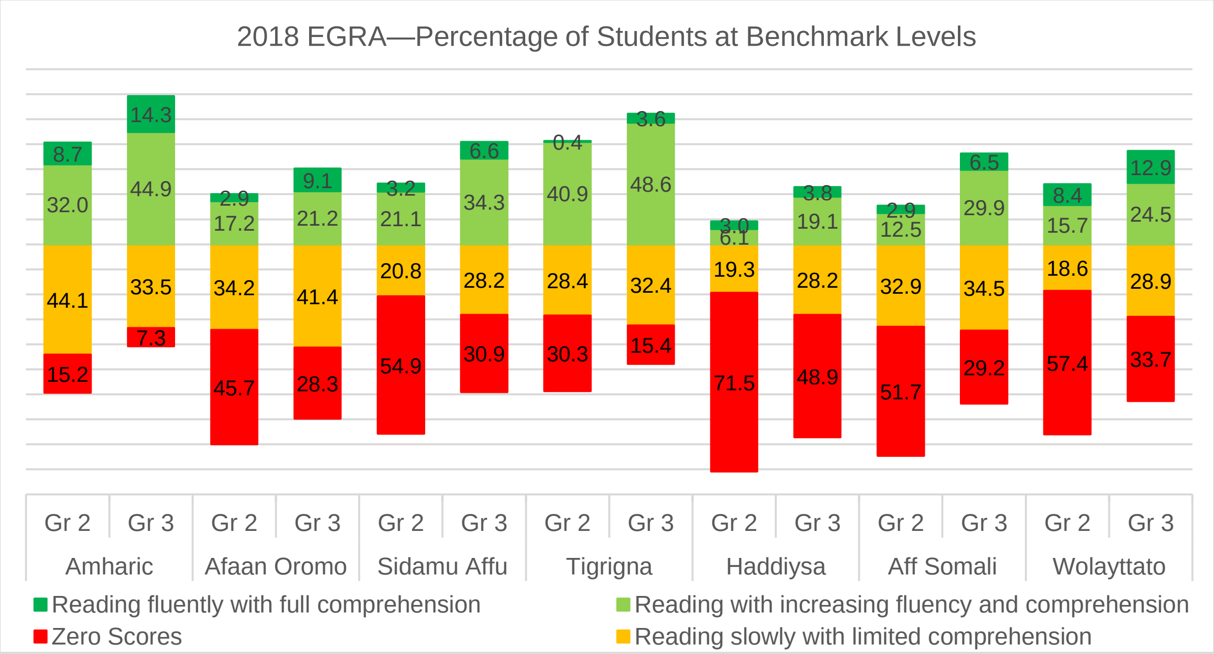


Figure 18. Percentage of Students at Benchmark Levels for Each Language, by Grade

Table 10. Percentage of Students at Benchmark Levels for Each Language, by Grade

Language	Grade	Zero Scores	Reading Slowly with Limited Comprehension	Reading with Increasing Fluency and Comprehension	Reading Fluently with Full Comprehension
Amharic	Gr 2	15.2	44.1	32.0	8.7
	Gr 3	7.3	33.5	44.9	14.3
Afaan Oromo	Gr 2	45.7	34.2	17.2	2.9
	Gr 3	28.3	41.4	21.2	9.1
Sidamu Affu	Gr 2	54.9	20.8	21.1	3.2
	Gr 3	30.9	28.2	34.3	6.6
Tigrigna	Gr 2	30.3	28.4	40.9	0.4
	Gr 3	15.4	32.4	48.6	3.6
Haddiysa	Gr 2	71.5	19.3	6.1	3.0
	Gr 3	48.9	28.2	19.1	3.8
Aff Somali	Gr 2	51.7	32.9	12.5	2.9
	Gr 3	29.2	34.5	29.9	6.5
Wolayttatto	Gr 2	57.4	18.6	15.7	8.4
	Gr 3	33.7	28.9	24.5	12.9

The percentage of students attaining the top two levels is larger in grade 3 (39.9%) than in grade 2 (25%), an increase of 14.9%. This may be interpreted as encouraging evidence, especially considering that benchmarks are designed to be aligned to the grade-level. In other words, even if the same percentage of students reached the benchmark levels, it would still indicate student growth between grades, as defined by the benchmark setters. The fact that a larger percentage of grade 3 students reached the combined benchmarks reflects the fact that growth between grades 2 and 3 is larger than expected by grade-level standards. This interesting finding suggests that reading instruction has a positive effect on progress across grades. On the other hand, it may suggest that the standards for grade-level expectations need to be revised. If the percentages of students attaining desired benchmark level would be about equal in two grades, it would mean children are progressing according to the expectations. Since we found a much higher percentage of students attaining the desired benchmark in grade 3 than in grade 2, it means that either children are progressing above the expectations, or the expectations may not be realistically paced.

3.2.3 PERCENTAGE OF ZERO-SCORES

For the timed EGRA components (letter name recognition, familiar words reading, invented words reading, and oral reading fluency), an auto-stop rule was implemented to discontinue the test if students could not correctly respond to a certain number of items (10 for letters and 5 for words) located at the start of the subtask. This rule was established to relieve stress and to reduce frustration among students. Students to whom the auto-stop rule is applied receive zero scores and are considered non-readers.

The oral reading fluency results reported in the preceding section are complemented by the information about non-reader rates in this section.

is based on the students who could not read at least one word in the ORF subtask. The total height of each bar represents the percentage of students who have zero scores and who therefore are considered non-readers. The largest proportion of non-readers is observed in Haddiysa (71.5% in grade 2 and 48.9% in grade 3), closely followed by Wolayttatto, Sidamu Affo, and Aff Somali languages. It is worrying that the rates of non-readers in those languages are over 50% in grade 2 and still over 30% in grade 3.

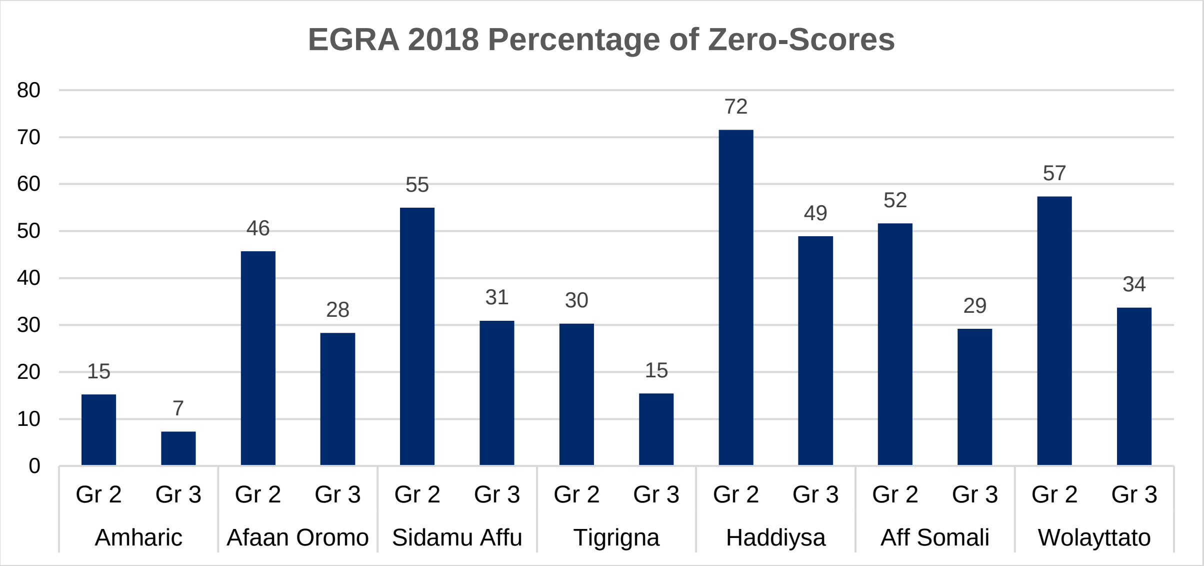


Figure 19. 2018 EGRA Percentage of Zero-Scores in Oral Reading Fluency, by Grade and Language

The lowest rates of non-readers are observed in Amharic (15.2% in grade 2 and 7.3% in grade 3). In all languages, the percentages of non-readers are larger in grade 2, indicating that certain improvement of reading skills happens from grade 2 to grade 3. Aggregated for all languages and grades assessed by the 2018 EGRA, there is an average of 37.2% of students with zero scores in ORF, which represents over one-third of the student population, a large gap that calls for action.

4. COMPARISON OF EGRA PERFORMANCE ACROSS YEARS

This chapter describes comparative results of EGRA performance in Ethiopia across three administration years: 2014 (baseline), 2016 (midline), and 2018 (endline). The comparability between the results attained in these three EGRA administrations is enabled through equating studies. The study for comparability between years 2014 and 2016 is described in the 2016 EGRA report, and the study for comparability between years 2016 and 2018 is described in Chapter 2 of this document. We should still hold some reservations about the comparability between 2014 and 2016 data, because they were obtained by different instrument modes (paper vs. tablets) and collected by different projects using different sampling strategies. On the other hand, comparisons between the data collected by the 2016 EGRA and the 2018 EGRA administrations can be considered reliable.

4.1 COMPARISON OF MEAN SCORES

4.1.1 DIFFERENCES IN TIMED SUBTASKS

Table 11 shows the mean scores on the timed subtasks attained in the midline and the endline administrations by language and grade. In both grade levels, some mean score increases, as well as decreases, are observed across different languages and subtasks, In the text that follows, we focus on interpretation of the ORF scores across the years.

Table 11. Mean Scores for EGRA Timed Tasks in Years 2016 and 2018

Language	Grade	Letter Name Recognition		Familiar Words Reading		Invented Words Reading		Oral Reading Fluency	
		2016	2018	2016	2018	2016	2018	2016	2018
Afaan Oromo	Two	39.3	41.8	10.0	12.8	5.0	4.9	9.8	11.1
	Three	53.9	56.9	19.0	20.9	9.8	9.4	21.2	21.3
Aff Somali	Two	23.6	37.6	6.4	11.6	6.4	10.5	6.4	10.6
	Three	42.4	52.1	14.3	19.0	14.1	17.4	16.5	20.2
Amharic	Two	49.0	29.6	30.1	27.3	21.4	18.6	28.7	24.9
	Three	60.7	42.8	41.0	37.9	27.7	24.9	40.5	38.1
Haddiysa	Two	34.0	28.8	8.8	7.3	7.4	5.2	7.5	5.9
	Three	48.9	46.7	15.6	13.9	12.7	10.6	14.4	12.5
Sidamu Affo	Two	54.8	38.2	16.2	10.7	13.5	7.7	16.3	10.3
	Three	70.0	57.6	25.6	18.5	21.8	14.5	27.1	20.8
Tigrigna	Two	33.9	30.1	23.5	25.3	14.8	11.7	16.4	15.8
	Three	42.8	42.5	37.2	35.9	19.4	16.1	26.2	25.3
Wolayttatto	Two	63.8	32.6	27.7	18.5	24.8	13.5	29.1	8.9
	Three	72.3	47.1	35.7	27.1	33.0	23.5	38.9	18.9

Although the differences between administration years are, in most cases, statistically significant due to large sample sizes, it is important to evaluate the practical educational significance of these differences, which is given by the size measure of Cohen's d.

Focusing on the ORF as the most pertinent timed EGRA subtask, Table 12 presents the ORF mean scores across all three EGRA administrations (2014, 2016, and 2018), which also includes evidence about the size of the differences based on the Cohen's d measure of effect size. The ORF results across three administration years are depicted in Figure 20. Appendix 9 contains all the results of statistical significance testing on ORF between the years 2014 and 2016, between the years 2016 and 2018, and corresponding sizes of difference.

Results for Aff Somali in 2014 and for Wolayttatto in 2016 may not be considered reliable, so the interpretations and comparisons involving these results should be taken with reservation. Therefore, they were dropped from computation of the overall scores in 2014 and 2016, respectively, and they are not presented in Figure 20.

It should be also noted that the fluency scores presented for 7 languages are for descriptive purposes only, not for evaluative comparison across languages, because the word-per-minute units do not have the same meaning in different languages. For the same reason, the overall scores, which represent averages across languages, are presented for descriptive purposes, applying the compensatory model for expressing overall performance at national level.

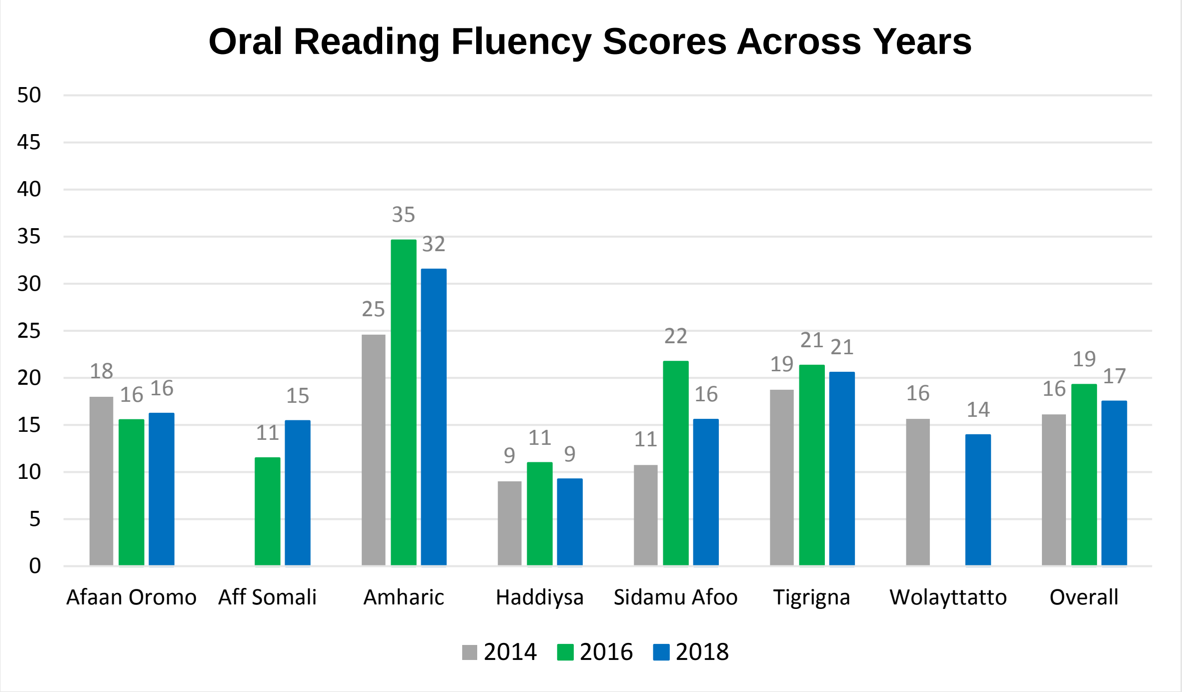
Table 12. Comparison of 2014, 2016, and 2018 Scores in Oral Reading Fluency

Language	Grade 2					Grade 3				
	2014	2016	2018	D 14→16	D 16→18	2014	2016	2018	D 14→16	D 16→18
Afaan Oromo	12.1	9.8	11.6	-.12	0.09	23.9	21.2	21.3	-.14	0
Aff Somali	20.4	6.4	10.6	-.89**	0.28*	32	16.5	20.2	-.98**	0.19
Amharic	19.2	28.7	24.9	.49*	-0.21	30	40.5	38.1	.54**	-0.11
Haddiysa	6.5	7.5	5.9	.06	-0.12	11.5	14.4	12.5	.17	-0.10
Sidamu Affo	7.1	16.3	10.3	.45*	-0.32*	14.4	27.1	20.8	.62**	-0.31*
Tigrigna	13.3	16.4	15.8	.17	-0.04	24.2	26.2	25.3	.11	-0.05
Wolayttatto	11.2	29.1	8.9	.85**	-0.87**	20.1	38.9	18.9	.81**	-0.82**
Overall	11.6	14.2	12.5			20.7	24.3	22.4		

Notes. * indicates that the size of difference is educationally significant.
** indicates a strong educational effect, something substantially changed (Wolf, 1986).

The grade 2 ORF scores between the baseline (2014) and midline (2016) EGRA administrations show substantial increases in three languages and a substantial decrease in one language, whereas the differences in three languages were practically marginal or negligible. However, when looking at changes of grade 2 ORF scores between the 2016 and 2018 administrations, it can be noted that there was a substantial decrease in two languages (Sidamu Affo and Wolayttatto) and decreases in four languages were either negligible or marginal, whereas a practically significant increase was observed in only one language (Aff Somali).

When looking at grade 3 trends between years, a similar pattern can be observed. Change in ORF scores from 2014 to 2016 is characterized by substantial increases in three languages, decrease in one language, and practically negligible differences in three languages. The pattern of differences of grade 3 ORF scores between the 2016 EGRA and the 2018 EGRA administrations is almost the same as in grade 2. There was a significant decrease in two languages (Wolayttatto and Sidamu Affo), a marginal increase in one language (Aff Somali), and negligible differences in four other languages.



Note. Overall average ORF scores across languages are for descriptive purposes because the ORF units (words per minute) may have different meaning imposed by the nature of each language.

Figure 20. Trend of Oral Reading Fluency Scores Across Years 2014, 2016, and 2018

The overall trend of ORF scores across years in different languages shows diverse patterns, with few substantial ups and downs, but in most cases with differences of relatively small and practically negligible size. When looking at the overall results averaged over the languages, it can be concluded that the changes in student ORF scores at an aggregated national level are very small and below what can be considered as practical significance. From 2014 to 2016, there was an average increase of 3 wpm, and from 2016 to 2018 there was an average decrease of only 2 wpm.

4.1.1 DIFFERENCES IN UNTIMED SUBTASKS

Table 13 shows comparisons of the mean scores in untimed subtasks (initial letter sound, listening comprehension, and reading comprehension) for the 2016 EGRA and the 2018 EGRA administrations for all assessed languages and grades. In most cases, small drops can be observed, and in a few cases, the mean scores increased. Further in the text, the focus of discussion is given to reading comprehension as the most pertinent EGRA untimed subtask.

Table 13. Mean Scores for EGRA Untimed Tasks in Years 2016 and 2018

Language	Grade	Initial Letter Sound		Listening Comprehension		Reading Comprehension	
		2016	2018	2016	2018	2016	2018
Afaan Oromo	Two	47.3	47.2	81.7	60.4	10.5	13.0
	Three	62.4	61.1	85.6	71.0	27.6	28.5
Aff Somali	Two	34.3	83.2	56.6	77.1	6.5	10.1
	Three	57.9	89.3	63.4	81.9	20.0	21.2
Amharic	Two	86.8	81.7	63.4	66.2	25.9	24.8
	Three	91.2	85.7	71.2	72.7	42.6	41.9
Haddiysa	Two	79.3	85.1	80.3	73.5	8.6	12.9
	Three	85.0	89.4	83.4	81.0	16.2	20.3
Sidamu Affo	Two	97.2	90.5	80.6	82.9	17.4	9.2
	Three	98.2	94.2	84.5	87.8	32.9	20.1
Tigrigna	Two	77.7	72.6	71.7	46.3	13.8	17.6
	Three	87.5	83.5	76.9	62.6	25.5	26.3
Wolayttatto	Two	73.9	74.1	77.9	52.5	41.4	11.4
	Three	78.4	75.1	80.3	54.4	53.1	21.4

Notes. ILS is initial letter sound; LC is listening comprehension; RC is reading comprehension.

The mean scores in reading comprehension obtained in all three EGRA administrations (2014, 2016, and 2018) are presented in Table 14, along with the evaluation of practical significance of differences between years given in terms of Cohen’s d measure of effect size. The results on the reading comprehension subtask across years are also depicted in Figure 21. All the results of statistical significance testing between years 2014 and 2016, as well as between 2016 and 2018 on reading comprehension (using t-tests), along with corresponding sizes of difference, are presented in Appendix 10.

Table 14. Comparison of 2014, 2016, and 2018 Scores in Reading Comprehension

Language	Grade 2					Grade 3				
	2014	2016	2018	D 16-14	D 18-16	2014	2016	2018	D 16-14	D 18-14
Afaan Oromo	14.0	10.5	13.0	-0.17	0.12	28.0	27.6	28.5	-0.01	0.03
Aff Somali	28.0	6.5	10.1	-1.33**	0.22*	44.0	20.0	21.2	-0.87**	0.04
Amharic	18.0	25.9	24.8	0.30*	-0.04	32.0	42.6	41.9	0.33*	-0.02
Haddiysa	12.0	8.6	12.9	-0.17	0.23*	22.0	16.2	20.3	-0.22	0.17
Sidamu Affo	20.0	17.4	9.2	-0.11	-0.38*	20.0	32.9	20.1	0.45*	-0.47*
Tigrigna	14.0	13.8	17.6	-0.01	0.21*	22.0	25.5	26.3	0.14	0.03
Wolayttatto	24.0	41.4	11.4	0.54**	-0.93**	40.0	53.1	21.4	0.44*	-0.84**
Overall	17.0	13.8	14.1			27.3	27.5	25.7		

Notes. * indicates that the size of difference is educationally significant.
** indicates a strong educational effect; something substantially changed (Wolf, 1986).

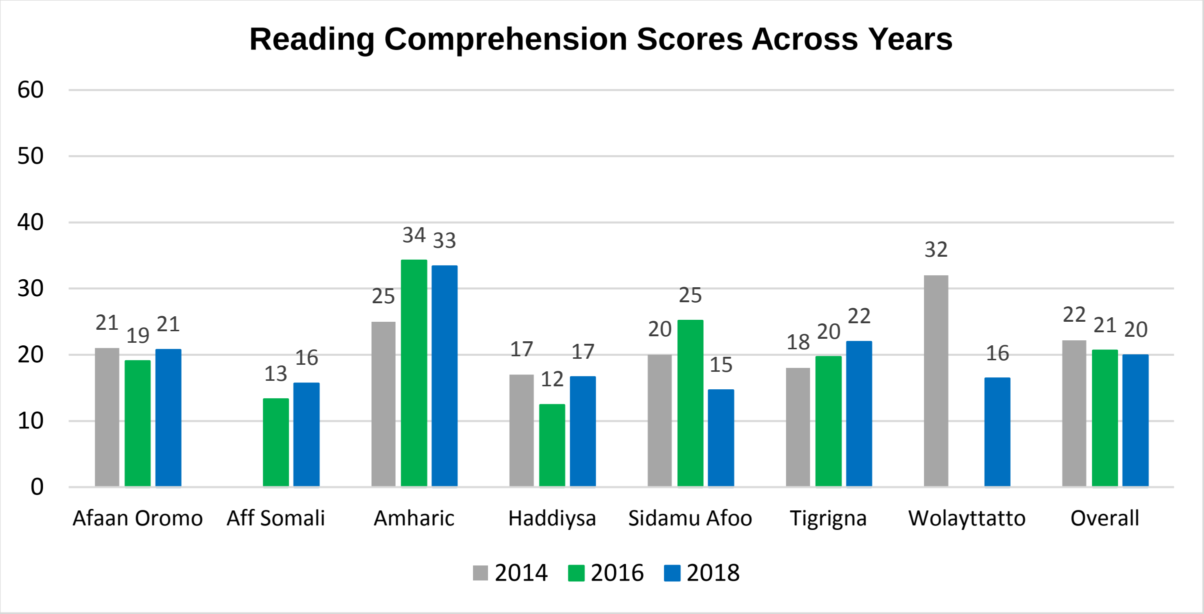


Figure 21. Trend of Reading Comprehension Scores Across Years 2014, 2016, and 2018

Similarly, as for ORF, educationally significant score gains in reading comprehension were observed between 2014 and 2016 in Wolayttatto and Amharic in both grades, and in Sidamu Affo in grade 3. However, a major decline between 2014 and 2016 was observed in Aff Somali in both grade levels. In the other three languages, the differences were negligible or too small to bear practical significance.

Regarding the differences between the 2016 and 2018 EGRA administrations, a major decline is observed in two languages (Sidamu and Wolayttatto) in both grades, there was a marginal increase in three languages (Somali, Haddiysa, and Tigrigna) in grade 2, and all other differences were practically negligible.

4.2 COMPARISON ACROSS YEARS IN ATTAINING BENCHMARK LEVELS

Results expressed in percentages of students attaining specific benchmark levels are available for the 2014 EGRA (baseline), the 2016 EGRA (midline), and the 2018 EGRA (endline) administrations. The comparative results obtained in these three administrations for all languages, showing the percentages of students in each proficiency level combined across both grades, are depicted in Figure 22.

As suggested in Section 3.2.2 of this document, it is meaningful to evaluate a joint attainment of Level 1 and Level 2 (full proficiency and increasing proficiency).For the sake of convenience and to facilitate comparisons, those two levels are positioned in the graph above the reference line, and two lower levels (limited proficiency and nonreaders) are located below the reference line.

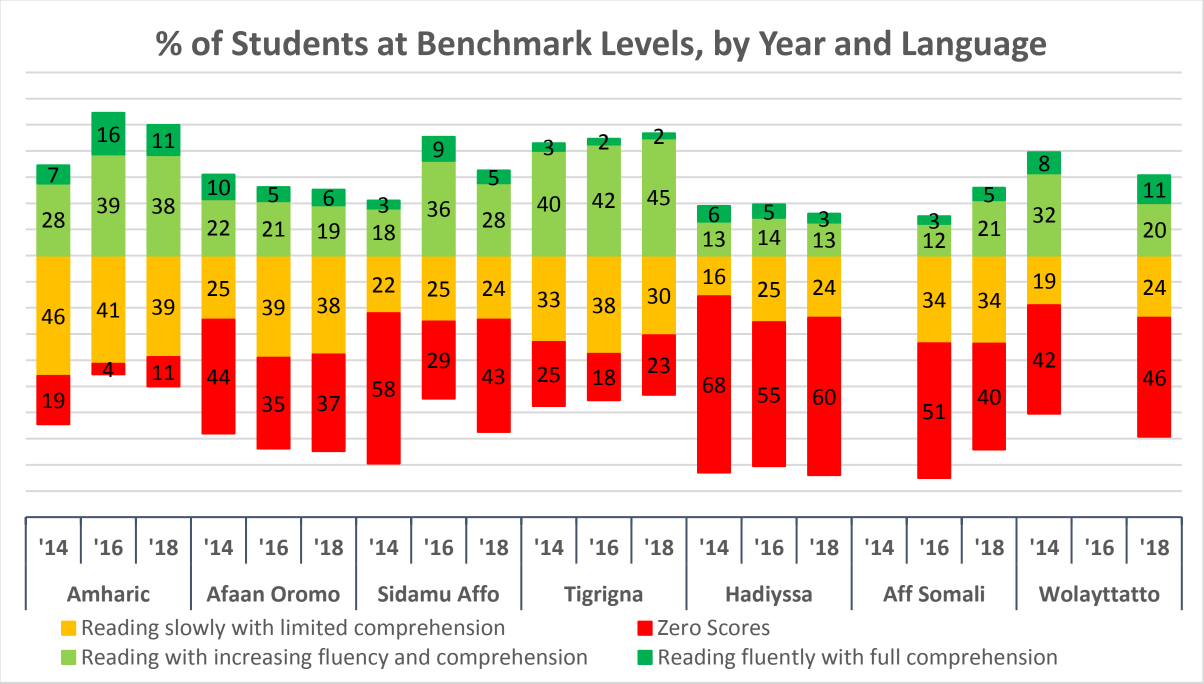


Figure 22. Percentage of Students at Benchmark Levels for Each Language, by Year

As Figure 22 illustrates, when comparing reading performance between years 2016 and 2018, a relatively steep drop in percentage of students attaining full or increasing reading proficiency was observed in Sidamu Affo; 12% less students reached this benchmark in 2018 than in 2016.

In the other languages, the changes in the percentage of students attaining the two upper levels of reading proficiency are rather small—practically negligible—though it is worth noting that in Aff Somali there was an increase of 11% of students reaching this referenced benchmark.

To facilitate comparisons between years at the overall Ethiopian level, the performance of students was aggregated across languages. This aggregation is justified when using the performance standards (benchmarks) that are set based on the common conceptual definitions of performance levels and the customized operational definitions (cut scores) for each language. In other words, the EGRA benchmarks levels have the same meaning in each language, but the cut scores were custom set to reflect specific characteristics of each language. Thus, we can use the aggregated percentages of students performing at each benchmark level as indicators of overall performance at the Ethiopian level.

Figure 23 shows the performance of students within each year and grade aggregated across languages to represent the overall performance at the Ethiopian level. When looking at the differences between grades within each EGRA year, it can be observed that the percentage of students attaining the upper two benchmark levels is consistently higher in grade 3 than in grade 2, as would be expected.

Considering that the benchmarks are set higher in grade 3 than in grade 2 to reflect different expectations for these two grade levels, this finding indicates that the growth of reading performance between grades is steeper than expected by benchmarks. It is not quite clear whether this should be interpreted as a good indicator of instruction being more effective than expected, or simply that the benchmark expectations were not set accurately to reflect the difference between the grade levels.

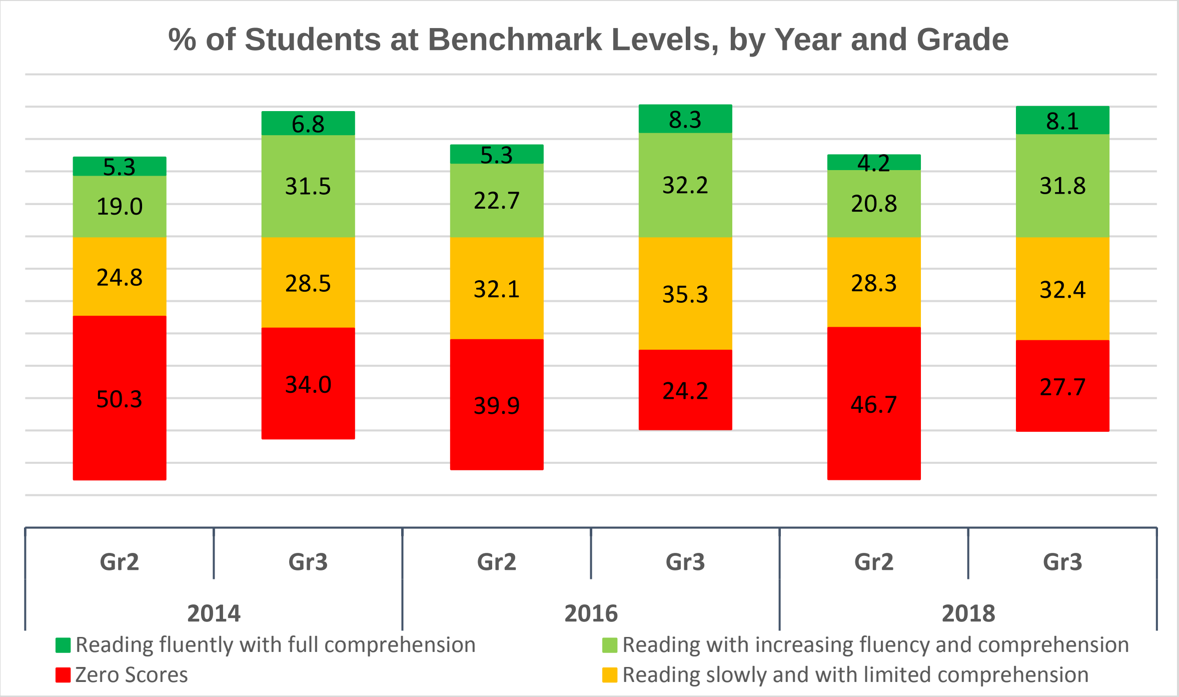


Figure 23. Overall Percentage of Students at Benchmark Levels, by Year and Grade

To enable direct insights into the overall trend in reading performance across years in Ethiopia, the percentage of students attaining each benchmark level were aggregated across languages and grades (Figure 24). It can be observed that there is a little variation in student reading performance over the three EGRA administrations in Ethiopia. The percentage of students reaching the two upper benchmark levels in reading (full and increasing fluency and comprehension) was 31.4% in year 2014; then it rose to 34.2% in year 2016 and then decreased to 32.5% in year 2018. Therefore, fluency and comprehension levels have remained relatively constant over these periods as these differences are small and not bearing practical significance.

This finding suggests that reading performance in primary grades in Ethiopia is still inert, not showing progress that one would hope for considering multiple years of reading intervention.

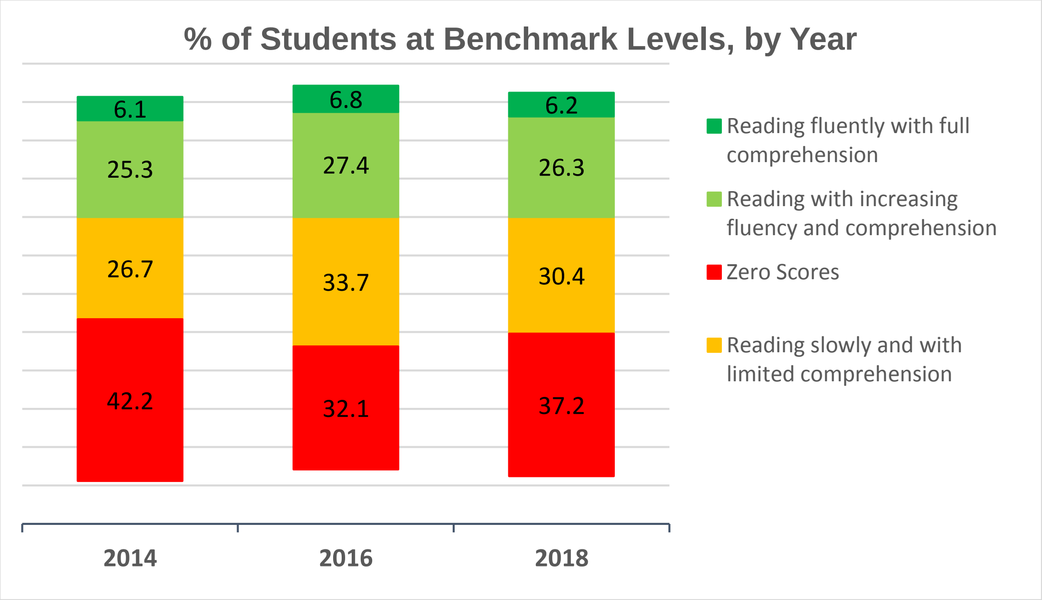


Figure 24. Overall Percentage of Students at Benchmark Levels, by Year

5. FACTORS ASSOCIATED WITH STUDENT PERFORMANCE

The following sections present findings on the associations between performance on the EGRA ORF subtask and various contextual factors assessed by means of the student, teacher, and director background questionnaires. These factors include, but may not be limited to, access to teaching and learning materials (mother tongue textbooks, teacher guides, supplementary reading materials, etc.), school support activities (supervising and mentoring, policy support, etc.), teachers' and principals' characteristics and activities (education level, teaching experience, monitoring student performance, etc.), and school environment factors (facilities, resources, etc.).

READ M&E analyzed the background factors in relation to performance in the ORF subtask. The ORF subtask is relevant to the reading comprehension subtask because there is a high correlation between those two measures that places them under the same dimension of reading. These analyses were carried out by comparing the ORF scores related to the respondents' answers to questions. For example, based on the question in the directors' questionnaire, *Have you received training to support MT teachers?* directors were divided into two groups: those who answered "Yes" and those who answered "No." Then, the mean student scores on the ORF subtask were computed for the two groups of schools corresponding to these two groups of directors. Finally, the difference between mean student scores on ORF for the two groups of schools was calculated, the statistical significance of this difference was tested, and the associated effect sizes were calculated using Cohen's d measure.

We used a similar approach for questions that have more than two options. One of the responses was taken as a reference—typically the one representing the lowest resource addressed by the question—and the rest of the options were compared against this reference. Appendix 10 contains bar charts for each analyzed background question to show the difference in ORF scores between the groups of respondents choosing different options, along with the corresponding statistical significances and effect sizes.

The contextual variables for teachers and directors were grouped into three categories: teachers' and directors' personal characteristics, their pedagogical activities, and school environment or resource factors. On the other hand, the contextual variables for students were grouped as student personal characteristics, home environment, and school environment.

For the analyses of all the contextual variables, data from both grade 2 and grade 3 were combined because associations between reading performance and background factors were not likely to change from one year to another. For example, the associations between ORF scores and variables indicating who helps at home, or availability of reading resources were not likely to change between grades 2 and 3 to the extent that our models would be able to detect any systemic differences. Similarly, data for each of the languages were combined under the assumption that background factors affecting student performance in one language will have the same effect in a different language. Preliminary analyses of associations between background factors and student performance in the ORF subtask, disaggregated by grade and language, confirmed these assumptions. For the analysis of contextual factors, READ M&E utilized the entire sample of schools assessed by the 2018 EGRA (a total of 459 schools).

As an important note for interpretation of the results, the reader should bear in mind that these analyses report the ***associations between reading performance and contextual variables***, which does not provide sufficient information for causal attribution, and these contextual variables ***should not be viewed as factors that influence*** reading achievement. The associations between two variables—let's say *X* and *Y*—may be due to *X* causing *Y*, or *Y* causing *X*, or some third factor may be affecting both variables, making them correlated with each other. Thus, causal interpretation of the associations between contextual variables and reading performance requires additional scrutiny and understanding of a myriad of circumstances that could contribute to associations between the analyzed variables.

In selecting significant associations, we used both statistical significance ($p < 0.05$) and practical significance (Cohen's d greater than 0.25). The tables in this section show the difference between the mean student scores of the two groups and the corresponding values of the statistical significance (t-test) and practical significance (Cohen's d).

5.1 DIRECTOR QUESTIONNAIRE

Four hundred and fifty-nine school directors responded to questionnaires that contain questions about themselves and mother tongue instruction in their schools. Of this number, 92% are male and 8% are female. Over half of the respondents (69%) reported holding a bachelor's degree; 2%, a master's degree; and the rest (29%), a diploma.

To review the associations between the contextual variables assessed by the directors' questionnaire and reading performance of students in their schools, this section presents selected background variables based on their content relevance and ***statistical or practical significance***.

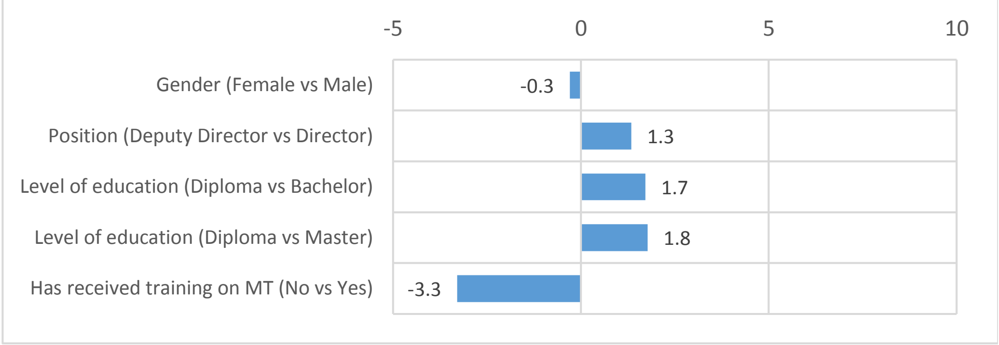
5.1.1 DIRECTOR CHARACTERISTICS

Among the directors' characteristics that were analyzed for potential association with reading performance in their schools were gender, position, level of education, and whether they received training on how to support mother tongue teachers.

The results of the analysis show that level of education and position at the school are positively associated with reading performance. Those respondents that identified themselves as Directors are associated with higher student performance than those that identified themselves as Deputy Directors. Regarding level of education, holding a degree higher than Diploma yielded a small but significant association with their students' reading performance.

Additionally, the results show that a director's gender is not significantly related to reading performance in their schools, likely because of the gender imbalance in the number of respondents. Interestingly, Directors that reported not receiving training on monitoring and supporting mother tongue teachers (about 71%) showed a statistically significant association with reading performance, possibly because of the large proportion of respondents in this group. The results of these analyses are presented in Table 15.

Table 15. Directors' Characteristics Associated With Reading Performance



Sig	D
0.62	-0.01
0.00	0.07
0.00	0.08
0.10	0.09
0.00	-0.16

5.1.2 DIRECTOR ACTIVITIES

Students' performance on ORF tasks was analyzed in relation to directors' responses about the activities they perform personally and activities of other school officials under the director's management.

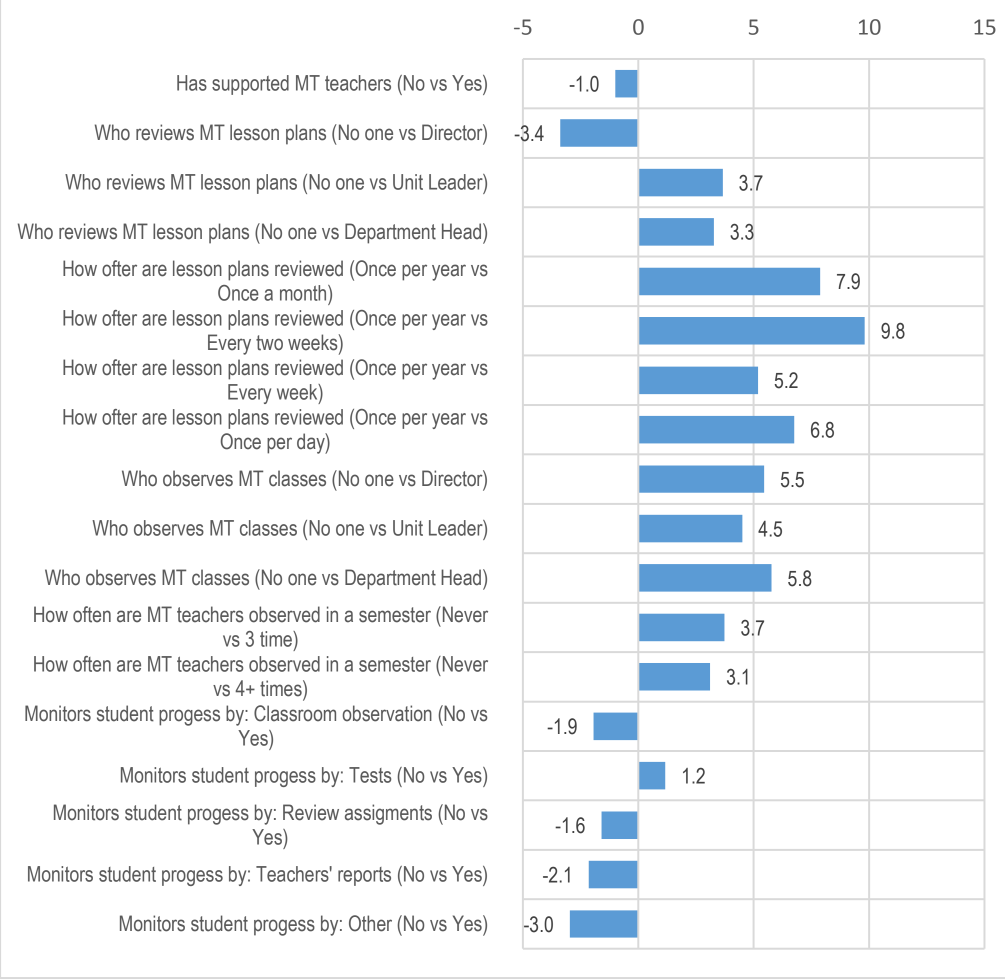
The results of the analysis indicate that there are several activities that are significantly and positively associated with student learning outcomes. For example, activities of other school officials (Deputy Directors, Unit Leaders, Department Heads) that are managed by school directors (thus, under director's responsibility) appeared to be significantly related to reading performance of students. These significant activities are: the existence of a person responsible for reviewing the mother tongue lesson plans, the frequency of the review of the mother tongue lesson plans, the existence of a person responsible for observing mother tongue teachers while teaching, and the frequency at which the mother tongue teachers are observed (see Table 16 below).

The existence of a person to review the mother tongue lesson plans is positively associated with student reading performance, only if a Unit Leader or Department Head performs this task. Also, reviewing lesson plans once a month or more often shows a strong positive association with reading performance. Classroom observation as an activity is also statistically significant when performed by a Director, Unit Leader or Department Head.

We also analyzed a range of activities that directors personally implement, specifically: supporting mother tongue teachers, and monitoring student progress by various techniques such as: classroom observation, reviewing test results, oral evaluation, reviewing students' assignments, and reviewing progress reports provided by teachers. In schools where directors reported that they have not supported mother tongue teachers (about 22%) a small statistically significant increase in reading performance was observed when compared to those schools where directors reported they have supported mother tongue teachers. This finding implies that when mother tongue teachers are effective they may not need directors' support.

Regarding the different techniques used by directors to monitor the progress of students, the only technique positively associated with reading performance is reviewing test results. Other techniques that are statistically significant but negatively associated with reading performance are: classroom observation, reviewing students' assignments, and reviewing progress reports provided by teachers.

Table 16. Directors' Activities Associated with Reading Performance

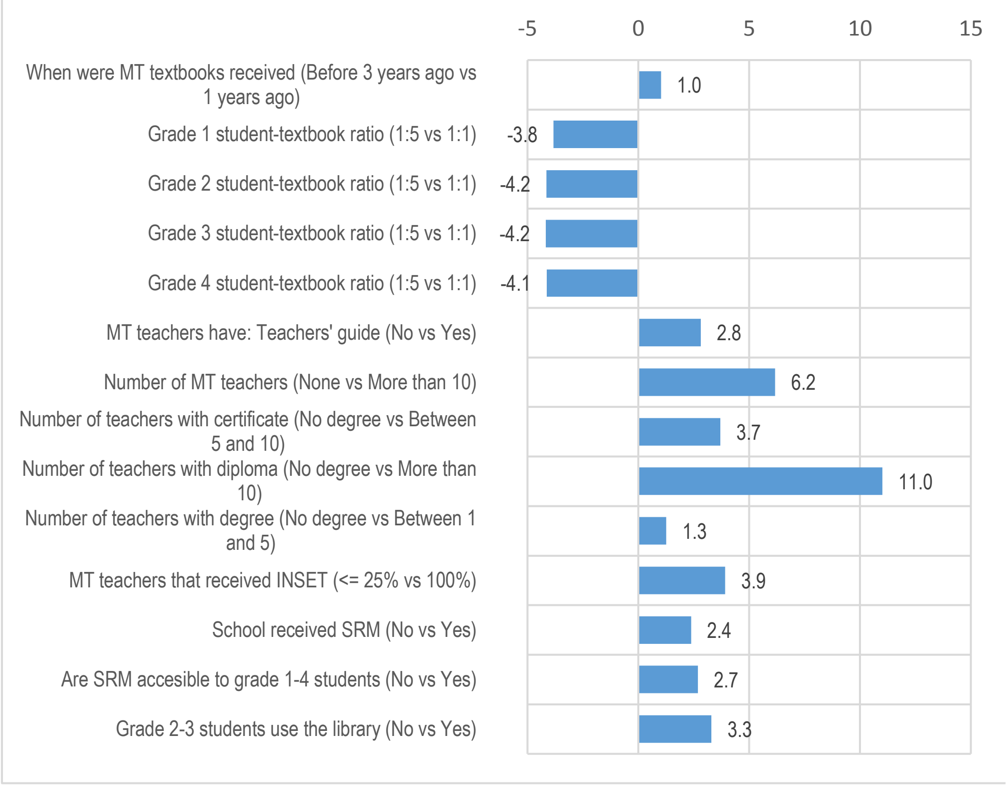


Sig	D
0.01	-0.05
0.04	-0.17
0.04	0.18
0.04	0.16
0.00	0.41
0.00	0.48
0.00	0.26
0.00	0.33
0.00	0.27
0.01	0.23
0.00	0.28
0.01	0.18
0.02	0.15
0.00	-0.10
0.00	0.06
0.00	-0.08
0.00	-0.11
0.00	-0.15

5.1.3 SCHOOL RESOURCES REPORTED BY DIRECTORS

We assessed the availability of school resources by administering questionnaires to directors. Results that yielded either statistically or practically significant associations with reading performance are presented in Table 17 below. These significant resources were: when the school received mother tongue textbooks, the ratio between mother tongue textbooks and students in grades 1–4, the availability of teachers' guides for mother tongue teachers, number of mother tongue teachers at the school, educational qualification of teachers, percentage of mother tongue teachers that have received in-service training, availability of supplementary reading materials, and grade 2 and grade 3 students making use of the school library.

Table 17. School Resources (Reported by Directors) Associated with Reading Performance



Sig	D
0.03	0.05
0.00	-0.19
0.00	-0.20
0.00	-0.20
0.00	-0.20
0.00	0.14
0.00	0.33
0.01	0.18
0.00	0.59
0.00	0.06
0.00	0.19
0.00	0.12
0.00	0.13
0.00	0.16

The schools that received mother tongue textbooks within the past year show increased student performance compared to those that received mother tongue textbooks 2 or more years ago. The availability of a mother tongue teachers' guide for teachers to use also was significantly positively associated with student reading performance, which is fully aligned with expectations that informed the READ TA intervention.

Regarding the number of mother tongue teachers in schools, when directors reported they have more than 10 mother tongue teachers, a practically significant difference in reading performance can be observed compared to schools with none. In addition, teachers' qualifications are significantly related to student reading performance in schools. Based on directors' responses about the number of teachers with qualifications, there was a strong positive association between number of teachers with a diploma or degree and student reading performance in their schools. Another finding is that, when 100% of mother tongue teachers receive in-service training, their schools see a positive association with student performance in reading comprehension when compared to schools that have a lesser percentage of teachers receiving training.

And finally, when directors report that their school received supplementary reading materials, these materials are available to students, and grade 2 and grade 3 students are using the school library or reading room, there was a positive association of these factors with higher performance in reading in corresponding schools.

A higher availability of textbooks for students (based on the textbook-to-student ratio) unexpectedly appeared to be negatively associated with performance in reading. In the schools where directors reported the textbook-student ratio was 1:5 show higher reading performance of students than the schools with ratios of 1:1. This unexpected result is likely due to respondents' or enumerators' misunderstanding of the question's meaning.

5.2 TEACHER QUESTIONNAIRE

Eight hundred fifty-three teachers responded to questionnaires that contain questions about themselves and the instruction of mother tongue in their schools. The distribution by gender is relatively balanced: 43% are male and 57% are female. The highest level of education reported by 87% of the respondents is Diploma, and the average years of service is 10 years.

5.2.1 TEACHER CHARACTERISTICS

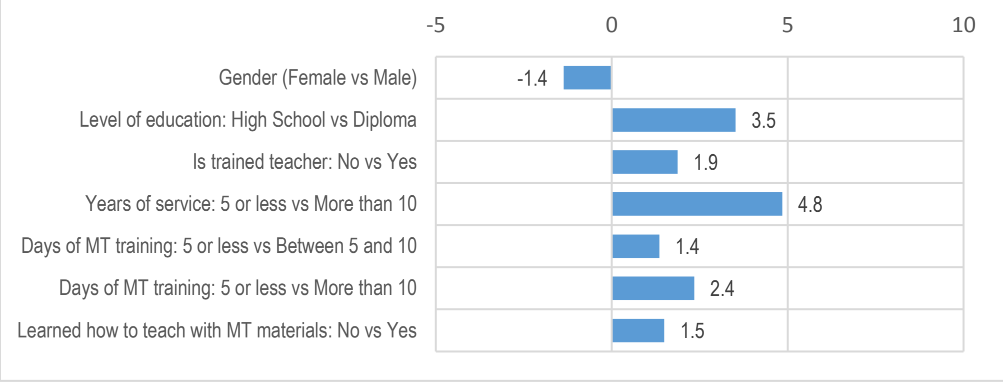
The teachers' characteristics associated with reading performance are gender, amount of training, level of education, years of service as a teacher, whether he or she has received training on mother tongue materials, length of the training, and if they know how to teach using the mother tongue materials, as shown in Table 18.

Schools where teachers have levels of education above a high school diploma showed increased performance in reading. Respondents who identified themselves as trained teachers (meaning that the focus of their degree is teaching) are associated with significantly higher reading performance in their schools; it can also be noted that longer trainings on the mother tongue curriculum (more than 10 sessions) are also positively associated with reading performance.

In terms of years of service, teachers who report having 10 or more years of service are associated with a higher reading performance for their students. Teachers who report to know how to teach with the new mother tongue materials are significantly associated with increased performance in reading.

Male teachers are associated with significantly lower reading performance in their schools than female teachers. This difference should not be interpreted as evidence that male teachers may be less effective in teaching reading than female teachers. In fact, the difference may be due to the fact that male teachers are more likely to serve in remote rural areas with greater poverty, availability of resources, and low student performance.

Table 18. Teachers' Characteristics Associated with Reading Performance



Sig	D
0.00	-0.07
0.01	0.17
0.00	0.09
0.00	0.24
0.00	0.07
0.00	0.12

5.2.2 TEACHER ACTIVITIES

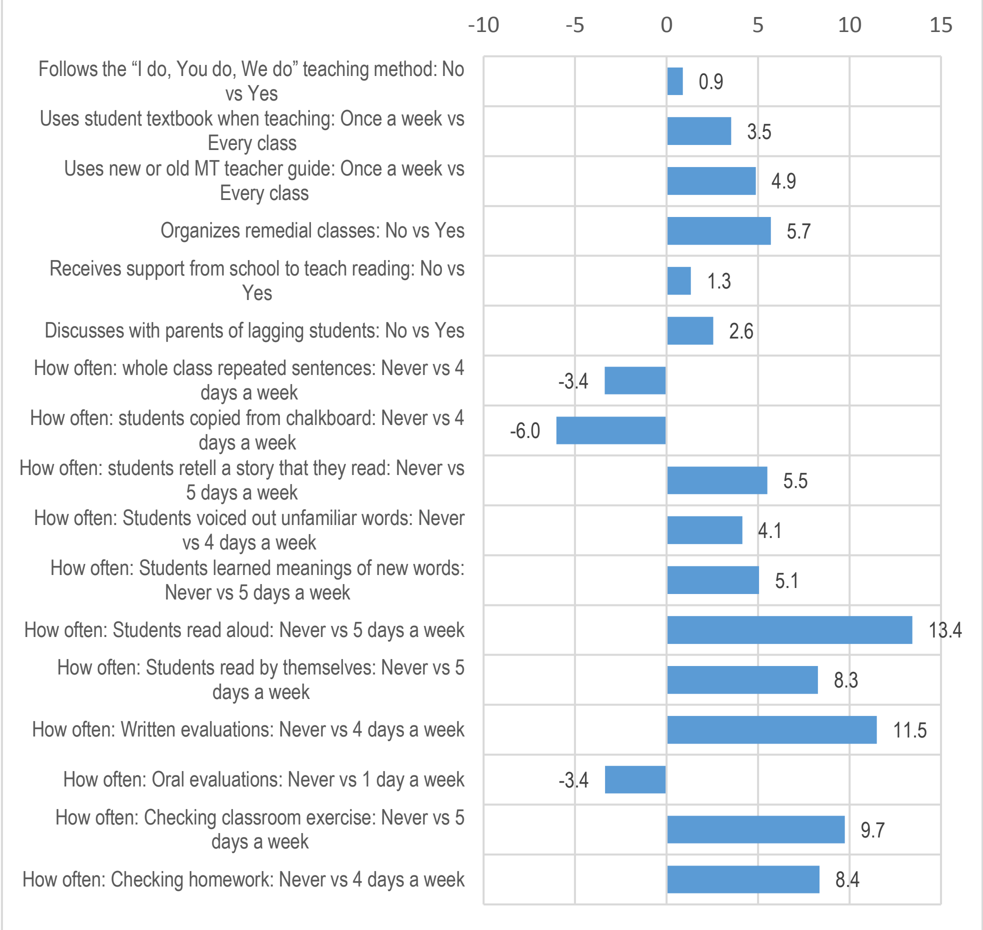
The teachers' activities that are significantly associated with reading performance in their schools are listed in Table 19, which shows the activities with the highest differences in means and the highest effect sizes.

Teachers that reported using the "I do, you do, we do" teaching method and the student textbook and the teacher guide every time they teach are associated with higher performance in reading. Organizing remedial classes for students who are lagging is also associated with higher performance in reading. Not surprisingly, receiving support from the school and discussing with parents when a student is lagging are also positively associated with reading performance.

Teachers were asked about how frequently they perform different activities with their students during the mother tongue class. Their responses were based on the past five school days. The most frequent activity was having students copy text from the chalkboard while the least frequent activity was having students sound out unfamiliar words. When looking at the effect of classroom activities based on their frequency, those that are positively associated with higher performance in reading are retelling a story they read, sounding out unfamiliar words, learning the meanings of new words, reading aloud, and reading by themselves.

Teachers were also asked about the different methods they used to monitor students' reading progress. Oral evaluation is the most frequently used, followed by checking classroom exercises. The methods that yield the highest effect sizes on reading performance are in-classroom written evaluations of student performance four days a week, assigning and checking classroom exercises five days a week, and assigning and checking homework four days a week.

Table 19. Teachers' Activities Associated with Reading Performance

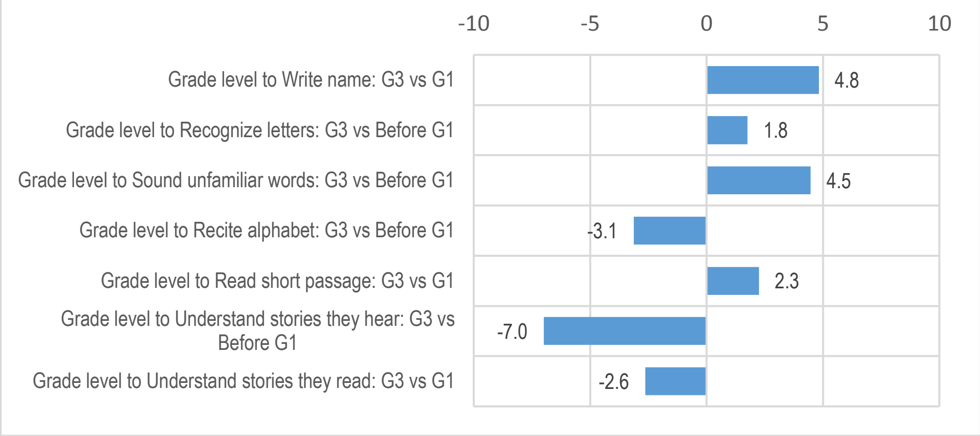


Sig	D
0.03	0.04
0.00	0.17
0.00	0.24
0.00	0.28
0.00	0.07
0.00	0.13
0.00	-0.16
0.00	-0.30
0.00	0.27
0.00	0.20
0.00	0.25
0.00	0.69
0.00	0.42
0.00	0.58
0.00	-0.17
0.00	0.48
0.16	0.37

Teacher activities that are not positively associated with reading performance are having students repeat sentences said by the teacher and having students copy from the chalkboard. In terms of techniques to monitor student progress, oral evaluations do not positively affect reading scores.

Table 20 shows the associations of teachers' responses on the question, *at what grade level should students first be able to demonstrate the listed reading skills?*

Table 20. Teachers' Expectations of Students' Performance



Sig	D
0.00	0.24
0.03	0.09
0.00	0.22
0.00	-0.15
0.00	0.11
0.00	-0.33
0.00	-0.13

Teachers having realistic expectations about the grade level in which students should first demonstrate different reading skills may be a contributing factor to increased reading performance. Teachers who expect students to write their name, recognize letters, sound out unfamiliar words, and read short passages with few mistakes before grade 1 or at grade 1 are positively associated with reading performance of their schools. On the other hand, teachers who expect students to demonstrate skills such as reciting the alphabet, understanding stories they hear, and understanding stories they read before or at grade 1 are negatively associated with reading performance.

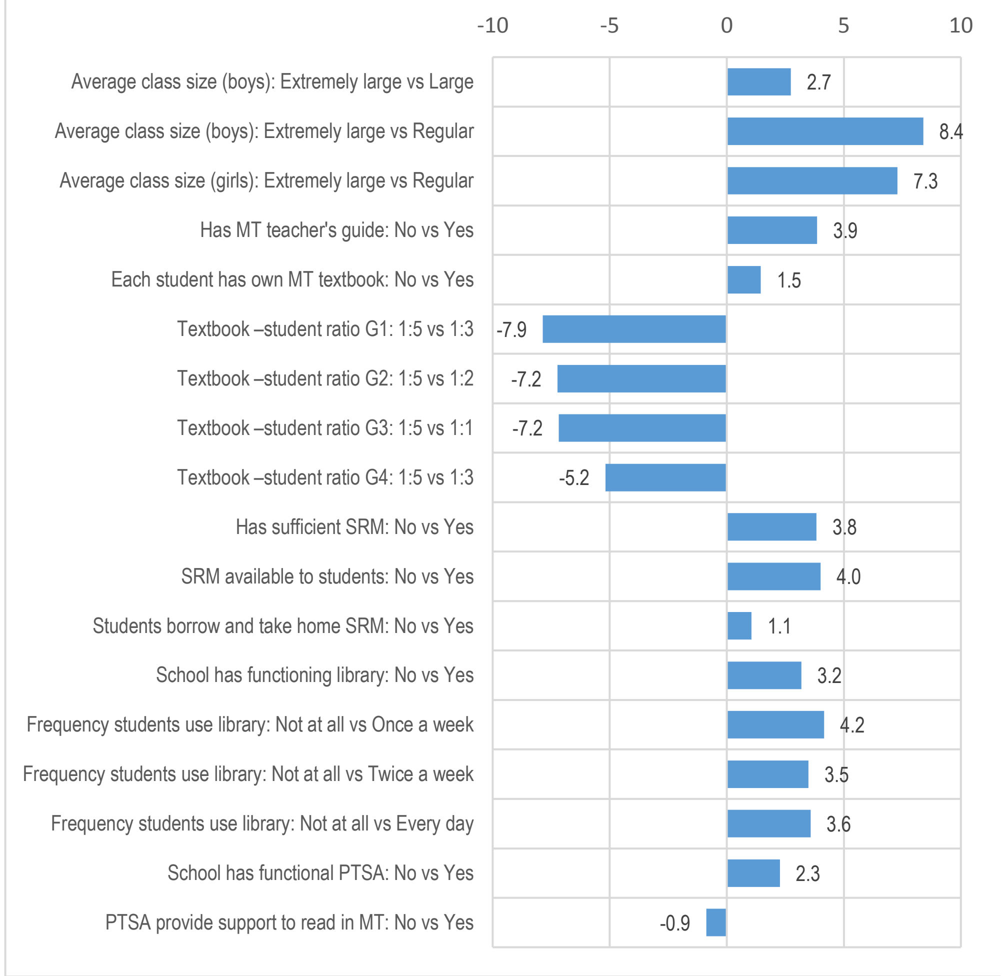
5.2.3 SCHOOL RESOURCES REPORTED BY TEACHERS

Teacher responses on school resources that are significantly associated with reading performance are presented in Table 21. The variables associated with student reading performance are average class size, availability of the teacher guide, availability of student textbooks, availability of supplementary reading materials, availability of a school library, use of the library by the students, and existence of a functional PTSA at the school.

Having regular (up to 30 students) or large (between 31 and 60 students) class sizes is associated with better reading performance than having extremely large class sizes (more than 61 students). The availability of grade-appropriate reading materials (mother tongue teachers' guide, student mother tongue textbooks, and supplementary reading materials), existence of a school library and reading corner, and use of the library or reading corner are all positively associated with student performance in reading.

As is the case with directors, a higher textbook-to-student ratio is negatively associated with reading performance for grades 1–4. For example, having one textbook for every five students is associated with higher reading performance than having one textbook for fewer students. Again, it is possible that there was a misunderstanding of this question, especially when looking at the responses to the question *Does every student in your class have the new mother tongue student textbook for his or her own?* which implies a 1:1 textbook-to-student ratio. Teachers who report that each of their students have his or her own mother tongue textbook are significantly and positively associated with reading performance.

Table 21. School Resources (Reported by Teachers) Associated with Reading Performance



Sig	D
0.00	0.14
0.00	0.42
0.00	0.36
0.00	0.19
0.00	0.07
0.00	-0.39
0.00	-0.35
0.00	-0.35
0.00	-0.25
0.00	0.19
0.00	0.20
0.00	0.05
0.00	0.16
0.00	0.21
0.00	0.18
0.00	0.18
0.00	0.11
0.00	-0.04

5.3 STUDENT QUESTIONNAIRE

Student contextual factors (as assessed by the student questionnaire) were analyzed in relation to their performance in the ORF subtask. A total of 12,986 students responded to the questionnaire; of these, 51% are male and 49% are female. Regarding language spoken at home, 95% reported that the medium of instruction and language spoken at home is the same, while 5% reported that the medium of instruction is different. Regarding literacy in the family, 82% report they have family members who can read and write, while 18% report that no one in the family can read or write. The following sections analyze these contextual factors for students.

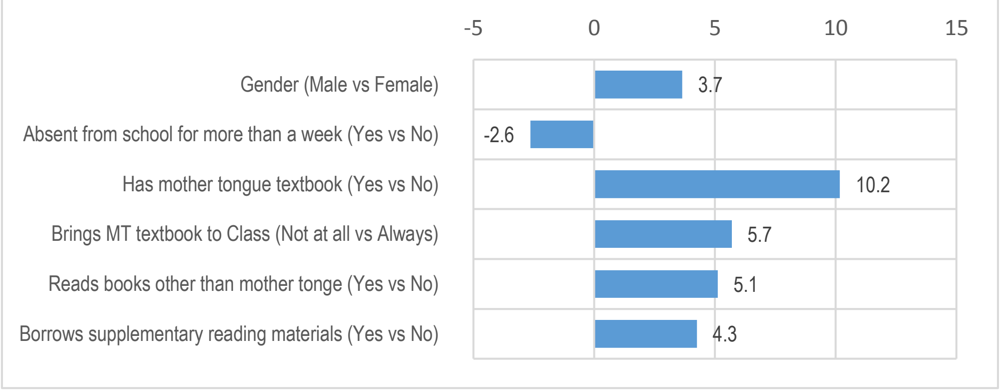
5.3.1 STUDENT CHARACTERISTICS

Student characteristics that are significantly associated with reading performance are gender, availability of the mother tongue textbook, bringing the mother tongue textbook to class every day, reading books in languages other than mother tongue, and borrowing supplementary reading materials. The corresponding significance levels and effect sizes are shown in Table 22.

Male students demonstrate higher reading performance than female students. Possessing the mother tongue textbook and taking it to school every day are associated with higher performance in reading. Regarding students' attitudes toward reading, those who report reading books in languages other than the mother tongue textbook and borrowing supplementary reading materials obtain ORF scores significantly higher than the students who do not report these activities.

A negative association with reading performance was found for students reporting that they were absent from school for more than a week.

Table 22. Student Characteristics Associated with Reading Performance



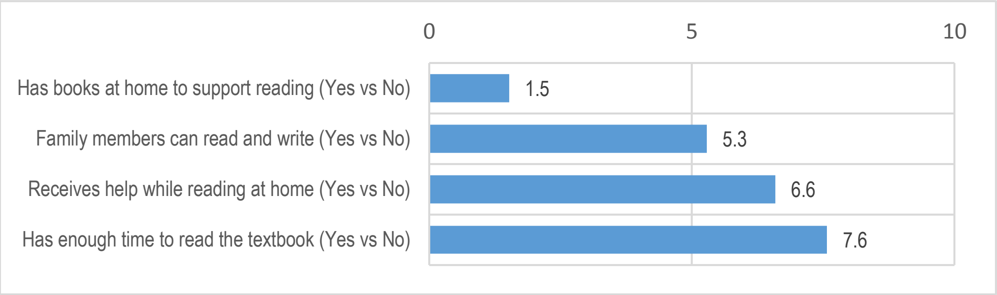
Sig	D
0.00	0.18
0.00	-0.13
0.00	0.51
0.00	0.27
0.00	0.25
0.00	0.21

5.3.2 HOME RESOURCES REPORTED BY STUDENTS

This section addresses the indicators of home resources obtained from student responses that are significantly associated with their reading performance. These variables are having books at home, having literate family members, receiving help with reading, and having enough time to read at home.

An interesting finding is in relation to the question *Is the language you speak at home and the language you learn at school the same?* The responses to this question do not show either positive or negative association with reading performance.

Table 23. Home Resources (Reported by Student) Associated with Reading Performance

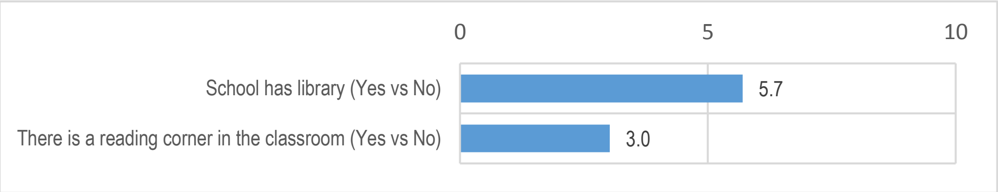


Sig	D
0.00	0.08
0.00	0.26
0.00	0.33
0.00	0.38

5.3.3 SCHOOL RESOURCES REPORTED BY STUDENTS

Indicators of school resources are frequently identified as highly relevant factors for the student learning progress. The two significant school environment variables reported by students are: existence of a school library in the community and existence of a reading corner at the school. Both variables are significantly and positively associated with reading performance as shown in Table 24.

Table 24. School Resources (Reported by Student) Associated with Reading Performance



Sig	D
0.00	0.28
0.00	0.15

6. DISCUSSION AND POLICY CONSIDERATIONS

The 2018 EGRA endline results offer critical information to inform discussions and decisions concerning policies, professional development, strategies and interventions to improve reading outcomes, especially reading comprehension among early grade students in Ethiopia. Prepared for Ethiopian policy makers, administrators and teacher practitioners, this section provides a summary of general conclusions from the findings and some key implications and recommendations, which must be discussed with these audiences for full interpretation in the Ethiopian context.

Endline results indicate several encouraging patterns, along with areas in need of intense improvement. The results are based on the sample of grade 2 and grade 3 students in five targeted regions and seven key mother tongue languages in Ethiopia. The presentation and discussion of the results revolves around two points of consideration relevant for policy decisions about the way forward: 1) expectations about student performance; and 2) resources and factors affecting the endline national aggregate scores.

The first point focuses on expectations about the level of student reading performance in the early grades of primary education. One would hope that by grade 2, students have acquired the prerequisite reading skills and are able to comprehend grade- and culturally-appropriate text materials. By the end of grade 3, students should be able to read adequately to ensure their ongoing involvement and success in the education system. By grade 4 we expect students to be able to “read to learn”, whereas the emphasis in earlier grades has been more on “learning to read”. When students fail to acquire effective reading skills by the end of grade 3, we typically see students abandoning the school system in the following grades. It should also be noted that the amount of time it takes to “learn to read” is language specific and orthographically dependent. The number of symbols and their combinations to be acquired are different in different languages and may affect the expectations about the grade level at which students achieve decoding fluency (Nag & Perfetti, 2014).

The second point of consideration is the fact that the scores reported in this READ M&E 2018 study relate to an endline evaluation; that is, a study of levels of achievement in aspects of early grade reading at the end of a period of interventions designed to improve reading skills. We would thus expect scores to be elevated in an endline evaluation compared to midline and baseline evaluations since they presumably reflect an increasing level of appropriate instructional methodology and resources for reading. There are multiple factors to be considered, however, in setting expectations about the progress in early grades reading at a broad, national level. These factors include, but may not be limited to, the degree to which the intervention was implemented and the availability of resources, such as reading materials and human resources, competency level of teachers and available training. Further, some contextual factors may interfere with intervention implementation at a national level, and factors in the macro environment, such as natural perils or political commotions, which might have had an unforeseen effect on instruction time, teacher and student attendance, and other variables affecting student performance.

The findings related to these two points should be taken into consideration in policymakers' deliberations about future actions to improve reading instruction and overall student learning outcomes.

6.1 DISCUSSION OF STUDENT PERFORMANCE ON EGRA

6.1.1 EGRA SUBTASK SCORES

There are two positive signs that suggest performance goals can be achieved if an effective strategy and sustained evidence-based reading interventions are implemented. The first of these is in listening comprehension, a fundamental early skill that supports the acquisition of comprehension-based reading skills. Scores are solid in the targeted grades: the overall listening comprehension average across seven languages in grade 2 is 66%, and in grade 3 it is 73%. These aggregated scores across languages are only comparable in a general descriptive sense, not in a statistical evaluation sense, because the listening tests are different for each language and are clearly based on quite different language and sociocultural contexts. We should conclude from these results that students in all languages acquire solid listening skills and are able to identify key information in grade-appropriate texts to which they listen. We can also conclude that text comprehension as a general construct does not appear to be a problem in learning for students in Ethiopia. They receive solid experiences in listening for information and develop useful skills in this area. We can also conclude that students' lack of success in the parallel skill of reading comprehension (14% in grade 2 and 27% in grade 3) has more to do with not having developed appropriate decoding skills, a key pre-comprehension skill, rather than the lack of ability to comprehend the information received. Ethiopian students can develop these decoding skills, given the right pedagogical approach and given the availability of appropriate materials and human resources. A study of the contextual factors points the way to potential strategies.

The other positive sign is that there is a clear trend toward improvement of all reading skills across the early grade levels. Improvements in scores from grade 2 to grade 3 aggregated for all seven languages are: increased ORF from 13 words-per-minute to 23 words-per-minute (effect size of 0.55, which counts as a substantial educational effect), and increased reading comprehension from 14% to 27% (a substantial educational effect size of 0.51). All other measures similarly improved from grade 2 to grade 3. This suggests a clear ability to improve, and one would assume that given the right approach and support to reading instruction, these increases could occur sooner and lead to appropriate levels of gain by the end of grade 3.

The endline results for grade 2 student scores predict that by the end of grade 3, students in the targeted regions and languages in Ethiopia will still lag in demonstrating grade-appropriate reading skills. There is plenty of evidence for this finding, including the ORF words-per-minute scores and reading comprehension percentage scores discussed above, which clearly show how large the gap is from acceptable levels of comprehension. The percentage of students who scored zero on ORF and reading comprehension is high. For example, 47% of grade 2 students scored zero in ORF, which falls to 28% in grade 3, and 64% of grade 2 students scored zero in reading comprehension, which falls to 45% in grade 3.

6.1.2 GENDER GAP

In recent decades, USAID and other international donors have focused on identifying and rectifying performance gaps by gender throughout the world. In many countries, emphasis has appropriately focused on closing access and achievement gaps between boys and girls.

In Ethiopia, the 2016 EGRA midline results showed that gender differences were relatively small and balanced across languages. Boys performed significantly better in Somali and Haddiysa; girls significantly outperformed boys in Afaan Oromo and Sidamu. In other languages, differences were negligible.

However, in the 2018 EGRA, the gender gap increased in favor of boys in all languages, except Amharic. Boys significantly outperformed girls in virtually all EGRA subtasks, revealing deep gaps in the more advanced skills of ORF and reading comprehension in three languages: Somali, Haddiysa, and Sidamu. Girls were only better in Amharic, but the size of the difference was small to negligible.

Important questions to consider at the policy level relate to how Ethiopia plans to address gender gaps and improve reading and other educational outcomes for girls, beginning in the early grades. These gaps would not necessarily require urgent measures if they indicated that lag decreased in later grades. Available data on educational attainment by gender in Ethiopia, however, indicate that the gender gap increases in grade 3 as students have to master increasingly difficult skills. Therefore, we recommend the following questions for consideration and discussion:

- Does girls' attendance drop off for some reason?
- Are there general pedagogical practices that hinder the growth of girls' skills? Are girls called upon in classrooms as frequently as boys?
- Are there any issues with learning styles that hinder their participation? Are teaching methods suitable to keep girls engaged?
- Are there activities and materials that could better motivate girls for learning?
- Are there cultural norms or domestic issues that disproportionally distract girls from active participation in learning activities, attending class, or completing homework assignments?
- Do girls have access to libraries and supplemental reading materials outside the classroom in the same way as boys?
- What is the role of each parent and other caregivers and community mentors in modeling language use and literacy activities at home? In which language(s)?

6.1.3 PERCENTAGE OF STUDENTS REACHING BENCHMARKS

Another piece of evidence underlining the low endline scores is the percentage of students who achieved the highest benchmark level, thus demonstrating reading fluency with full or almost full comprehension. Since the benchmarks were set specifically for each language using the same conceptual definitions, the benchmark levels provide a common framework that enable comparisons among languages, as well as aggregations of results across languages. Using locally set benchmarks overcomes the issue of comparability due to linguistic and orthographic differences between languages.

The percentage of students who reached the most desirable benchmark, *reading fluently with full or almost full comprehension*, aggregated for all languages, was just 4% in grade 2, rising to 8% in grade 3. This evidence indicates that less than 10% of students in Ethiopia are reaching the fully functional grade-appropriate reading level by the end of grade 3. This performance result is a call for action on intensifying ongoing efforts in improving reading levels across the country.

However, for monitoring and evaluation purposes, it is more informative to take into consideration the two upper benchmark levels combined: *reading fluently with full or almost full comprehension* and *reading with increasing fluency and comprehension*. When aggregated across all languages, 25% of grade 2 and 40% of grade 3 students fall into these two upper benchmark levels combined. Thus, it can be stated that, on average, 40% of Ethiopian students are reaching either fully functional or partially functional grade-appropriate reading levels at the end of grade 3.

At the bottom of the distribution, 60% of students have insufficient reading skills at the end of grade 3, falling into either the non-reader category (28% of students with *zero scores*) or having insufficient fluency and comprehension (32% of students *reading slowly with limited comprehension*). Again, this finding calls for further actions to improve reading performance of Ethiopian students.

When looking at student performance across different languages, we observe considerable differences. The percentages of students with reading skills reaching the upper two benchmark levels range from as high as 50% in Amharic and 47% in Tigrigna to as low as 16% in Haddiysa. These differences show policymakers which regions require more concerted efforts to improve students' early grade reading performance.

6.2 DISCUSSION OF CONTEXTUAL FACTORS ASSOCIATED WITH STUDENT PERFORMANCE

READ M&E analyzed the associations between student ORF scores and background factors based on the responses in directors', teachers', and students' questionnaires. The questionnaires collected information about factors such as access to teaching and learning materials, school support activities, teachers' and principals' characteristics and activities, and home and school environment factors. The results of these analyses revealed many significant associations relevant for policymakers' consideration to improve strategies for reading throughout the country.

6.2.1 BACKGROUND FACTORS ASSESSED BY QUESTIONNAIRES

The background variables based on responses from ***school directors*** showed a pattern of associations useful for policy considerations. Directors' level of education yielded a small but significant association with reading performance of their students. The directors' management of activities performed by other school officials (department heads and unit leaders) showed a substantial positive association with reading performance of students in their schools. Directors' training on mother tongue education, as well some of their own activities, such as supporting mother tongue teachers, did not show positive association with reading performance. It is perhaps because in the schools with low resources (thus lacking dedicated staff for reviewing and supporting mother tongue teachers), directors need to conduct those activities by themselves. This finding also implies that it may be more effective when directors manage school activities performed by other officials, than when they conduct these activities themselves. Obviously, management and support are the best roles to play for directors. Recommendations follow:

- Stipulate that every school must have a person in charge to review mother tongue lesson plans (unit leader or department head) and that they do so regularly (at least once per month or more often).
- Stipulate that every school must have a person in charge to observe mother tongue classes (director, department head, or unit leader) and have him or her observe mother tongue teachers at least three times per semester.
- Monitor student progress by tests and consider implementing a comprehensive formative assessment system, including brief classroom assessments.
- Empower school principals and community stakeholders to engage parents and community members in creating overlapping home, community, and school environments that support learning and gender equality.

Directors' responses to questionnaires about availability of school resources appeared to be significantly related to student performance. For example, when the mother tongue textbooks were received, the availability of mother tongue teacher's guides, number of mother tongue teachers at the school, educational qualification of teachers, percentage of mother tongue teachers who have received in-service training, availability of supplementary reading materials, and students making use of the school library all made a difference. Recommendations follow:

- Ensure that mother tongue teachers have teacher's guides and students receive their own textbooks on time.
- Based on school size, hire a sufficient number of mother tongue teachers, especially those with certificates, diplomas, and those who received in-service training.
- Ensure that schools receive supplementary reading materials, make them accessible to grade 1-4 students, and motivate or provide incentives for students to use the school library or reading room.

Background variables based on ***teachers' responses*** showed many significant associations with students' reading performance. For example, teachers' characteristics significantly associated with reading performance were: whether he or she is a trained teacher, level of education, years of service as a teacher, whether they received training on mother tongue materials, length of the training, and whether they know how to teach using the mother tongue materials. Recommendations follow:

- Provide schools with a sufficient number of teachers who hold pre-service diplomas and who have at least 5 years of teaching experience.
- Ensure that all teachers receive extensive training on using mother tongue materials.
- Assess teachers' self-confidence and actual competencies in teaching the new mother tongue curriculum and using the new mother tongue materials.

Various teachers' activities show a strong association with performance of their students. For example, teachers using the student textbook and the teacher's guide every time they teach, providing remedial classes, and holding discussions with parents of lagging students make a positive difference. Significantly higher reading performance was found in schools where teachers report that they frequently ask students to retell the story they read, voice out unfamiliar words, learn the meaning of new words, read aloud, and read by themselves. Also, when teachers report that they frequently perform written evaluations, check classroom exercises, and check homework, there was an increase in student reading performance. Students also demonstrate higher reading performance in schools where teachers have appropriate expectations of the reading skills that should be acquired in earlier grade levels. Recommendations follow:

- Stipulate that teachers must use student textbooks and the teacher's guide in every class they teach.
- Make sure that teachers provide remedial classes and meet with parents for students falling behind.
- Encourage teachers to conduct the following activities frequently (4-5 days a week): ask students to retell the stories they read, sound out unfamiliar words, learn the meaning of new words, read texts aloud, and read texts by themselves.
- Encourage teachers to frequently evaluate students (4-5 times a week) through: in-classroom written evaluations, checking classroom exercises, and checking homework.
- Build higher expectations among teachers regarding student reading performance.

Data from teachers' responses indicate that various school resources are significantly associated with increased student reading performance. These include average class size (up to 30 students), availability of the teacher's guide, availability of student textbooks, availability of supplementary reading materials, availability of school library, use of the library by the students, and existence of a functional PTSA at the school. Recommendations follow:

- Provide resources for building more classroom space or mobile classrooms and hire more teachers to ensure regular class sizes (preferably up to 30 students) and to avoid extremely large classes (over 60 students).
- Provide all students with their own textbooks.
- Ensure that schools have sufficient supplementary reading materials that are available to students.
- Provide all schools with functional libraries or reading rooms accessible to students and require students to use the library.
- Establish functional parent-teacher-student associations in schools.

Based on ***students' responses***, the background factors that showed significant association with their reading performance include gender, owning a mother tongue textbook, bringing the mother tongue textbook to class every day, reading books in languages other than the mother tongue, borrowing supplementary reading materials, and school absenteeism. Recommendations follow:

- Conceptualize and implement activities that contribute to closing the gender gap (discussed earlier in this chapter).
- Create measures for mitigating the effects of absenteeism and loss of instructional time.
- Provide all students with mother tongue textbooks and require that they bring them to every class.
- Motivate students to read books in languages other than mother tongue and to borrow supplementary reading materials.

The home contextual variables significantly related to student performance were having books at home, having literate family members, receiving help with reading, and having enough time to read at home. The significant school resource variables evaluated through students' responses include existence of school library and existence of a reading corner at the school. The following policy recommendations can be formulated:

- Organize community actions aimed to supply homes with books and other learning tools.
- Increase adult literacy programs in communities where needed.
- Build partnerships between schools and civil society to empower and guide parents to help children with reading.
- Organize activities aimed at reducing child domestic chores to allow them to read at home.

As the final note, these background factors were assessed through self-report instruments (questionnaires) and that their associations with student reading performance as measured by EGRA does not represent causal relationship. The factors' associations could be attributed to other variables not controlled in the study.

On the other hand, some background factors (for example school resources) presented in this section were assessed through multiple respondents (director, teachers, and student) and they yielded reasonably aligned results, which underscores both the reliability and the significance of these results for policy considerations.

6.2.2 CONTEXTUAL FACTORS BASED ON TEAM OBSERVATIONS

The factors discussed in this section are based on the READ M&E team's informal data collection through occasional classroom observations, anecdotes, and team discussions on these topics. These contextual factors and events could have influenced the reading performance of students in recent years. They should be considered when making policy decisions to improve reading performance.

Information gained about contextual factors in these informal ways overlap somewhat with data collected on the same topics through questionnaires. Informal data, in any case, complement questionnaire findings. Specific issues related to student reading performance are as follows:

- Performance suffers when reading is not taught as a subject and when it is embedded within the mother tongue subject. Reading is not taught explicitly, and sufficient time is not allotted in class and outside classroom.

- Teachers are not well acquainted with how to teach reading. They have received short-term trainings, but the duration and frequency of these trainings is insufficient. There is no systematic evidence concerning whether teachers implemented the trained methods at the desired level.

- There is very high teacher turnover. Therefore, untrained teachers are replacing trained ones. A teacher who is teaching another subject may become a mother tongue teacher without receiving the required training.

- School environments rich with books are not yet created. Schools and classrooms are not child-friendly, and they do not stimulate students to practice reading.

- Libraries and reading rooms are nonexistent or they are just nominal. In most cases they are not accessible to early grade students. Most of the books are reference materials for higher grade levels.

- Age-appropriate reading materials are not readily available to children. In some places, the new curriculum materials (teacher's guide and textbooks) are not yet available.

- In one case (Afaan Oromo), the newly developed materials are totally abandoned and no longer in use due to the RSEB's decision.

- We did not evaluate the mother tongue materials, but anecdotally, there are issues to be addressed and the materials may need some revisions. Some teachers complain that there is not sufficient time to implement scripted lesson plans as recommended. Others suggest that letter names and letter sounds should be fully addressed earlier in pre-primary classes or in grade 1.

- Lots of instructional time was missed due to drought and floods that contributed to high rates of absenteeism among both teachers and students. In some locations, it was difficult to get enough students to attend class during the assessment week.

- In some places, school directors and supervisors are not familiar with the new teaching methods and the curriculum materials. Teachers, therefore, may not be getting the necessary support from these officials.

- In teaching reading, teachers need to receive coaching by reading specialists, including modeling and continuous follow-up. However, the Ethiopian school system does not have reading coaches.

To conclude, this study provides two categories of findings: 1) information about the level of student reading performance in early grades, and 2) insights into the background factors associated with reading performance. It is important to carefully consider all the lessons learned from this study in planning strategies and approaches to improving early grade student reading performance in Ethiopia. The READ M&E team, in collaboration with MOE and RSEBs, will organize and conduct dissemination activities that will present the 2018 EGRA findings, as well as co-interpretation of data and recommendations for action, that will be customized to each region and language.

6.3 EGRA COMPARISONS ACROSS YEARS[4]

As stressed earlier, this READ M&E 2018 study is viewed as an endline evaluation of early grade reading designed to improve reading skills. One would expect that reading scores would demonstrate gains in an endline evaluation compared to midline and baseline evaluations. However, this expectation may not be justified without taking into consideration multiple factors related to the spread and depth of intervention activities and contextual factors specific to the locations and periods being observed. The expectation that improvements in reading scores would be observed at a national level assumes a widely increased level of appropriate instructional methodology and resources for reading throughout and across the country.

In this regard, there is little information about 1) the magnitude of appropriate use of reading instruction, 2) the extent to which instruction affects positive learning, 3) the evidence base undergirding reading practices in use, and 4) the extent to which evidence-based practices are implemented with integrity and fidelity.

When looking at the overall trend in reading performance at a national level, little variation can be observed between the EGRA 2014, 2016, and 2018 administrations.

[4] Data obtained for Aff Somali in 2014 and for Wolayttatto in 2016 are not included in these comparisons as the data were not considered sufficiently reliable.

The percentage of students whose reading performance falls within the upper two benchmark levels (*increasing* and *full comprehension*) was 31% in 2014, 34% in 2016, and 32% in 2018—thus revolving around one-third of the student population.

Considering each language separately, the comparison between the baseline (2014) and midline (2016) EGRA administrations reveals that reading performance significantly increased in Amharic and Sidamu Affo, whereas differences in other languages were practically negligible. When looking at changes of ORF scores between the 2016 and 2018 administrations, student performance in Sidamu Affo and Amharic decreased, but stayed significantly above the 2014 level. A significant increase was observed only in Aff Somali, and changes in other languages were either negligible or marginal.

Therefore, the main conclusions based on the variation of student reading scores across years in Ethiopia become apparent:

- Changes in student reading performance across administration years 2014, 2016, and 2018 at overall national level are relatively small and do not exhibit a desired trend of improvement.

- In some languages, significant changes in student reading performance across years were recorded, but no systematic trend of improvement was detected.

In the context of stagnating reading performance in Ethiopia, understanding the state of reading outcomes in the early grades is the crucial first step toward improving reading instruction and outcomes. It is essential to understand what is happening in this area, to monitor progress at all levels, to adapt and calibrate interventions and supports, and to capitalize on observations about contextual factors. Initiatives, reforms, and proposed changes to the status quo in any one area of the system must be tightly aligned with other parts of the system.

6.4 THE WAY FORWARD

Results from the endline evaluation suggest that current intervention activities must be strengthened and improved to make more dramatic improvements in early grade reading outcomes. Ethiopia’s MoE and supporting donors may consider a range of constructive strategies to plan a systematic approach in improving student reading proficiency.

Based on the results and discussion of factors associated with student performance, three major strategies for going forward emerged. Some represent the strengthening of ongoing processes, and others represent entirely new intervention strategies focusing on teacher competencies and school-based formative assessments. Policymakers might consider the following strategies and actions:

1. Increase the breadth and depth of the current intervention. This is a continuation of the current activities primarily focusing on two major resources—human (teacher training at both pre-service and in-service levels) and materials (supplies for teaching reading, books, libraries, reading rooms, etc.). Also, focus on USAID’s proposed five T’s as the key to reading success:

    - Time: More time devoted to teaching reading;

- Techniques: Better techniques for teaching reading;
- Texts: More texts in the hands of children;
- Tongue: Teach children in the mother tongue; and
- Testing: Monitor and evaluate children's reading progress.

READ M&E recommends that all the T's be continuously evaluated and improved, and that formative classroom assessments and remediation activities be more frequently conducted. For example, regarding the first two T's, consideration should be given to questions such as: how much time is good enough to teach reading? Should that time be spent on teaching the code, improving vocabulary, or teaching phonological awareness? How should teachers be trained to teach reading – direct training, cascaded, peer mentoring, coaching, or scripted lessons vs. not scripted? What are the best techniques for teaching reading in fidel-based languages vs. latin-based languages? Finally, we should ask ourselves, has our monitoring and evaluation unpacked the shortcomings (and strengths) of the current intervention program(s) on the reading sub-skills of interest?
Also, READ M&E proposes that the fifth T (testing) be elaborated and expanded upon, as explained further in the text that follows.

2. Develop the instruments and procedures to evaluate corresponding intervention outcomes. That is, we recommend developing and implementing two types of outcome indicators: teacher competencies and school competencies. This recommendation is aligned with the previous recommendation since the intervention primarily focuses on teachers and schools.

   a. Develop and administer assessments of *teacher competencies* for teaching reading. This is an extremely valuable tool that provides insights into the strengths and weaknesses of the primary recipients of intervention (teachers) and informs where adjustments are necessary. Teachers deliver the interventions to students. Measuring student reading performance by EGRA captures the student outcomes of reading intervention. However, there is a major gap in assessing the primary aspects of intervention – teacher competencies for teaching reading. Thus, READ M&E recommends that assessment of teacher competencies become a routine part of the monitoring and evaluation protocols for reading and other educational interventions.

   b. Develop a system for evaluating *school competencies* for teaching reading. This entails creating an index to measure the degree to which the school environment and climate will enable effective teaching of reading -- the index that will include both human (teachers, directors, and other school officials) and material resources of the school (libraries, books, equipment, etc.). This development may lead towards the concept of school accountability system for the resources and support they provide. The index of school competencies should be structured to include the roles of the other school officials, not just teachers, as effective learning cannot happen without effective leadership and management support.

3. Establish a comprehensive, but easy to use, formative assessment system that empowers teachers to monitor and promote reading proficiency of students in early grades. This system will augment the current Formative Continuous Assessment (FCA) activity designed by READ M&E. The formative assessment system will include both instruction-embedded assessments (current FCA) as well as periodic EGRA-like assessments that teachers can use for monitoring and fostering reading performance of their students. Implementation of these strategies will entail the following systemic activities:

    a. Develop the assessment capacity of school principals so that they lead their administrative and teaching staff in the process of setting, monitoring, and achieving reading goals that reflect high expectations for all students.

    b. Build school capacity to monitor student progress regularly and train school data teams to implement formative assessment.

All three strategies above should be thoroughly discussed and customized to the diverse needs of regions and unique characteristics of the mother tongue languages in Ethiopia. READ M&E is already providing substantial assistance to MoE and USAID in conceptualizing and implementing the strategies related to the development and utilization of assessment of teacher and school competencies, as well as development and implementation of the comprehensive formative assessment system. Both assessment-based strategies play an important role in establishing the educational accountability system, which should be one of the crucial turns on the road towards improving education quality, and ultimately, student learning outcomes in Ethiopia.

7. APPENDICES

APPENDIX 1A. PILOTING PROCEDURES

7.1 PILOT ADMINISTRATION OF THE 2018 EGRA TOOL

Considering that the EGRA piloting served a dual purpose—evaluation of new forms and equating using a common-persons design, READ M&E developed the data collection design presented in Table 25.

Table 25. The Counterbalanced Data Collection Design for Piloting the 2018 EGRA Tools

Subtask Sequence	Set 1	Set 2	Set 3	Set 4
	Form A + Form REF	Ref Form + Form A	Form B + Form REF	Ref Form + Form B
1	FWR form A	FWR form REF	FWR form B	FWR form REF
2	IWR form A	IWR form REF	IWR form B	IWR form REF
3	ORF form A	ORF form REF	ORF form B	ORF form REF
4	RC form A	RC form REF	RC form B	ORF form REF
5	FWR form REF	FWR form A	FWR form REF	FWR form B
6	IWR form REF	IWR form A	IWR form REF	IWR form B
7	ORF form REF	ORF form A	ORF form REF	ORF form B
8	RC form REF	RC form A	ORF form REF	RC form B

Notes. FWR is familiar words reading; IWR is invented words reading; ORF is oral reading fluency; RC is reading comprehension. A is new form A; B is new form B; REF is reference form (midline).

Because the nature of common-persons equating design requires that the assessor administers two different forms of the test to each child, to avoid the effect of the order in which the two forms are administered it is necessary to apply a counterbalanced data collection design. This design makes a provision that half of the examinees receive the two forms in one order (e.g., form A followed by form REF), and the other half is assessed in reverse order (e.g., form REF followed by form A).

The two pilot forms (A and B) that READ M&E developed for the EGRA 2018 administration, were each combined with the 2016 midline form that served as a reference for equating (A+REF and B+REF). To apply the counterbalanced design, these two form combinations were further broken down into 4 sets of EGRA tasks, as presented in Table 25 above. All 4 sets were administered at each school to 40 randomly selected examinees, 20 from grade 3 and 20 from grade 4, so that 20 students per school each completed the two pilot forms A and B.

READ M&E conducted the piloting of the 2018 EGRA tools in seven languages spoken in five regions of Ethiopia (Tigray, Amhara, Oromia, Somali, and languages of the Southern Nations, Nationalities, and People's Region), which took place in November 2017. The test administration involved 48 test assessors deployed in 12 teams of four individuals to collect data in 56 sample schools in five regions in seven languages (eight schools representing each language). Considering that 40 students were selected at each school, the number of assessed students was 320 per language, yielding a total pilot sample of 2,240 students. The 2018 EGRA pilot was conducted on Nexus 7 tablets loaded with Tangerine software.

7.2 PILOT DATA ANALYSIS

Piloting the tools enabled READ M&E to assess item parameters for difficulty of stimuli (letter, words, or questions) and to establish the equating relationship between newly developed forms (A and B) and the 2016 EGRA taken as a reference form (R).

Pilot data analysis entailed computation of conditional P-values as a measure of item difficulty for timed tasks. “Conditional” in this context means that it was taken into consideration that not all students reach all the items in the timed tasks—as the end of stimulus list is approaching, the smaller number of students with increasingly higher ability is reaching each item. Thus, to enable comparability of item difficulty indicators across the entire length of the stimuli list, we applied an adjustment that yields “conditional” P-values estimated as if each item were taken by the entire sample of students. Figure 25 shows an example of conditional P-values for ORF in Amharic language.

Conditional P-values were used as one of the indicators for evaluating the quality of pilot forms and making decisions about which pilot form would be selected for operational administration. The form that exhibits a smoother curve (no large drops or bumps) of conditional P-values is considered as better structured form, not containing the items that could confuse a student.

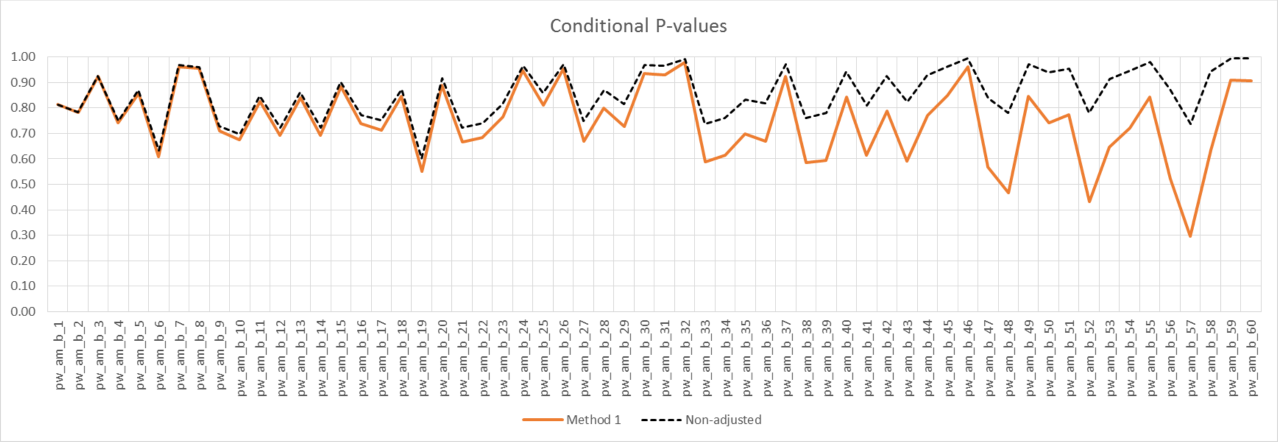


Figure 25. Example of Conditional *P*-Values: Oral reading fluency Form B in the Amharic Language

Another part of data analysis for each of the four piloted subtasks involved a computation of means for the new forms A and B, as well as for the reference form R. Testing of statistical significance was carried out for selected pairs of those means, as displayed in Table 26.

Table 26. Means for Oral Reading Fluency in Amharic Language

Amharic—Oral Reading Fluency (wpm)					
Compared Forms	**Compared Means**		**Difference**	**Significance**	**Comments**
A - B	31.6	37.7	n/a	n/a	Different forms, different persons, significance not tested
A - Ra	31.6	35.7	-4.1	0.00	Different forms, same persons, paired samples t-test
B - Rb	37.7	37.0	0.7	0.12	Different forms, same persons, paired samples t-test
Ra - Rb	35.7	37.0	-1.2	0.59	Same form, different persons, independent samples t-test

Notes. A is pilot form A; B is pilot form B; Ra is reference form taken by group A; Rb is reference form taken by group B.

The following pairs of the means were evaluated:

- **A – B:** Means for form A and form B. Since two different groups completed each of the two different forms, we do not know if the difference between means is attributable to differences in the difficulty level of the forms or to differences in the ability of the groups. Significance was not evaluated (no meaningful interpretation, and no decision needs to be made based on this difference).

- **A – Ra:** Means for form A and form R taken by the same persons (group A). In this common-persons situation, the difference between means can be attributed to the difference between form difficulty. Statistical significance was evaluated by a paired samples t-test. If the difference were statistically significant, it would be taken as equating constant.

- **B – Rb:** Means for form B and form R taken by the same persons (group B). The interpretation is the same as described in the preceding paragraph.

- **Ra – Rb:** Means for form R taken by two groups of different persons (groups A and B). In this case, the difference between means can be attributed to the difference between abilities of the two groups taking the same test. Statistical significance was evaluated by an independent samples t-test. Because the students were randomly assigned to groups A and B, it is reasonable to expect that this difference would not be significant.

The results of the evaluations of differences between the means A – Ra and the means B – Rb were used as a major criterion for making a decision about which of the pilot forms, A or B, to select for operational 2018 EGRA administration. The means for all piloted EGRA subtasks in both pilot forms A and B, for all languages, along with tests of significance and indication of which pilot forms were selected for operational 2018 EGRA endline administration, are presented in Appendix 1B.

7.3 COMPARABILITY ACROSS ADMINISTRATION YEARS

To establish the comparability between the 2018 and 2016 EGRA tools, READ M&E used a common-persons piloting design. The purpose was to: 1) evaluate the EGRA

2018 newly developed forms, and 2) enable computation of equating relationship between the reference form (2016) and each of the new forms (A and B).

In this design, the same pupils sit for more than one form of the assessment, as presented in Table 25 in Section 2.2.3. The rationale of common-persons design is that the pupils who take two different forms of the instrument have the same underlying distribution of ability. Thus, any difference between the results collected from these two forms can be attributed to the form characteristics rather than to the student characteristics.

READ M&E established comparability of the new forms A and B with the reference form using the classical test theory equating, specifically the mean equating method. This equating method was selected as the simplest and most transparent, suitable for equivalent-groups data, including data collected by common-persons design. It relies on the plain assumption that difference in difficulty between two forms *X* and *Y* is constant throughout the entire score range. Although this assumption may not always be met in practice, the mean equating appears to be robust and, according to our preliminary analyses, yielding virtually the same outcomes as obtained by more complicated methods, such as linear or equipercentile equating. The computation of equated scores using mean equating is based on the following formula:

$$X - \bar{X} = Y - \bar{Y}$$
$$Y = X + (\bar{Y} - \bar{X})$$
$$Y = X + EC$$

Where: *X* is the score on new form *Y*, *Y* is the score on reference form; $\bar{X}$ is the mean of new form; $\bar{Y}$ is the mean of reference form; and *EC* is the equating constant.

Thus, using mean equating method, the scores obtained by the newly piloted forms are made comparable with the reference form by making a simple linear transformation: adding the ***equating constant*** to the scores obtained by the new form. The equating constant is defined as a difference between the means of the reference form and the new form of the test $(\bar{Y} - \bar{X})$. In such a way, the scores obtained by the new form are placed on the same scale as the scores from the reference form, so they become directly comparable. If the difference between the means of the reference and the new form is not statistically significant, it indicates that the difficulties of both the forms are equivalent and no equating adjustment is necessary.

Based on the results of equating, as well as considering the conditional P-values of these forms, READ M&E selected the pilot forms that have smaller equating constant (thus, closer in difficulty to the reference form) to serve as operational forms in the 2018 endline administration. The obtained equating constants and forms selected for operational administration are presented in Table 27.

Table 27. Selected Forms and Corresponding Equating Constants

Language	FWR		IWR		ORF		RC	
	Form	EC	Form	EC	Form	EC	Form	EC
Afaan Oromo	B	3.9	B	n/s	B	1.8	B	4
Af Somali	A	2.8	A	1.9	A	n/s	A	3.1
Amharic	B	5.3	B	1.3	B	n/s	B	n/s
Haddiysa	B	0.7	B	n/s	B	n/s	B	7.6
Sidamu Affo	B	1.7	A	n/s	B	-1.4	B	-3
Tigrigna	B	8.4	A	2.1	A	-2.5	A	n/s
Wolayttatto	A	n/s	B	1.2	A	n/s	A	n/s

Notes. FWR is familiar words reading; IWR is invented words reading; ORF is oral reading fluency; RC is reading comprehension; EC is equating constant.

APPENDIX 1B. MEANS FOR PILOTED EARLY GRADE READING ASSESSMENT SUBTASKS

AMHARIC

Familiar Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	31.6	32.0	-0.3	n/a
A - Ra	31.6	37.2	-5.6	0.00
B - Rb	32.0	37.3	-5.3	0.00
Ra - Rb	37.2	37.3	-0.1	0.98

Invented Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	21.2	21.8	-1.5	n/a
A - Ra	21.2	23.9	-2.7	0.00
B - Rb	21.8	23.0	-1.3	0.00
Ra - Rb	23.9	23.0	0.9	0.58

Passage Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	31.6	37.7	-4.9	n/a
A - Ra	31.6	35.7	-4.1	0.00
B - Rb	37.7	37.0	0.7	0.12
Ra - Rb	35.7	37.0	-1.2	0.59

Reading Comprehension				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	21.5	38.8	-12.7	n/a
A - Ra	21.5	34.2	-12.7	0.00
B - Rb	38.8	38.8	0.0	1.00
Ra - Rb	34.2	38.8	-4.5	0.15

HADDIYSA

Familiar Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	13.9	7.9	-0.6	n/a
A - Ra	13.9	15.2	-1.3	0.01
B - Rb	7.9	8.6	-0.7	0.04
Ra - Rb	15.2	8.6	6.6	0.00

Invented Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	13.3	6.8	-0.8	n/a
A - Ra	13.3	14.4	-1.1	0.00
B - Rb	6.8	7.0	-0.2	0.43
Ra - Rb	14.4	7.0	7.3	0.00

Passage Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	15.3	8.1	-0.6	n/a
A - Ra	15.3	16.0	-0.8	0.01
B - Rb	8.1	8.3	-0.2	0.60
Ra - Rb	16.0	8.3	7.8	0.00

Reading Comprehension				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	21.4	8.3	5.4	n/a
A - Ra	21.4	23.6	-2.3	0.05
B - Rb	8.3	15.9	-7.6	0.00
Ra - Rb	23.6	15.9	7.8	0.02

OROMO

Familiar Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	9.4	16.2	-2.7	n/a
A - Ra	9.4	16.1	-6.7	0.00
B - Rb	16.2	20.1	-3.9	0.00
Ra - Rb	16.1	20.1	-4.0	0.09

Invented Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	6.6	9.5	0.9	n/a
A - Ra	6.6	5.8	0.7	0.02
B - Rb	9.5	9.7	-0.1	0.75
Ra - Rb	5.8	9.7	-3.8	0.00

Passage Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	13.8	19.0	-1.3	n/a
A - Ra	13.8	16.8	-3.1	0.00
B - Rb	19.0	20.8	-1.8	0.00
Ra - Rb	16.8	20.8	-4.0	0.13

Reading Comprehension				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	11.5	22.0	-5.0	n/a
A - Ra	11.5	20.5	-9.0	0.00
B - Rb	22.0	26.0	-4.0	0.00
Ra - Rb	20.5	26.0	-5.5	0.14

SIDAMA

Familiar Words Reading

Compared Forms	Compared Means		Difference	Significance
A - B	14.8	15.3	-0.9	n/a
A - Ra	14.8	17.3	-2.6	0.00
B - Rb	15.3	17.0	-1.7	0.00
Ra - Rb	17.3	17.0	0.4	0.82

Invented Words Reading

Compared Forms	Compared Means		Difference	Significance
A - B	12.4	14.0	1.1	n/a
A - Ra	12.4	13.3	-0.9	0.05
B - Rb	14.0	16.0	-2.0	0.00
Ra - Rb	13.3	16.0	-2.8	0.08

Passage Words Reading

Compared Forms	Compared Means		Difference	Significance
A - B	19.7	17.7	1.8	n/a
A - Ra	19.7	16.5	3.3	0.00
B - Rb	17.7	16.3	1.4	0.00
Ra - Rb	16.5	16.3	0.2	0.92

Reading Comprehension

Compared Forms	Compared Means		Difference	Significance
A - B	21.4	17.5	1.1	n/a
A - Ra	21.4	17.3	4.1	0.00
B - Rb	17.5	14.5	3.0	0.04
Ra - Rb	17.3	14.5	2.7	0.25

SOMALI

Familiar Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	17.9	17.3	-0.5	n/a
A - Ra	17.9	20.7	-2.8	0.00
B - Rb	17.3	19.6	-2.3	0.00
Ra - Rb	20.7	19.6	1.1	0.62

Invented Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	18.3	16.3	-0.2	n/a
A - Ra	18.3	20.2	-1.9	0.00
B - Rb	16.3	18.0	-1.7	0.00
Ra - Rb	20.2	18.0	2.2	0.29

Passage Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	21.9	22.5	-2.6	n/a
A - Ra	21.9	22.1	-0.2	0.63
B - Rb	22.5	20.2	2.3	0.00
Ra - Rb	22.1	20.2	1.9	0.43

Reading Comprehension				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	21.9	35.1	-14.5	n/a
A - Ra	21.9	25.0	-3.1	0.01
B - Rb	35.1	23.7	11.4	0.00
Ra - Rb	25.0	23.7	1.3	0.72

TIGRIGNA

Familiar Words Reading				
Compared Forms	Compared Means		Difference	Significance
A - B	21.3	30.5	-5.4	n/a
A - Ra	21.3	35.1	-13.8	0.00
B - Rb	30.5	38.9	-8.4	0.00
Ra - Rb	35.1	38.9	-3.8	0.17

Invented Words Reading				
Compared Forms	Compared Means		Difference	Significance
A - B	17.3	16.5	2.2	n/a
A - Ra	17.3	19.5	-2.1	0.00
B - Rb	16.5	20.8	-4.3	0.00
Ra - Rb	19.5	20.8	-1.3	0.40

Passage Words Reading				
Compared Forms	Compared Means		Difference	Significance
A - B	27.4	32.6	-2.4	n/a
A - Ra	27.4	24.9	2.5	0.00
B - Rb	32.6	27.8	4.9	0.00
Ra - Rb	24.9	27.8	-2.8	0.20

Reading Comprehension				
Compared Forms	Compared Means		Difference	Significance
A - B	22.0	32.5	-4.8	n/a
A - Ra	22.0	21.4	0.6	0.71
B - Rb	32.5	27.0	5.4	0.00
Ra - Rb	21.4	27.0	-5.7	0.05

WOLAYTATTO

Familiar Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	18.8	11.6	1.6	n/a
A - Ra	18.8	18.6	0.2	0.79
B - Rb	11.6	13.0	-1.5	0.00
Ra - Rb	18.6	13.0	5.6	0.00

Invented Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	15.8	10.5	0.5	n/a
A - Ra	15.8	16.5	-0.7	0.22
B - Rb	10.5	11.7	-1.2	0.01
Ra - Rb	16.5	11.7	4.8	0.01

Passage Words Reading				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	21.9	11.8	3.0	n/a
A - Ra	21.9	21.6	0.3	0.61
B - Rb	11.8	14.5	-2.7	0.00
Ra - Rb	21.6	14.5	7.1	0.00

Reading Comprehension				
Compared Forms	**Compared Means**		**Difference**	**Significance**
A - B	27.3	12.3	4.2	n/a
A - Ra	27.3	28.2	-0.9	0.46
B - Rb	12.3	17.4	-5.1	0.00
Ra - Rb	28.2	17.4	10.9	0.00

APPENDIX 2. INDEPENDENT SAMPLE T-TEST OF THE TIMED TASKS BY GRADE

Language = Afaan Oromo

	Grade	N	Mean	Std. Deviation	Mean Difference	Std Dev	Cohen's d
LNR	Two	1,118	41.8	25.5	15.1	27.0	0.6
	Three	1,214	56.9	28.4			
FWR	Two	1,118	12.8	11.5	8.1	14.1	0.6
	Three	1,214	20.9	16.7			
IWR	Two	1,118	4.9	8.7	4.5	11.4	0.4
	Three	1,214	9.4	14.2			
ORF	Two	1,117	11.1	14.0	10.2	17.7	0.6
	Three	1,214	21.3	21.5			

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Aff Somali

	Grade	N	Mean	Std. Deviation	Mean Difference	Std Dev	Cohen's d
LNR	Two	817	37.6	29.9	14.5	29.4	0.5
	Three	737	52.1	28.9			
FWR	Two	817	11.6	13.6	7.4	15.0	0.5
	Three	737	19.0	16.4			
IWR	Two	817	10.5	13.8	6.9	14.8	0.5
	Three	737	17.4	15.7			
ORF	Two	817	10.6	17.7	9.6	19.3	0.5
	Three	737	20.2	20.8			

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Amharic

	Grade	N	Mean	Std. Deviation	Mean Difference	Std Dev	Cohen's d
LNR	Two	782	29.6	22.4	13.2	23.8	0.6
	Three	803	42.8	25.2			
FWR	Two	782	27.3	16.0	10.6	17.3	0.6
	Three	804	37.9	18.6			
IWR	Two	782	18.6	11.6	6.3	12.1	0.5
	Three	804	24.9	12.6			
ORF	Two	782	24.9	17.8	13.3	19.5	0.7
	Three	804	38.1	21.2			

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Haddiysa

	Grade	N	Mean	Std. Deviation	Mean Difference	Std Dev	Cohen's d
LNR	Two	952	28.8	28.7	17.9	30.7	0.6
	Three	943	46.7	32.7			
FWR	Two	952	7.3	12.7	6.6	14.5	0.5
	Three	943	13.9	16.3			
IWR	Two	952	5.2	10.5	5.4	12.1	0.4
	Three	943	10.6	13.7			
ORF	Two	952	5.9	12.0	6.6	14.4	0.5
	Three	943	12.5	16.8			

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Sidamu Affo

	Grade	N	Mean	Std. Deviation	Mean Difference	Std Dev	Cohen's d
LNR	Two	981	38.2	31.1	19.4	31.8	0.6
	Three	880	57.7	32.6			
FWR	Two	981	10.7	12.3	7.8	13.6	0.6
	Three	880	18.5	15.0			
IWR	Two	981	7.7	11.6	6.8	13.3	0.5
	Three	880	14.5	15.0			
ORF	Two	981	10.3	15.4	10.5	17.5	0.6
	Three	880	20.8	19.6			

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Tigrigna

	Grade	N	Mean	Std. Deviation	Mean Difference	Std Dev	Cohen's d
LNR	Two	924	30.1	27.2	12.3	29.1	0.4
	Three	1,001	42.5	31.0			
FWR	Two	924	25.3	15.8	10.6	18.3	0.6
	Three	1001	35.9	20.9			
IWR	Two	924	11.7	10.3	4.4	11.4	0.4
	Three	1,001	16.1	12.5			
ORF	Two	924	15.8	14.3	9.5	16.5	0.6
	Three	1,001	25.3	18.8			

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Wolayttatto

	Grade	N	Mean	Std. Deviation	Mean Difference	Std Dev	Cohen's d
LNR	Two	931	32.7	28.0	14.4	28.6	0.5
	Three	901	47.1	29.2			
FWR	Two	931	18.5	15.0	8.6	17.1	0.5
	Three	901	27.1	19.1			
IWR	Two	931	13.5	18.0	9.9	20.4	0.5
	Three	901	23.5	22.7			
ORF	Two	931	8.9	18.0	9.9	20.4	0.5
	Three	901	18.9	22.7			

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

APPENDIX 3. INDEPENDENT SAMPLE T-TEST OF THE TIMED TASKS BY GENDER

Language = Afaan Oromo, Grade = 2

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
LNR	Female	572	39.6	25.3	-2.930	1116	0.003	-4.5	-0.18
	Male	546	44.1	25.5					
FWR	Female	572	11.8	11.0	-2.817	1116	0.005	-1.9	-0.17
	Male	546	13.8	11.8					
IWR	Female	572	4.2	8.5	-2.632	1116	0.009	-1.4	-0.16
	Male	546	5.6	8.8					
ORF	Female	571	9.9	13.7	-2.857	1115	0.004	-2.4	-0.17
	Male	546	12.3	14.2					

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Afaan Oromo, Grade = 3

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
LNR	Female	606	52.3	27.4	-5.722	1212	0.000	-9.2	-0.33
	Male	608	61.5	28.6					
FWR	Female	606	18.2	15.7	-5.584	1212	0.000	-5.3	-0.32
	Male	608	23.5	17.2					
IWR	Female	606	7.1	11.0	-5.529	1212	0.000	-4.5	-0.32
	Male	608	11.6	16.5					
ORF	Female	606	17.8	20.1	-5.781	1212	0.000	-7.0	-0.33
	Male	608	24.8	22.3					

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Aff Somali, Grade = 2

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
LNR	Female	340	31.5	28.5	-4.952	815	0.000	-10.4	-0.35
	Male	477	41.9	30.1					
FWR	Female	340	8.9	10.5	-4.809	815	0.000	-4.6	-0.36
	Male	477	13.5	15.1					
IWR	Female	340	7.9	11.8	-4.470	815	0.000	-4.3	-0.33
	Male	477	12.3	14.9					
ORF	Female	340	7.2	12.4	-4.664	815	0.000	-5.8	-0.35
	Male	477	13.0	20.3					

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Aff Somali, Grade = 3

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
LNR	Female	285	45.4	27.3	-5.096	735	0.000	-10.9	-0.39
	Male	452	56.3	29.0					
FWR	Female	285	14.1	13.7	-6.592	735	0.000	-8.0	-0.51
	Male	452	22.1	17.3					
IWR	Female	285	12.7	13.6	-6.565	735	0.000	-7.6	-0.51
	Male	452	20.3	16.2					
ORF	Female	285	15.1	18.4	-5.391	735	0.000	-8.3	-0.42
	Male	452	23.4	21.7					

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Amharic, Grade = 2

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
LNR	Female	403	30.8	22.4	1.495	780	0.135	2.4	0.11
	Male	379	28.4	22.4					
FWR	Female	403	28.4	16.0	2.017	780	0.044	2.3	0.14
	Male	379	26.1	15.9					
IWR	Female	403	19.2	11.6	1.569	780	0.117	1.3	0.11
	Male	379	17.9	11.6					
ORF	Female	403	26.0	17.8	1.901	780	0.058	2.4	0.14
	Male	379	23.6	17.7					

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Amharic, Grade = 3

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
LNR	Female	398	44.3	26.1	1.746	801	0.081	3.1	0.12
	Male	405	41.2	24.2					
FWR	Female	398	38.9	18.7	1.562	801	0.119	2.0	0.11
	Male	405	36.9	18.5					
IWR	Female	398	25.3	12.6	1.033	801	0.302	0.9	0.07
	Male	405	24.4	12.5					
ORF	Female	398	39.3	21.0	1.534	801	0.125	2.3	0.11
	Male	405	37.0	21.3					

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Haddiysa, Grade = 2

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
LNR	Female	469	23.0	25.4	-6.206	950	0.000	-11.3	-0.40
	Male	482	34.3	30.5					
FWR	Female	469	5.0	10.6	-5.513	950	0.000	-4.5	-0.36
	Male	482	9.5	14.0					
IWR	Female	469	3.6	9.4	-4.681	950	0.000	-3.1	-0.31
	Male	482	6.7	11.2					
ORF	Female	469	3.7	9.7	-5.514	950	0.000	-4.2	-0.36
	Male	482	8.0	13.6					

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Haddiysa, Grade = 3

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
LNR	Female	478	41.0	32.4	-5.462	940	0.000	-11.5	-0.36
	Male	464	52.5	32.0					
FWR	Female	478	11.0	14.7	-5.638	940	0.000	-5.9	-0.37
	Male	464	16.9	17.3					
IWR	Female	478	8.4	12.6	-4.936	940	0.000	-4.3	-0.32
	Male	464	12.8	14.4					
ORF	Female	478	9.3	14.8	-6.065	940	0.000	-6.5	-0.40
	Male	464	15.8	18.0					

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Sidamu Affo, Grade = 2

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
LNR	Female	487	32.8	29.1	-5.411	978	0.000	-10.6	-0.35
	Male	493	43.4	31.9					
FWR	Female	487	8.6	10.6	-5.499	978	0.000	-4.2	-0.35
	Male	493	12.8	13.4					
IWR	Female	487	6.1	10.3	-4.468	978	0.000	-3.3	-0.29
	Male	493	9.3	12.5					
ORF	Female	487	7.6	12.8	-5.489	978	0.000	-5.3	-0.35
	Male	493	12.9	17.2					

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Sidamu Affo, Grade = 3

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
LNR	Female	441	51.3	33.0	-6.023	874	0.000	-13.0	-0.41
	Male	436	64.3	30.7					
FWR	Female	441	15.4	14.1	-6.290	874	0.000	-6.2	-0.43
	Male	436	21.7	15.2					
IWR	Female	441	11.9	14.2	-5.125	874	0.000	-5.1	-0.35
	Male	436	17.1	15.4					
ORF	Female	441	17.1	18.7	-5.860	874	0.000	-7.6	-0.40
	Male	436	24.7	19.7					

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Tigrigna, Grade = 2

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
LNR	Female	448	28.0	26.8	-2.346	922	0.019	-4.2	-0.15
	Male	476	32.2	27.3					
FWR	Female	448	24.3	15.7	-1.821	922	0.069	-1.9	-0.12
	Male	476	26.2	15.9					
IWR	Female	448	11.4	10.4	-1.019	922	0.309	-0.7	-0.07
	Male	476	12.1	10.2					
ORF	Female	448	15.0	14.3	-1.661	922	0.097	-1.6	-0.11
	Male	476	16.5	14.3					

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Tigrigna, Grade = 3

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
LNR	Female	497	39.9	31.4	-2.636	999	0.009	-5.2	-0.17
	Male	504	45.0	30.5					
FWR	Female	497	33.8	20.4	-3.081	999	0.002	-4.0	-0.19
	Male	504	37.9	21.1					
IWR	Female	497	15.0	12.6	-2.824	999	0.005	-2.2	-0.18
	Male	504	17.2	12.3					
ORF	Female	497	23.4	18.8	-3.087	999	0.002	-3.6	-0.20
	Male	504	27.1	18.6					

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Wolayttatto, Grade = 2

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
LNR	Female	453	31.3	29.1	-1.473	929	0.141	-2.7	-0.10
	Male	478	34.0	26.9					
FWR	Female	453	9.1	15.2	-1.939	929	0.053	-1.9	-0.13
	Male	478	11.1	14.9					
IWR	Female	453	8.6	12.8	-1.401	929	0.161	-1.2	-0.09
	Male	478	9.7	12.9					
ORF	Female	453	10.4	18.3	-1.793	929	0.073	-2.1	-0.12
	Male	478	12.5	17.7					

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

Language = Wolayttatto, Grade = 3

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
LNR	Female	445	44.4	29.0	-2.722	898	0.007	-5.3	-0.18
	Male	455	49.7	29.2					
FWR	Female	445	16.0	17.1	-4.171	898	0.000	-5.3	-0.28
	Male	455	21.3	20.5					
IWR	Female	445	13.8	14.4	-4.710	898	0.000	-4.9	-0.32
	Male	455	18.7	16.8					
ORF	Female	445	17.9	20.5	-4.478	898	0.000	-6.7	-0.30
	Male	455	24.6	24.2					

Notes. LNR is letter name recognition; FWR is familiar word reading; IWR is invented word reading; ORF is oral reading fluency.

APPENDIX 4. INDEPENDENT SAMPLE T-TEST OF THE UNTIMED TASKS BY GRADE

Language = Afaan Oromo

	Grade	N	Mean	Std. Deviation	Mean Difference	Std Dev	Cohen's d
RC	Three	1,117	13.0	18.9	15.5	25.1	0.6
	Two	1,214	28.5	31.2			
ILS	Three	1,118	47.2	41.7	14.0	40.9	0.3
	Two	1,214	61.1	40.1			
LC	Three	1,118	60.4	29.5	10.6	27.6	0.4
	Two	1,214	71.0	25.7			

Notes. RC is reading comprehension; ILS is initial letter sound; LC is listening comprehension.

Language = Aff Somali

	Grade	N	Mean	Std. Deviation	Mean Difference	Std Dev	Cohen's d
RC	Three	817	10.1	17.6	11.1	22.3	0.5
	Two	737	21.2	27.0			
ILS	Three	817	83.2	24.8	6.1	22.0	0.3
	Two	737	89.4	19.1			
LC	Three	817	77.1	25.9	4.9	25.1	0.2
	Two	737	81.9	24.3			

Notes. RC is reading comprehension; ILS is initial letter sound; LC is listening comprehension.

Language = Amharic

	Grade	N	Mean	Std. Deviation	Mean Difference	Std Dev	Cohen's d
RC	Three	782	24.8	25.4	17.1	29.0	0.6
	Two	804	41.9	32.5			
ILS	Three	782	81.7	23.2	4.0	21.6	0.2
	Two	804	85.7	20.0			
LC	Three	782	66.2	25.9	6.5	25.2	0.3
	Two	804	72.7	24.5			

Notes. RC is reading comprehension; ILS is initial letter sound; LC is listening comprehension

Language = Haddiysa

	Grade	N	Mean	Std. Deviation	Mean Difference	Std Dev	Cohen's d
RC	Three	952	12.9	13.6	7.5	17.2	0.4
	Two	943	20.3	20.8			
ILS	Three	952	85.1	27.3	4.3	25.1	0.2
	Two	943	89.4	22.9			
LC	Three	952	73.5	26.6	7.6	24.5	0.3
	Two	943	81.0	22.4			

Notes. RC is reading comprehension; ILS is initial letter sound; LC is listening comprehension.

Language = Sidamu Affo

	Grade	N	Mean	Std. Deviation	Mean Difference	Std Dev	Cohen's d
RC	Three	981	9.2	17.9	10.9	21.4	0.5
	Two	880	20.1	24.9			
ILS	Three	981	90.5	22.4	3.7	19.6	0.2
	Two	880	94.2	16.9			
LC	Three	981	82.9	21.7	4.9	20.1	0.2
	Two	880	87.8	18.5			

Notes. RC is reading comprehension; ILS is initial letter sound; LC is listening comprehension.

Language = Tigrigna

	Grade	N	Mean	Std. Deviation	Mean Difference	Std Dev	Cohen's d
RC	Three	924	17.6	17.7	8.8	21.8	0.4
	Two	1,001	26.3	25.9			
ILS	Three	924	72.6	35.4	10.9	32.2	0.3
	Two	1,001	83.5	28.9			
LC	Three	924	46.3	28.5	16.4	29.0	0.6
	Two	1,001	62.6	29.6			

Notes. RC is reading comprehension; ILS is initial letter sound; LC is listening comprehension.

Language = Wolayttatto

	Grade	N	Mean	Std. Deviation	Mean Difference	Std Dev	Cohen's d
RC	Three	931	11.4	18.0	9.9	20.4	0.5
	Two	901	21.4	22.7			
ILS	Three	931	74.1	29.2	1.0	29.0	0.0
	Two	901	75.1	28.9			
LC	Three	931	52.5	26.4	1.9	26.6	0.1
	Two	901	54.4	26.8			

Notes. RC is reading comprehension; ILS is initial letter sound; LC is listening comprehension.

APPENDIX 5. INDEPENDENT SAMPLE T-TEST OF THE UNTIMED TASKS BY GENDER

Language = Afaan Oromo, Grade = 2

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
PA (%)	Female	572	43.6	40.7	-2.902	1116	0.004	-7.2	-0.17
	Male	546	50.9	42.4					
RC (%)	Female	571	12.1	19.0	-1.505	1115	0.133	-1.7	-0.09
	Male	546	13.8	18.8					
LC (%)	Female	572	60.6	29.9	0.183	1116	0.855	0.3	0.01
	Male	546	60.3	29.2					

Notes. PA is phonological awareness; RC is reading comprehension; LC is listening comprehension.

Language = Afaan Oromo, Grade = 3

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
PA (%)	Female	606	55.5	40.6	-4.955	1212	0.000	-11.3	-0.28
	Male	608	66.8	38.8					
RC (%)	Female	606	24.0	29.5	-4.973	1212	0.000	-8.8	-0.29
	Male	608	32.9	32.3					
LC (%)	Female	606	69.9	26.8	-1.508	1212	0.132	-2.2	-0.09
	Male	608	72.1	24.4					

Notes. PA is phonological awareness; RC is reading comprehension; LC is listening comprehension.

Language = Aff Somali, Grade = 2

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
PA (%)	Female	340	82.1	25.1	-1.080	815	0.280	-1.9	-0.08
	Male	477	84.0	24.6					
RC (%)	Female	340	7.9	13.2	-3.091	815	0.002	-3.8	-0.23
	Male	477	11.7	19.9					
LC (%)	Female	340	74.4	26.8	-2.501	815	0.013	-4.6	-0.18
	Male	477	79.0	25.0					

Notes. PA is phonological awareness; RC is reading comprehension; LC is listening comprehension.

Language = Aff Somali, Grade = 3

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
PA (%)	Female	285	87.8	20.3	-1.723	735	0.085	-2.5	-0.13
	Male	452	90.3	18.3					
RC (%)	Female	285	15.5	24.5	-4.606	735	0.000	-9.3	-0.35
	Male	452	24.8	27.8					
LC (%)	Female	285	76.1	27.5	-5.279	735	0.000	-9.5	-0.39
	Male	452	85.6	21.2					

Notes. PA is phonological awareness; RC is reading comprehension; LC is listening comprehension.

Language = Amharic, Grade = 2

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
PA (%)	Female	403	82.2	22.9	0.688	780	0.491	1.1	0.05
	Male	379	81.1	23.6					
RC (%)	Female	403	26.9	26.5	2.378	780	0.018	4.3	0.17
	Male	379	22.6	24.0					
LC (%)	Female	403	67.7	25.2	1.594	780	0.111	2.9	0.11
	Male	379	64.7	26.6					

Notes. PA is phonological awareness; RC is reading comprehension; LC is listening comprehension.

Language = Amharic, Grade = 3

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
PA (%)	Female	398	86.2	18.9	0.708	801	0.479	1.0	0.05
	Male	405	85.2	21.0					
RC (%)	Female	398	44.1	33.4	1.850	801	0.065	4.2	0.13
	Male	405	39.8	31.5					
LC (%)	Female	398	72.6	25.1	-0.059	801	0.953	-0.1	0.00
	Male	405	72.7	23.8					

Notes. PA is phonological awareness; RC is reading comprehension; LC is listening comprehension.

Language = Haddiysa, Grade = 2

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
PA (%)	Female	469	83.1	29.3	-2.272	950	0.023	-4.0	-0.15
	Male	482	87.1	25.1					
RC (%)	Female	469	10.8	11.2	-4.635	950	0.000	-4.0	-0.30
	Male	482	14.9	15.3					
LC (%)	Female	469	70.4	28.0	-3.556	950	0.000	-6.1	-0.23
	Male	482	76.5	24.8					

Notes. PA is phonological awareness; RC is reading comprehension; LC is listening comprehension.

Language = Haddiysa, Grade = 3

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
PA (%)	Female	478	88.8	23.8	-0.886	940	0.376	-1.3	-0.06
	Male	464	90.1	22.0					
RC (%)	Female	478	17.4	18.5	-4.407	940	0.000	-5.9	-0.29
	Male	464	23.3	22.7					
LC (%)	Female	478	78.2	24.7	-3.983	940	0.000	-5.8	-0.26
	Male	464	83.9	19.4					

Notes. PA is phonological awareness; RC is reading comprehension; LC is listening comprehension.

Language = Sidamu Affo, Grade = 2

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
PA (%)	Female	487	89.4	23.9	-1.421	978	0.156	-2.0	-0.09
	Male	493	91.5	20.8					
RC (%)	Female	487	6.7	15.0	-4.338	978	0.000	-4.9	-0.28
	Male	493	11.7	20.0					
LC (%)	Female	487	82.9	22.1	-0.050	978	0.960	-0.1	0.00
	Male	493	83.0	21.2					

Notes. PA is phonological awareness; RC is reading comprehension; LC is listening comprehension.

Language = Sidamu Affo, Grade = 3

	Gender	**N**	**Mean**	**Std. Deviation**	**t**	**DF**	**Sig. (2-tailed)**	**Mean Difference**	**Cohen's d**
PA (%)	Female	441	94.7	15.4	0.931	874	0.352	1.1	0.06
	Male	436	93.6	18.3					
RC (%)	Female	441	16.7	23.6	-4.205	874	0.000	-7.0	-0.28
	Male	436	23.7	25.8					
LC (%)	Female	441	89.7	16.6	2.953	874	0.003	3.7	0.20
	Male	436	86.0	20.1					

Notes. PA is phonological awareness; RC is reading comprehension; LC is listening comprehension.

Language = Tigrigna, Grade = 2

	Gender	**N**	**Mean**	**Std. Deviation**	**t**	**DF**	**Sig. (2-tailed)**	**Mean Difference**	**Cohen's d**
PA (%)	Female	448	69.7	36.5	-2.409	922	0.016	-5.6	-0.16
	Male	476	75.3	34.1					
RC (%)	Female	448	15.9	16.7	-2.842	922	0.005	-3.3	-0.19
	Male	476	19.2	18.4					
LC (%)	Female	448	44.4	28.3	-1.965	922	0.050	-3.7	-0.13
	Male	476	48.0	28.6					

Notes. PA is phonological awareness; RC is reading comprehension; LC is listening comprehension.

Language = Tigrigna, Grade = 3

	Gender	**N**	**Mean**	**Std. Deviation**	**t**	**DF**	**Sig. (2-tailed)**	**Mean Difference**	**Cohen's d**
PA (%)	Female	497	81.3	30.8	-2.425	999	0.015	-4.4	-0.15
	Male	504	85.7	26.8					
RC (%)	Female	497	24.6	25.9	-2.165	999	0.031	-3.5	-0.14
	Male	504	28.1	25.8					
LC (%)	Female	497	60.8	29.5	-1.958	999	0.051	-3.7	-0.12
	Male	504	64.4	29.5					

Notes. PA is phonological awareness; RC is reading comprehension; LC is listening comprehension.

Language = Wolayttatto, Grade = 2

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
PA (%)	Female	453	72.0	30.3	-2.081	929	0.038	-4.0	-0.14
	Male	478	76.0	28.0					
RC (%)	Female	453	11.9	24.4	-1.417	929	0.157	-2.3	-0.09
	Male	478	14.2	25.3					
LC (%)	Female	453	52.3	27.0	-0.199	929	0.842	-0.3	-0.01
	Male	478	52.6	25.8					

Notes. PA is phonological awareness; RC is reading comprehension; LC is listening comprehension.

Language = Wolayttatto, Grade = 3

	Gender	N	Mean	Std. Deviation	t	DF	Sig. (2-tailed)	Mean Difference	Cohen's d
PA (%)	Female	445	75.0	28.7	0.025	898	0.980	0.0	0.00
	Male	455	75.0	29.0					
RC (%)	Female	445	22.7	29.3	-3.809	898	0.000	-8.2	-0.26
	Male	455	30.8	34.8					
LC (%)	Female	445	54.8	25.8	0.471	898	0.638	0.8	0.03
	Male	455	54.0	27.7					

Notes. PA is phonological awareness; RC is reading comprehension; LC is listening comprehension.

APPENDIX 6. MEAN SCORES OF ORAL READING FLUENCY BY ZONES

	Grade 2			Grade 3		
Zone	Mean	N	Std. Deviation	Mean	N	Std. Deviation
Agew Awi	34.7	24	21.1	49.3	23	16.4
East Gojjam	25.1	165	15.6	37.2	158	21.1
North Shewa	33.4	83	15.4	43.2	96	19.6
North Wollo	18.0	62	15.4	27.1	55	19.5
South Gondar	22.7	128	19.3	33.6	82	20.8
South Wollo	25.5	140	18.8	36.1	179	20.9
West Gojjam	22.9	180	17.3	41.6	210	21.5
Arsi	12.3	370	13.6	23.1	344	20.3
Bale	11.5	188	13.5	23.3	216	22.7
East Hararghe	13.1	89	16.5	17.9	76	18.9
East Shewa	11.0	45	14.2	18.3	37	19.5
East Welega	9.8	107	15.7	20.9	112	21.1
Guji	8.4	145	11.9	14.5	194	18.2
Illubabor	8.0	63	11.4	19.6	77	19.8
West Haraghe	11.9	68	17.2	28.0	106	28.2
West Welega	9.3	43	11.0	23.1	52	21.6
Fafan	10.7	684	18.3	21.5	585	21.7
Sitti	9.6	133	14.0	15.1	152	16.4
Hadiya	5.9	952	12.0	12.5	943	16.8
Sidama	10.3	981	15.4	20.8	880	19.6
Wolayita	11.4	931	18.0	21.4	901	22.7
Central	18.8	218	15.2	26.5	181	19.8
Northwestern	14.4	419	13.9	23.3	480	18.6
Southern	15.5	287	14.0	27.3	340	18.2
Total	12.5	6,504	16.5	22.5	6,481	21.4

APPENDIX 7. MEAN SCORES OF READING COMPREHENSION BY ZONES

	Grade 2			Grade 3		
Zone	**Mean**	**N**	**Std. Deviation**	**Mean**	**N**	**Std. Deviation**
Agew Awi	38.95	24	28.501	64.00	23	35.175
East Gojjam	25.57	165	23.888	41.71	158	31.454
North Shewa	33.87	83	26.209	45.86	96	32.166
North Wollo	13.74	62	19.132	23.44	55	28.689
South Gondar	22.85	128	24.412	32.73	82	28.742
South Wollo	25.95	140	27.856	38.86	179	32.054
West Gojjam	22.33	180	24.660	48.88	210	32.649
Arsi	14.23	370	20.394	31.77	344	30.333
Bale	14.26	188	19.225	28.58	216	31.970
East Hararghe	13.69	89	16.634	23.68	76	26.772
East Shewa	13.43	45	19.638	23.72	37	29.166
East Welega	11.34	107	19.708	26.24	112	31.118
Guji	9.62	145	14.650	19.27	194	24.986
Illubabor	10.66	63	17.365	26.79	77	29.185
West Haraghe	14.66	68	22.295	40.80	106	40.519
West Welega	10.44	43	14.785	32.75	52	33.362
Fafan	10.22	684	17.903	23.15	585	28.093
Sitti	9.54	133	15.658	13.68	152	20.505
Hadiya	12.86	952	13.586	20.31	943	20.833
Sidama	9.23	981	17.895	20.08	880	24.933
Wolayita	13.11	931	24.877	26.91	901	32.575
Central	20.22	218	19.478	27.14	181	26.645
Northwestern	15.98	419	16.508	24.25	480	25.734
Southern	17.93	287	17.694	28.87	340	25.488
Total	14.13	6,504	20.186	26.43	6,481	28.963

APPENDIX 8. RELATIONSHIP BETWEEN ORAL READING FLUENCY AND READING COMPREHENSION

Grade	RC (%)	Afaan Oromo	Aff Somali	Amharic	Haddiysa	Sidamu Affo	Tigrigna	Wolayttatto
Gr 2	0%	4.8	4.7	7.7	1.5	2.9	1.6	2.2
	20%	22.6	34.9	26.9	21.0	21.5	20.7	25.2
	40%	31.8	34.8	34.9	31.5	31.0	28.5	27.5
	60%	42.5	55.8	43.5	49.7	39.6	39.7	45.6
	80%	45.2	55.6	68.5	57.2	59.2	53.9	50.6
	100%	68.4	93.6	64.0	63.0	77.1	54.4	70.6
Gr 3	0%	5.43	3	6.96	0	11.37	8	2.5
	20%	20.88	23	32.02	20	30.24	28	22.26
	40%	31.75	43	35.86	40	38.82	48	31.76
	60%	45.33	63	50.26	60	46.68	68	51.41
	80%	52.36	83	63.77	80	65.7	88	49.04
	100%	67.78	100	70.65	100	65.45	100	68.33

APPENDIX 9. COMPARISON OF 2014, 2016, AND 2018 SCORES IN ORAL READING FLUENCY

		2016			2014					
	Language	**ORF Mean**	**Std. Dev.**	**N**	**ORF Mean**	**Std. Dev.**	**N**	**Sig. (2-tailed)**	**Mean Difference**	**Cohen's d**
Gr 2	Afaan Oromo	9.8	14.8	898	12.1	14.8	898	0.00	-2.3	-0.16
	Aff Somali	6.4	12.9	709	20.4	12.9	709	0.00	-14	-1.09
	Amharic	28.7	18.4	903	19.2	18.4	903	0.00	9.5	0.52
	Haddiysa	7.5	15.27	898	6.5	15.27	898	0.17	1	0.07
	Sidamu Affo	16.3	19.98	973	7.1	19.98	973	0.00	9.2	0.46
	Tigrigna	16.4	16.77	886	13.3	16.77	886	0.00	3.1	0.18
	Wolayttotta	30.8	24.05	952	11.2	24.05	952	0.00	19.6	0.82
Gr 3	Afaan Oromo	21.2	22.5	900	23.9	22.5	900	0.01	-2.7	-0.12
	Aff Somali	16.5	18.6	662	32	18.6	662	0.00	-15.5	-0.83
	Amharic	40.5	20.6	881	30	20.6	881	0.00	10.5	0.51
	Haddiysa	14.4	19.16	895	11.5	19.16	895	0.00	2.9	0.15
	Sidamu Affo	27.1	20.84	819	14.4	20.84	819	0.00	12.7	0.61
	Tigrigna	26.2	18.94	900	24.2	18.94	900	0.03	2	0.11
	Wolayttotta	40.8	24.39	848	20.1	24.39	848	0.00	20.7	0.85

		2018			2016					
	Language	**ORF Mean**	**Std. Dev.**	**N**	**ORF Mean**	**Std. Dev.**	**N**	**Sig. (2-tailed)**	**Mean Difference**	**Cohen's d**
Gr 2	Afaan Oromo	11.05	13.96	572	9.8	14.8	898	0.11	1.25	0.09
	Aff Somali	10.56	17.68	340	6.4	12.9	709	0.00	4.16	0.28
	Amharic	24.86	17.76	403	28.7	18.4	903	0.00	-3.84	-0.21
	Haddiysa	5.86	12.02	469	7.5	15.27	898	0.04	-1.64	-0.12
	Sidamu Affo	10.31	15.43	487	16.3	19.98	973	0.00	-5.99	-0.32
	Tigrigna	15.77	14.3	448	16.4	16.77	886	0.50	-0.63	-0.04
	Wolayttotta	11.44	18.02	453	30.8	24.05	952	0.00	-19.36	-0.87
Gr 3	Afaan Oromo	21.29	21.5	546	21.2	22.5	900	0.94	0.09	0.00
	Aff Somali	20.17	20.84	477	16.5	18.6	662	0.00	3.67	0.19
	Amharic	38.13	21.16	379	40.5	20.6	881	0.06	-2.37	-0.11
	Haddiysa	12.5	16.77	482	14.4	19.16	895	0.07	-1.9	-0.10
	Sidamu Affo	20.82	19.56	493	27.1	20.84	819	0.00	-6.28	-0.31
	Tigrigna	25.25	18.76	476	26.2	18.94	900	0.37	-0.95	-0.05
	Wolayttotta	21.37	22.72	478	40.8	24.39	848	0.00	-19.43	-0.82

APPENDIX 10. COMPARISON OF 2014, 2016, AND 2018 SCORES IN READING COMPREHENSION

		2016			2014					
	Language	RC Mean	Std. Dev.	N	RC Mean	Std. Dev.	N	Sig. (2-tailed)	Mean Difference	Cohen's d
Gr 2	Afaan Oromo	10.5	20.59	898	14	20.59	898	0.00	-3.5	-0.17
	Aff Somali	6.5	16.22	709	28	16.22	709	0.00	-21.5	-1.33
	Amharic	25.9	26.27	903	18	26.27	903	0.00	7.9	0.30
	Haddiysa	8.6	20.47	898	12	20.47	898	0.00	-3.4	-0.17
	Sidamu Affo	17.4	23.24	973	20	23.24	973	0.01	-2.6	-0.11
	Tigrigna	13.8	18.28	886	14	18.28	886	0.82	-0.2	-0.01
	Wolayttotta	43.4	35.82	952	24	35.82	952	0.00	19.4	0.54
Gr 3	Afaan Oromo	27.6	32.41	900	28	32.41	900	0.79	-0.4	-0.01
	Aff Somali	20	27.74	662	44	27.74	662	0.00	-24	-0.87
	Amharic	42.6	31.88	881	32	31.88	881	0.00	10.6	0.33
	Haddiysa	16.2	26.36	895	22	26.36	895	0.00	-5.8	-0.22
	Sidamu Affo	32.9	28.98	819	20	28.98	819	0.00	12.9	0.45
	Tigrigna	25.5	25.57	900	22	25.57	900	0.00	3.5	0.14
	Wolayttotta	55.3	34.54	848	40	34.54	848	0.00	15.3	0.44

		2018			2016					
	Language	RC Mean	Std. Dev.	N	RC Mean	Std. Dev.	N	Sig. (2-tailed)	Mean Difference	Cohen's d
Gr 2	Afaan Oromo	12.97	18.94	572	10.5	20.59	898	0.02	2.47	0.12
	Aff Somali	10.11	17.55	340	6.5	16.22	709	0.00	3.61	0.22
	Amharic	24.78	25.41	403	25.9	26.27	903	0.47	-1.12	-0.04
	Haddiysa	12.86	13.59	469	8.6	20.47	898	0.00	4.26	0.23
	Sidamu Affo	9.23	17.89	487	17.4	23.24	973	0.00	-8.17	-0.38
	Tigrigna	17.59	17.68	448	13.8	18.28	886	0.00	3.79	0.21
	Wolayttotta	13.11	24.88	453	43.4	35.82	952	0.00	-30.29	-0.93
Gr 3	Afaan Oromo	28.45	31.22	546	27.6	32.41	900	0.62	0.85	0.03
	Aff Somali	21.19	26.96	477	20	27.74	662	0.47	1.19	0.04
	Amharic	41.91	32.52	379	42.6	31.88	881	0.73	-0.69	-0.02
	Haddiysa	20.31	20.83	482	16.2	26.36	895	0.00	4.11	0.17
	Sidamu Affo	20.08	24.93	493	32.9	28.98	819	0.00	-12.82	-0.47
	Tigrigna	26.34	25.88	476	25.5	25.57	900	0.56	0.84	0.03
	Wolayttotta	26.91	32.58	478	55.3	34.54	848	0.00	-28.39	-0.84

REFERENCES

Almond, R. G., & Sinharay, S. (2012). *What can repeated cross-sectional studies tell us about student growth?* ETS Research Report Series. Princeton, NJ: ETS.

Cohen, J. (1977). *Statistical power analysis for behavioral sciences (revised ed.)*. New York, NY: Academic Press.

Fuchs, L., Fuchs, D., Hosp, K., & Jenkins, J. (2001). *Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis.* Scientific Studies of Reading, 5(3), 239–256.

Gove, Amber. (2009). *Effective Reading for All and the Early Grade Reading Assessment*.

Government of Ethiopia, UN Children's Fund, Save the Children, Education Cluster (2017). *Education in Emergency Strategic Response Plan - 2017*

Hirsch, E. D., Jr. (2003). *Reading Comprehension Requires Knowledge--of Words and the World. American Educator*, v27 n1 p10-13

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

Kamhi, A.G., & Catts, H. W. (1991). *Language and reading: Convergences, divergences, and development*. In A. G. Kamhi & H. W. Catts (Eds.), Reading disabilities: A developmental language perspective (pp. 1–34). Toronto, Ontario, Canada: Allyn & Bacon.

McBride-Chang, C., & Kail, R. V. (2002). *Cross-cultural similarities in the predictors of reading acquisition*. Child Development, 73(5), 1392-1407

McBride-Chang, C., & Ho, C. S.-H. (2000). *Developmental issues in Chinese children's character acquisition*. Journal of Educational Psychology, 92(1), 50-55

National Reading Panel (U.S.), & National Institute of Child Health and Human Development (U.S.). (2000). *Report of the National Reading Panel: Teaching children to read: an evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: reports of the subgroups*. Washington, D.C.: National Institute of Child Health and Human Development, National Institutes of Health.

O’Maggio, A. (1986). *A proficiency-oriented approach to listening and reading*. In A. O’Maggio (Ed.), Teaching Language in Context, (pp. 121–174). Boston, MA: Henile & Heinle.

Seymour H.K., Philip & Aro, Mikko & M Erskine, Jane. (2003). *Foundation literacy acquisition in European orthographies [Electronic version]*. British Journal of Psychology. 94. 143-174

Snow, C.E., Burns, M.S., & Griffin, P. (eds.) (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press, 432 pp.

Thomas, S. M., Salim, M., & Jung Peng, W. (2013). *Monitoring and evaluating school effectiveness: The case of longitudinal data sets.* In L. Tikley & A. M. Barrett (Eds.), Education Quality and Social Justice in the Global South: Challenges for Policy, Practice and Research. New York, NY: Routledge.

USAID Ethiopia, Global Reading Network, Ministry of Education Ethiopia (2015). *Results of the Early Grade Reading Benchmarking Workshop in Ethiopia*

Wolf, F. M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Beverly Hills, CA: Sage.