USAID | ETHIOPIA
FROM THE AMERICAN PEOPLE

USAID Reading for Ethiopia's Achievement Developed Monitoring and Evaluation (READ M&E)

Early Grade Reading Assessment (EGRA) 2016 Midterm Report

October 17, 2016

Submitted by
American Institutes for Research (AIR)

Contract No. AID 663 C 15 00001

USAID Reading for Ethiopia's Achievement Developed Monitoring and Evaluation (READ M&E)

Early Grade Reading Assessment (EGRA)

2016 Midterm Report

Submitted by:

American Institutes for Research®

October 17, 2016

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

Acknowledgements

1 Overview

1.1 Executive Summary of 2016

READ M&E administered the Early Grade Reading Assessment (EGRA) in June 2016, to evaluate the reading fluency and comprehension levels of Grades 2 and 3 students in seven languages: Afan Oromo, Aff Somali, Amharic, Hadiyissa, Sidamu Afoo, Wolayttatto and Tigrigna- in five regions. READ M&E collected data from 315 schools (45 schools from each language) and assessed a total of 12,242 Grade 2 and 3 students.

On the 2016 EGRA, 16% percent of Amharic speaking students and 37% of Wolayttatto students achieved the top level of reading proficiency – reading fluently with full comprehension. In the other five of the seven languages, less than 10% of the students in assessed grades achieved that desirable benchmark of reading proficiency. However, considering the top two benchmark levels combined (reading fluently with full comprehension and reading with increasing fluency and comprehension), it can be observed that overall about 40% of students in assessed grades and languages are exhibiting relatively functional reading proficiency levels (ranging from 15% in Aff Somali to 68% in Wolayttatto).

Comparison between 2014 and 2016 oral reading fluency results: Comparisons of the average ORF scores between the 2014 baseline and the 2016 midterm EGRA by language shows somewhat mixed results. Substantial gains were observed in three languages: Wolayttatto, Sidamu Afoo, and Amharic, with difference sizes (Cohen's D) of 0.83, 0.54, and 0.51, respectively, which are considered as practically significant effect sizes in educational field indicating that something really changed (Wolf, 1986). Relatively negligible differences were observed in three languages: Tigrigna, Hadiyissa, and Afan Oromo (difference sizes of 0.14, 0.11, and -0.13, respectively), whereas a substantial drop in ORF scores was observed in one language: Aff Somali, with difference size of -0.94. Similar pattern is observed in Reading Comprehension (RC) scores. Thus, a substantial gain in three languages, no differences in three languages, and a substantial drop only in one language could be considered as a relatively positive overall score.

Grade level results: In each language, the mean differences between the two grade levels in all EGRA sub-tasks 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, which indicates positive grade gain. The average size of grade differences across all languages measured by Cohen's D is 0.54 for Oral Reading Fluency (ORF) and0.51 for Reading Comprehension (RC), which falls deep into educationally significant

effect size indicating that substantial reading gains are happening between grades 2 and 3. It is worthwhile to mention that reading gains from grade 2 to 3were also strong in 2014, but still smaller than in 2016 (average Cohen's D across languages was 0.51 for ORF and 0.43 for RC), which suggests that educational effect on reading in grade 3 is on increase.

Gender: In two languages (Somali and Hadiyissa) boys performed significantly better than girls (average effect sizes across all EGRA subtasks are 0.29 and 0.25, respectively), and in two languages (Oromo and Sidamu) girls significantly outperformed boys (average effect sizes 0.25 and 0.23, respectively). In the other three languages, gender differences were small or negligible.

Comparisons between Reading Comprehension and Listening comprehension scores: Listening comprehension is higher than reading comprehension in all seven languages (average percent-correct rate for reading comprehension questions is 25%, whereas for listening comprehension it is 75%). This difference is a consequence of the amount of information received through these two perception modes, in other words, inability to read a given text restricts the exposure to information, so there is less material to comprehend compared to listening mode where children are exposed to full information contained in the stimulus text.

Relationship between oral reading fluency and reading comprehension scores: In each language, a strong relationship exists between oral reading fluency and reading comprehension scores, which is again due 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 oral reading fluency as a condition for learning.

Performance on Subtasks: In all the seven EGRA subtasks, Wolayttatto speaking children performed the highest. Amharic speaking children followed and Af-Somali speaking children scored the lowest.

The mean score for correct words per minute in oral reading fluency range from 11.5 (Aff Somali) to 35.8 (Wolayttatto). The percentage of students who stopped trying to read words (non-readers) ranges from 4.4% in Amharic to 54% in Af-Somali. High proportions of students from Hadiyissa (37.7%) and Afan Oromo (26.2%) also discontinued the subtask.

In all seven languages, children were most able to complete the 'naming the letter' subtask, 33.0 (Aff Somali) to 68.5 (Wolayttatto). This indicates a strong alphabetic knowledge.

The untimed subtasks are Phonemic Awareness (PA), Reading Comprehension (RC) and Listening Comprehension (LC). The mean reading comprehension scores range from 9.8% (Aff Somali) to 51% (Wolayttatto). In each language, the mean reading comprehension scores were less than the other two subtasks. In phonemic awareness, the mean scores range from 38% (Aff Somali) to 97% (Sidamu Afoo) and in listening comprehension the mean scores range from 56.7% (Aff Somali) to 83.2% (Afan Oromo).

The subtask that showed the largest proportions of discontinuity was invented words reading. Students discontinued this task in the range of 12.7% in Amharic to 65.2% in Aff Somali. In a positive trend, when viewed by grade level, the proportion of students who discontinued in the oral reading fluency subtasks decreases in Grade 3.

Conclusions: Considering that the ultimate goal of reading is to construct meaning, the findings of the 2016 EGRA suggest that, although some students were able to respond correctly to the Reading Comprehension subtask questions, it is worrying their number is low. A high percentage of Ethiopian students cannot read enough words with in 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. However, in spite of methodological differences between baseline and mid-term EGRA, it is encouraging impression that reading scores improved over time in about half of analyzed languages. In addition, relatively large differences between student performance in grades 2 and 3 suggest that these substantial gains in reading skills could be considered as positive effects of reading instruction.

1.2 History and Purpose of EGRA in Ethiopia

The EGRA instrument consists of a set of subtasks designed to assess foundational reading skills crucial to becoming a fluent reader. EGRA was designed to help USAID Partner countries measure systematically how well primary school children were acquiring reading skills. EGRA is designed to be a method-independent approach to assessment (i.e., the instrument does not reflect a particular method of reading instruction). Instead, EGRA measures the basic skills that a child must build to be able to read fluently and with comprehension—the ultimate goal of reading. EGRA subtasks are based on research regarding a comprehensive approach to reading acquisition across languages.

EGRA, as an assessment, is not designed to provide information on individual children's progress toward learning to read and/or measure the early reading performances in the education systems as a whole. It cannot be used to evaluate teachers or headmaster performance. Nor can the EGRA be used to measure

gains within a short time period. EGRA is a one to one, orally administered assessment targeted at measuring the pre-reading and reading skills foundational to later reading (and academic success).

The EGRA tool used in Ethiopia has four-timed (Letter Name Knowledge, Familiar Words Reading, Invented Words Reading, Passage Reading Fluency) and three untimed activities (Phonemic Awareness, Reading Comprehension and Listening Comprehension). These skills are phonological awareness, decoding, reading fluency, reading comprehension, and listening comprehension. The EGRA is designed to help teachers, parents, education officials and policy makers, as well as donors to prioritize areas of intervention to improve foundational skills in early grades. However, EGRA cannot be a high-stakes accountability tool and cannot be used for direct cross-language comparisons.

Since 2007, EGRA has been adapted and administered in over 120 languages by more than 30 organizations in over 70 countries in the world. USAID funded projects have conducted EGRA in 36 languages in 23 countries between the year2006 and mid of 2015.

1.2.1 EGRA in Ethiopia

In Ethiopia, the EGRA was conducted for the first time in May and June 2010, by the collaborative efforts of RTI and IQPEP with the MOE. The EGRA was conducted in eight regions, for six languages: Tigrigna, Afaan Oromo, Amharic, Aff Somali, Sidamu Afoo, and Hararigna (approximately 90% of the population speaks at least one of these languages).After the results of the 2010 EGRA, USAID, the Ministry of Education, 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 early grades in order to enable greater learning in upper grades.

IQPEP in collaboration with the MoE and RSEBs conducted a second EGRA in May 2013 to assess formatively 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 pre-reading skills in Ethiopia, though the progress is slow and shows significant variation depending on the language and region. In June 2014, RTI conducted a baseline EGRA for Hadiyissa and Wolayittatto languages. The results obtained revealed that some Hadiyissa and Wolayittatto-speaking students were only beginning to learn to read in their respective language by Grade 3.

The chart below shows the history of the EGRA in Ethiopia:

Table 1: EGRA in Ethiopia

	Date	Conducted by	Languages	Sample size	Data collection period
1	2010	RTI, IQPEP and MoE	Six (Amharic, Afaan Oromo, Tigrigna, Sidamu Afoo, Hararigna, Aff Somali)	8 regions, 90 woreda, 338 schools, 13,079 Grade 2 &3 students	May 10- June 16, 2010
2	2013	FHI360/ IQPEP	Six (Amharic, Afaan Oromo, Tigrigna, Sidamu Afoo, Aff Somali)	8 regions, 53 WEO, 240 (120 controlled) schools, 9406 (4699 controlled) students	May 2013
3	2014	FHI 360/IQPEP	Six (Amharic, Afaan Oromo, Tigrigna, Sidamu Afoo, Hararigna, Aff Somali)	8 regions, 53 WEO, 240 (120 controlled) schools, 9406 (4699 controlled) students	May 2014
4	2014	RTI	Two (Hadiyissa and Wolaytatto)	Two Zones (Hadiya and Wolayta) 2000.00 students	June 2014
5	2016	AIR/READ M&E	Seven (Amharic, Afaan Oromo, Aff Somali, Tigrigna, Sidamu Afoo, Hadiyissa and Wolaytatto)	5 Regions, 7 languages, 13,475 grades 2 & 3 students	May 23-June 12, 2016

1.3 EGRA limitations

EGRA is a set of subtasks that measure foundational skills that are predictive of later reading success. It is not intended to be a high-stakes accountability measure to determine student grade promotion or to evaluate individual teachers. EGRA is designed to complement, rather than replace, existing curriculum-based pencil-and-paper assessments. However, due to the constraints imposed by children's limited attention span and stamina, neither EGRA nor any other single instrument is capable of measuring all skills required for students to read with comprehension. EGRA is not intended to be an instructional program, but rather is capable of informing instructional programs. EGRA cannot fully determine background or literacy behaviors that could influence a student's ability to read.

1.4 Seven subtasks of the EGRA

Skills measured by the EGRA are phonological awareness, decoding, reading fluency, reading comprehension, and listening comprehension. Below is a brief description of each of the seven EGRA components:

The *Letter Name Recognition* (LNR) subtask assessed knowledge of the alphabetic principle, the foundation of learning to read. The alphabetic principle is the understanding that words are composed of sounds (i.e., phonemes). Letters (i.e. graphemes) are symbols that represent those sounds. When children understand that sounds correspond to letters (develop phonological awareness), they can begin to learn to decode words (McBride-Chang & Kail, 2002; 2004; McBride-Chang & Ho, 2000). Research in other languages has suggested that reading skills progress only after 80% of letters (fidels) are mastered (Seymour et al., 2003). EGRA measures the ability to read the letters of the alphabet without hesitation and naturally. This timed test assesses automaticity and fluency of letter names. It is timed to 1 minute, which saves time and prevents children having to spend time on something that is difficult for them.

The *Initial Letter Sound* (ILS) subtask was an assessment of phonemic awareness. 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 primarily on literacy in the English language) determined that skill in phoneme identification and phonological awareness is 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 ILS subtask, the student stimulus included a list of the 10 most frequently used letters randomly arranged. The frequency of letters in everyday use was determined during development of the subtask by text analysis and calculations of word count frequencies. The administrator read each word two times and asked the pupils to make the first sound of the word. If a pupil did not answer within 3 seconds, a response "no answer" was recorded. The maximum score for this section was 10 points, 1 point for each correct answer.

Familiar Word Recognition (FWR) assessed the ability to recognize and read high-frequency words. Frequency of words was determined through a word count analysis of the most commonly used words in textbooks of appropriate level. The list of words was derived from the fifty most frequently used words in the Grade 2 textbooks. For this task, EGRA assessors were able to attain a measure of decontextualized

decoding skills 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: The words were then randomly arranged in the pupil stimulus. The FWR tasks were scored on a *words per minute* calculation that called for the administrator to determine how many words were attempted, how many were read correctly, and in what time over the course of 60 seconds.

Invented Word Recognition (IWR) assessed the ability of pupils to decode one- and two-syllable non-words that could plausibly exist in the language in question. The NWR task provided a measure of decoding related to that of the *Familiar Word Recognition* task but had the advantage of not allowing respondents to *sight-read* words. To achieve in reading, pupils need to acquire both sight-reading and decoding skills. According to Hirsch (2003), there is significant evidence that an over reliance on "sight word vocabulary" often leads to regression in reading development by age 9 or 10.

Fifty non-words were randomly arranged on a list in the pupil booklets and participants were asked to read as many as they could. The NWR task was graded on a *words per minute* calculation that called for administrator to determine how many words were attempted, how many were read correctly, and in what time frame on this 60 second task.

Oral Reading Fluency can be best understood as the ability to read with speed, accuracy, and proper expression. The purpose of the timed *Oral Reading Fluency* subtask was to examine whether pupils in Grades 2 and 3 were able to 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 was 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* reading fluency is considered an important precursor to reading comprehension, fluency alone is not an indicator of reading comprehension; nonetheless, it is an important foundational skill.

In 2016, the *Oral Reading Fluency* task included paragraphs with about 60 words. In subtask design, test developers conducted textbook reviews to determine what words could be considered grade appropriate. The stories were created to be appropriate for particular 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, how many were read correctly over a 60-second period.

The *Reading Comprehension* subtask, which relied on questions about the text, read in the *Oral Reading Fluency* subtask, determined understanding of the text and the ability of pupils to answer factual questions and make inferences based on what they read. After a pupil completed the *Oral Reading Fluency* subtask, the administrator then moved to the *Reading Comprehension* task that was a series of questions about the passage just read.

Research indicates that the ability to correctly understand and interpret oral stimuli (linguistic comprehension) and make meaning from what is heard is a core skill related to reading comprehension (Hoover & Gough, 1986; Kamhi & Catts, 1991). In this EGRA subtask, the child demonstrated *Listening Comprehension* 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), some of the core dimensions of listening are retaining parts of language in short-term memory, discriminating among distinctive sounds, detecting key ideas, and guessing meaning from context.

The subtasks included a paragraph of approximately 40 words. The test administrator read the passage aloud only once at a pace of about one word per second. When the text was completed, students were asked five oral comprehension questions.

2 2016 EGRA Design

This section details how the 2016 EGRA tools was developed, the methods used for equating the 2014 and 2016 tools and how the 2016 tool was piloted. As this was the first time in Ethiopia that the EGRA data was collected electronically significant discussion is given to how the new method of data collection performed. This section also covers the sample size, and zones and school selection.

2.1 Objectives of the 2016 Ethiopian EGRA

The 2016 EGRA measured the emerging reading skills of approximately 14,000 children in seven mother tongues in five regions of Ethiopia. The 2016 EGRA tool was equated to the 2014 EGRA tool in order to compare the two administrations. It is not possible to compare across the seven languages as each language has their own unique span of time toward fluency. As the READ suite of interventions is not yet

fully implemented, this 2016 EGRA cannot be taken as a measurement of impact nor can it be directly tied to implementation. However, scores can be compared to previous EGRAs.

2.2 Development of the 2016 EGRA tool

As the EGRA tool had been used several times prior to 2016, AIR/READ M&E revised the EGRA tools in seven mother tongues. To do this, AIR hired local mother tongue language experts who had worked on the development of the 2010 EGRA tool. These experts each constructed two EGRA tools: one tool A) had completely new items and the other tool B) had between 15%-74% new items, depending on the language. In November, READ M&E brought together representatives from the Ministry of Education (MoE), Regional State Education Bureaus (RSEB), National Examination Assessment and Evaluation Agency (NEAEA), USAID, and 2-3 local curriculum and assessment specialists per language. The MoE officials and language experts at this workshop decided not to engage with the completely new tool (Tool A) but to focus solely on the revised version (Tool B). Language experts and MoE reviewed each item of the EGRA (Tool B) until consensus was reached for each language's tool. The participants in the workshop compared items with the 2014 baseline and checked for compatibility with the newly developed mother tongue curriculum.

The participants in the workshop took into consideration that alpha syllabaries, such as Amharic and Tigrigna, are written with symbols called fidels, which are represented as syllables (consonant and vowel), and not at the phoneme level 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 EGRA 2015 measures phonemic awareness, syllabic awareness, letter sound fluency, word-naming fluency, unfamiliar word naming fluency, oral reading fluency, reading comprehension, and listening comprehension.

During the revision process, READ M&E and the workshop participants systematically reviewed the level of difficulty of each letter in both tools, position and distribution of the letters and words within the test in both tools, the nature of comprehension passage in terms of number of words and structure between both tools.

The participant of the workshop altered each tool (language) by a different amount. There is structural variation among the languages. Some of the languages are widely used and have rich

experience serving as medium of instruction while others do not. There is also variation in experience among the language experts. Hence, uniform type and magnitude of alteration on the tools was not expected. While one language considered big changes on one of the sub tasks, the other did not do any changes. This was left for the discretion of the Mother Tongue experts and the position of MOE and RSEBs. On the other hand, almost all of the toll developers were also involved during the development of the baseline tool. Hence, the workshop altered each tool by different type and amount. The table below describes the observed change from 2014 to 2016 EGRA instruments:

Table 2: Change per language per item from 2014 to 2016 EGRA

Sub task	Amharic	Afan-Oromo	Tigrigna	Aff-Somali	Sidamu-Afoo	Hadiyissa	Wolayttatto
Letter Identification	31%	10%	43%	0%	16%	4%	15%
Initial/End letter-Sound identification	30%	80%	30%	0%	100%	40%	50%
Familiar words reading	30%	54%	10%	40%	80%	22%	20%
Invented words reading	22%	60%	10%	26%	98%	30%	26%
Oral reading	0%	0%	0%	100%	100%	100%	100%
Listening comprehension	0%	0%	0%	50%	50%	100%	100%
Average Change	**18.83%**	**34%**	**15.50%**	**36%**	**74%**	**49.33%**	**51.83%**

2.2.1 Equating the 2014 and 2016 EGRA tools

To establish the comparability between the 2014 and 2016 Oral Reading Fluency (ORF) subtasks, READ M&E conducted a small research project within the 2016 data collection. The study employed a common-persons research design, which is a recommended research paradigm for equating between EGRA forms. In this design, the same pupils sit for more than one version of the assessment forms. The rationale of common-persons design lays in that the pupils who sit for two sets of data collected by two different forms of the instrument have the same underlying distribution of ability. Thus, any difference between the results collected from these two instruments can be attributed to the instrument characteristics rather than to student characteristics.

By using the common persons design, READ M&E not only controlled for differences in content on ORF but also verified the comparability of two different administration modes- paper (2014) and tangerine tablets (2016). After comparing the common-persons results from two instruments, analysis revealed no practical significant differences between the 2014 and 2016 EGRA ORF subtasks. Thus, the 2016 EGRA result required no statistical adjustment in ORF. The common-person results rule out "instrumentation" as one of the major threats to internal validity of program evaluation studies.

To compare the 2014 EGRA and the 2016 ORF subtasks through the common-person design method, READ M&E sampled five schools/200 pupils in each language (total 1,400 pupils). Twenty pupils of Grade 2 and another twenty of Grade 3 students represented by equal number of boys and girls were selected from the total grade 2 and grade 3 students in attendance. The selected children were given both the 2014 EGRA in its original hard copy version and the 2016 EGRA on the tablet. To control for a possible effect of practicing, the two instruments were administered using the counterbalanced order. One group (A – a random selection of 50% of students from the study sample) had the 2014 EGRA first and the 2016 EGRA second. The other group (B – the second 50% of students from the study sample) had the 2016 EGRA first and the 2014 EGRA second. This method eliminated any bias from practicing and the child being familiar with the test. It is also important that the same assessor administer both forms to the same students to rule out possible influence of a person who is administering the test.

Table 3: Common-Person Equating Administration method

Group	1st administration	2nd administration
A (subsample of 50% randomly selected students)	2014 EGRA pencil and paper	2016 EGRA tablet
B (subsample of 50% randomly selected students)	2016 EGRA tablet	2014 EGRA pencil and paper

Despite concerns that the administration of two EGRAs back to back might be require too much attention for a young child, data collectors and observers were relieved to find that the children remained interested and engaged during the approximately twenty-thirty minute assessment process.

2.2.2 EGRA pilot of 2016 tool

The EGRA tool development has been described prior in the 1.2.2 section of this report. Post development, READ M&E conducted a pilot EGRA to test the revised tools. Piloting the tools, allowed READ M&E to assess item parameters for difficulty and discrimination. It also allowed for the correction of any typos,

grammar errors, and other small items that may be overlooked outside of the actual assessment activity. This improves sub-task accuracy.

READ M&E conducted the pilot EGRA in seven languages and a Mini EGRA of three language in five regions of Ethiopia, Tigray, Amhara, Oromia, Somali and SNNP, from 21 November to 01 December 2015. The test administration required fifty-two test assessors who were deployed in thirteen teams of four people to collect data in sixty-five sample schools in five regions in seven languages (15 schools each from Tigrigna, Af- Somali and Sidamu-Afoo languages and 5 schools each from Amharic, Afan Oromo, Hadiyissa and Wolaytatto languages. The EGRA pilot was conducted on Nexus 7 tablets loaded with Tangerine software.

The READ M&E team, with the IT subcontractor, conducted a five-day training workshop from November 16-20, 2015. The purpose of the workshop was to equip the assessors on techniques on how to administer EGRA 2016 orally on a one-on-one basis using tablets. The first day and a half was devoted to understanding the EGRA tasks. For most of the participants, the first sessions were mostly a review as they have previously administered the EGRA on paper and pencil versions. The remaining days were devoted to practicing how to administer the EGRA on the Nexus 7 2013 tablet with Tangerine software. The final two days, assessors visited schools to practice administering the EGRA to students in the classroom thus providing them with an authentic experience prior to entering the field. Participants reported that visits to the schools were extremely useful.

As this was the first time for EGRA to be administered using tablets in Ethiopia, READ M&E surveyed the participants to see how many were familiar with tablet usage. While all of the participants were familiar with either desktop or laptop computers, seven participants were not familiar with smart phones and only three participants had used a tablet prior to the training. At the end of the workshop, the majority of the assessors reported feeling very comfortable with the new technology. Only few participants were dropped from participating in the data collection because of their self-perceived lack of comfort with the tablet.

READ M&E selected assessors for the pilot from Colleges of Teacher Education (CTEs), Universities, the MoE, Regional State Education Bureaus (RSEBs), National Education Assessment and Examinations Agency (NEAEA), Zone Education Departments (ZEDs) and preparatory schools. The minimum qualification was a M.Ed. or MA.

The pilot prepared the READ M&E team to conduct the larger sample size. Lesson learned included being more specific as to the required quality of vehicles proved, the need for one person to be responsible for

'replacement schools', preparing EGRA assessors for 'low morale' on days when most children score poorly and other valuable preparation.

2.3 Using Nexus 7 tablets and Tangerine Software

READ M&E used Nexus 7 Tablets loaded with Tangerine software to complete the EGRA data collection. Tablets were programmed with the seven EGRAs in November-December 2015 and the programming was modified after the pilot. The modifications included fixing typos, glitches in counting protocols, and issues with timing. These issues were satisfactorily resolved prior to the EGRA training in May.

During the training, some potential assessors struggled with the new technology in the beginning but by the end of the week, through constant practice and mentoring, all but a few participants had mastered the use of the tablets. READ M&E assured that only the best-qualified assessors were deployed by evaluating potential assessors in a series of exercises. For more details on the training process, please see section 3.1 Ensuring high quality data collectors.

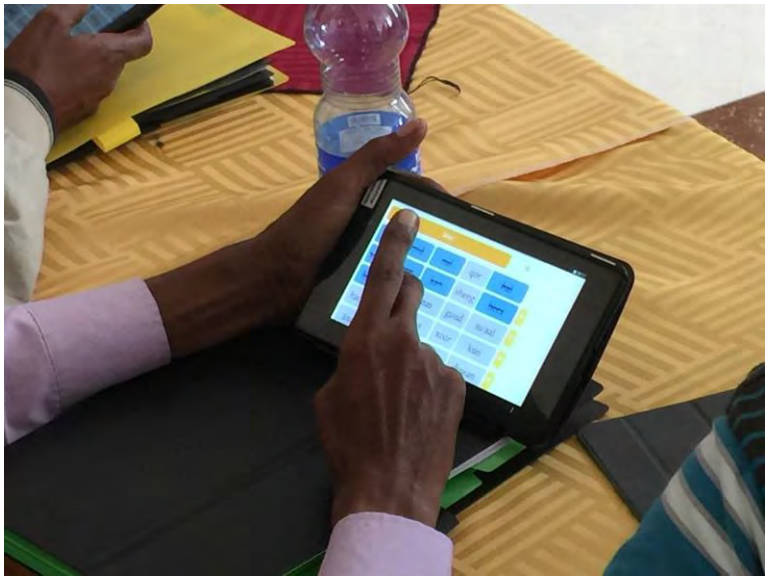

Figure 1: EGRA tablet

Assessors familiar with both the tablets and paper EGRA overwhelming prefer the tablets. During data collection, only one tablet failed. Fortunately, READ M&E was able to replace it the following day. Team leaders and assessors used hotel Wi-Fi, when available to upload data. More often, they used their own cell phones as Wi-Fi hot spots uploading data on a daily basis.

2.4 EGRA sample size

Using calculations from the 2014 baseline, READ M&E found that 300 schools was the sample size needed to enable the desired power of statistical analysis. This allowed for a representative sample size by language and gender. This comes out to roughly 42 schools per language (total schools 294) or 12,000 students. In order to conduct the common-person design study to establish the comparability between the baseline ORF tool (EGRA 2014) and the mid-term ORF tool (EGRA 2016), we assessed an additional 1,400 students from 35 schools or 5 additional schools per language. The common-person design process for establishing comparability is further explained in section 2.2.1.

Assessing 294 schools (42 per language) required for representative sampling and 35 schools (5 per language) required for comparability study comes to 329 schools. READ M&E added three schools per language (21 schools) to bring the total up to 350 schools. Therefore the final sample size was 45 schools per languages (total 315) for the EGRA 2016 and 5 schools per language for the purpose of establishing comparability (35) bringing the total to 50 schools per language or 350 schools total.

READ M&E sampled 350 schools from all five regions and seven languages (fifty schools in each language). Assessors tested forty students per school; twenty pupils of Grade 2 and another twenty of Grade 3 students. Assessors randomly selected an equal number of girls and boys from grade two and grade three students in attendance on the day of data collection. READ M&E's goal was to assess 14,000 children.

READ M&E was able to exceed the required number of students for desired power of statistical analysis (12,000 students) by 1,475 students. The table below illustrates the desired number (target #) vs the number of students assessed by grade and gender:

Table 4: Target number vs actual number

Language	Total # of Schools	Children assessed in Grade 2		Children assessed in Grade 3		Total	% of target # of children assessed (2000 per language)	% of children assessed for statistical representation (1,680 per language)
		Male	Female	Male	Female			
Wolaytatto	50	511	489	504	496	2000	100%	119%
Hadiyissa	50	504	495	495	498	1992	99.6%	118%
Sidama Affo	50	500	500	500	500	2000	100%	119%
Afaan Oromo	50	501	497	492	503	1993	99.65%	118.6%
Af-Somali	50	490	303	444	273	1510	75.5%	90%
Tigrigna	50	497	499	500	500	1996	99.8%	118.8%
Amharic	50	486	511	491	496	1984	99.2%	118%
Total	**350**	**3489**	**3294**	**3426**	**3266**	**13475**	**96.25%**	

The Somali region has been particularly hard hit by the El Nino effect and the subsequent drought and drought related migration. Assessors were able to reach fifty schools but unfortunately, attendance in the schools was too low to meet the targets. This does not affect the results of the 2016 EGRA however, as READ M&E was able to reach 90% of the needed population to make the sample significant.

2.5 Zone, schools, and student selection process

Zones and Schools: READ M&E selected zones according to their accessibility, security level, and status of El Nino effect. A complete random sampling procedure for EGRA 2016 was not done with issues of security and safety along with the impact on education of the drought in the forefront. READ M&E avoided areas that are listed as priority one zones by the Emergency Education Cluster report. UNICEF (April 2016 Fast Fact) estimates that 2.1 million school aged children are unable to access quality education at present. Many schools have closed and those that are kept open have minimal time on learning. In some areas girls are not attending schools as they are tasked with walking far distances to fetch water. READ M&E has consulted with RSEBs and USAID officials and all have agreed that the EGRA testing should not be done in areas affected by the drought.

Over the last six months, the tensions in the Oromia and North Gondor in Amhara region have increased. Schools in some areas have been closed for a good part of the last semester. Furthering the tension is an increase in internal migration. UNICEF (April 2016 Fast Facts) estimates that 105,300 people will be

internally displaced leading to conflict among ethnic groups over shelter, food, and water. READ M&E selected schools from regions where safety is ensured and from those zones that are not severely affected by the drought. Unfortunately, despite best efforts, areas in Amhara and Oromo required a high number of replacement schools.

Fifty schools were selected randomly based on the latest available school list by language. Replacement schools were selected at the same time. Replacement schools were used in case of inaccessibility and difficult weather (Rain and Flooding). Inaccessibility was defined as schools needing a walk of over an hour or those that required boats or motorcycles. The sample schools were confirmed for availability by the RSEBs. A chart of zones from which we have drawn a random sample of schools is attached (Annex).

During data collection, approximately 19% of schools were replaced. When data collectors found that the selected school was not available, they called Deputy Chief of Party, Dr. Solomon, and received the name of a different school that had been previously selected according to requirements. The Oromia region required the most replacement schools. Primary reasons for replacement included distance to walk to the school, school closed, and flooding.

Table 5: Replacement schools used by language

Region	Language	# of Replaced schools	Percentage
SNNP	Wolaytatto	0	0%
	Sidama Afo	7	14%
	Hadiyissa	1	2%
Amhara	Amharic	17	34%
Tigray	Tigirigna	8	16%
Somali	Af Somali	11	22%
Oromia	Afan Oromo	22	44%
		66	18.8%

Student selection: From grade 2 and 3 students from each grade were selected randomly in a random lottery method. The assessors systematically selected students from each selected sample school. In each of the sampled schools, 20 pupils of Grade 2 and another 20 of Grade 3 were selected to participate in the assessment. Where there were 20 or fewer children in a given class because of the aforementioned reasons (drought, bad weathered, waterborne disease) all children in that class were assessed. In each of these classes, an equal number of girls and boys were 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 deaf or blind pupils. Replacement of such children was done after sample selection and was done by the assessor in consultation with the supervisor (not the teacher).

There had been some concern about teachers and or principals swapping out lower performing students for higher performers but this did not happen according to the assessors. Concern about student leaving the testing site or misbehaving did not prove 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.'

3 Implementation

Implementation of the EGRA in seven languages in five diverse regions was a logistical challenge. However, the staff of READ M&E was well experienced with the necessary procedures and arrangements. In this section, some of the more interesting details of the procedures are explained. Particularly interesting, is a narrative describing a "Typical day of EGRA data collection." This section highlights from a minute-by-minute perspective exactly what happens when a team goes to conduct an EGRA at a school. The section ends with a frank discussion of the administrative challenges READ M&E encountered.

3.1 Ensuring high quality assessors

READ M&E held the training of assessors from May 16-22 in Bishoftu, Ethiopia. The five-day training had one hundred and forty needed assessors, twenty-five replacement assessors along with eight READ M&E staff, two additional EGRA trainers, and USAID representatives. READ M&E held the training in three sections, by language group, of approximately fifty to fifty-five participants each group.

Table 6: EGRA trainees and their organization

Trainee's Organization	Number of Trainees
CTE	11
Private consultants	22
MoE	43
NEAEA	2
Other NGO	1
RSEB	60
ZED	7
Total	**146**

One hundred and sixty-five participants were invited to the training however there were only 140 needed positions. This over-selection was intentional in order to account for poor performing assessors or the voluntary withdrawal of others. A recommended percentage of over-inviting is contextual but 15-20% seems to be a minimum (if not drawing heavily on experienced administrator cadres).

By the end of the training assessors were able to:

- Understand the purpose of EGRA and acquire knowledge of content (EGRA subtasks).
- Attain skill with tablet and paper based means of data collection.
- Acquire knowledge of how to save and upload, or deliver data.

- Acquire knowledge of how to select the students at schools.
- Demonstrate ability to administer EGRA (by end of training, there will be evidence of performance).
- Understand the communications, trouble shooting, and emergency protocols, when to connect with EGRA hub (READ M&E office) for assistance.
- Understand how to resolve key logistic and administrative issues (travel, money, appropriate vehicle use, etc.).

Most session designs employed interactive activities and tasks. For example, participants first worked in pairs to practice EGRA administration; then they observed their peers in larger group "fish bowl" type activities where they observed and evaluated live demonstrations; then they evaluated pre-recorded videos, etc. Debriefs were held in large and small groups. Such practice activities were conducted for both hard copy and Tangerine tablet administration.

Respondents on the pre-training survey were asked about their level of comfort with technology. As shown in the chart below-approximately 47% checked that they are 'easily frustrated by technology' or 'would rather use paper and pencil.' On the opposite side, approximately 56% expressed ease and comfort with technology.

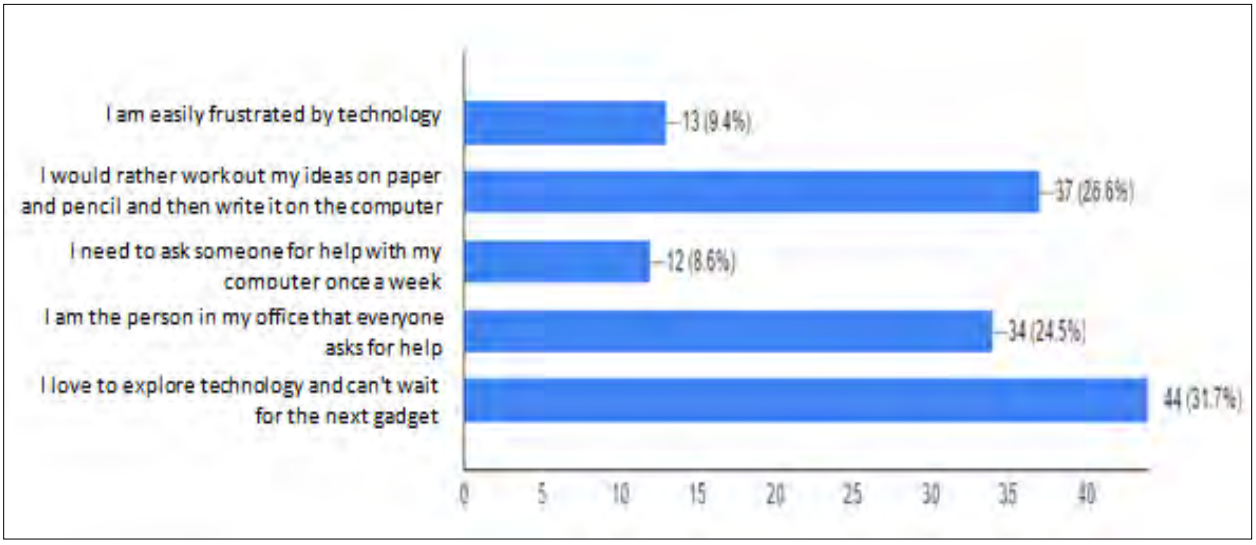

Figure 2: Comfort with technology

Many of the participant cohort had never held a handheld electronic device in their hands. They would make basic mistakes like placing their palm on buttons while trying to mark other sections with their

fingers. Considerable attention was needed to get everyone up to speed with the tablets. Trainers recommend that future trainers not take skill with touch screen devices for granted and to make sure adequate time was planned for practicing “basic functions” like turning on and off.

While there was variation in the proportion of experienced participants by language group, the trainers were keenly aware of the value of peer learning and did their best to encourage broad ownership of the learning. Veteran trainers made strategic use of the EGRA experience they had in the room by intentionally spreading the veterans around the groups and empowering them to serve as “coaches and mentors” for newcomers.

To select the needed 140 assessors out of the 165 present at the training, READ M&E used empirical evaluations to determine suitability of administrator candidates. There was a significant time investment made to develop measurement tools to assess candidates and provide instant feedback: An important component to ensure quality administration. Potential assessors were given four EGRA sample assessments, one on paper and three via the tablets. The assessments were scored against prepared answer keys. The assessors either watched a video of an assessment session with one person playing the assessor and the other playing the child or watched a live session. The videos were prepared in advance by the READ M&E team in all seven languages. The ‘child’ in the video or live assessment responded to the assessment according to preplanned errors. The assessors watching the assessment then scored their own sheets according to the errors they observed.

READ M&E team then calculated and reported to the participants the percent correct score as measured against an EGRA answer key (what was referred to as “the gold standard”). Each individual’s score was marked and tabulated, and the counts for participants falling into certain ranges were presented. The output presented on posters was, for example, the number of raters who scored between 60-70% correct (e.g. n = 5), number of raters who scored between 70-80% correct (e.g. n = 13), etc. for each language group (see poster below). Because READ M&E did not select the EGRA trainees personally but rather largely depended on the RSEBs to send the most qualified candidates, there was some disagreement among participants about using the sample EGRA assessments as criteria for hiring assessors. Some participants argued that due to their status in the office or their seniority, they should be given automatic preference for the position of assessors. Dr. Solomon, Deputy COP for READ M&E handled the difficult task of dismissing twenty-five participants gracefully and with good will.

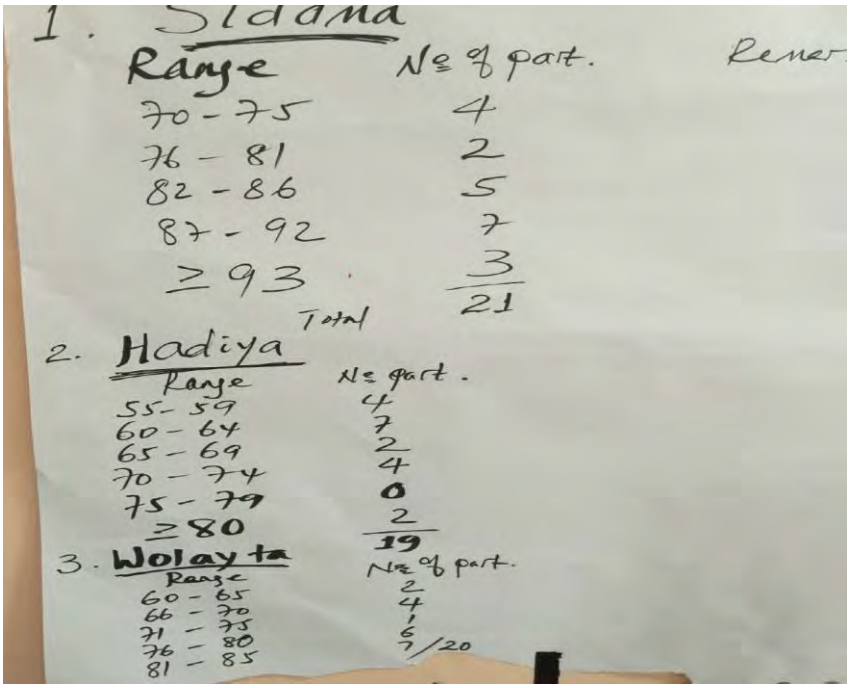

Figure 3: Scoring chart of EGRA assessors during training

The post-survey revealed that all in all the participants believed that they were well prepared and ready to administer the EGRA. Other details from the post-training survey include:

- 64% of the participant strongly agreed and the rest agreed on the well preparedness of the trainers. In the comments, they said that the trainers were well prepared and know very well about the content and language of the training. Most of the participants said that the skill, the preparation about the content of the presentation/topics, the approach, and the support provided from the trainers was encouraging.

- 82.6% and 75% of participants strongly agreed and 16.7% and 24.3% of the participants agreed on the training provided on EGRA has sufficiently prepared them to use and administer the EGRA using the tablets with the Tangerine software for data collection, respectively. Only 0.7% remains neutral on both cases.

- 64.3% of the participants strongly agreed and while 29.4% agreed on the time allotted for the training was sufficient.

3.2 Typical day of EGRA data collection

A typical day of EGRA data collection begins with the assessors leaving their hotel very early in the morning. Having arranged with the principal and woreda officials earlier in the week, they may stop to pick up a woreda official to guide them to a remote school. Sometimes the walk to the school is difficult- 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 EGRA. Usually the team arrives prior to 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 moves them to separate lines. The team writes down the children's names and gives each child a unique number. The selected children move to a comfortable visible area to wait their turn.

Figure 4: Children in Bahar Dar lining up for selection and receiving numbers

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 as the children are happy to have this free time to play with their mates.

Figure 5: EGRA assessor and child Hawasa area

The team leader sets up the other three assessors with their first children. When called, the child repeats their name and the assessor verifies the name and number on the master list. After the EGRA is finished, the child receives a pencil and runs back to their 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 coordinates, assists in identifying 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.

3.3 Supervision of EGRA data collection

READ M&E staff members and one additional consultant, Ato Bekalu, visited as many teams as possible. Supervision ensured that the EGRA proceeded as planned and problems with the tablets, if any, were resolved quickly and efficiently. Addis Yigzaw and Marc Bonnenfant from USAID joined the supervision team for Somali and Dr. Todd Drummond joined the team for Sidamu Affo near Hawasa.

Table 7: EGRA supervision staff, zone, and duration

SN	Name	Region/zone	# of days	Duration
1	Dr. Jordene Hale and Dr. Todd Drummond	Sidama zone	3	May 22-24/16
2	Dr. Jordene Hale	Amhara/ BahirDar zuria	4	June 1-5/16
2	Ato Daniel Tefera	Sidama zone	9	May 22-29/16
3	Ato Zewdu G/kidan	Somali and Oromia	10	May 30- June 7/16
4	Ato Endalamaw Teka	Oromia	9	May 21-29/16
5	W/ro Selam Wudu	Tigray	9	May 21-29/16
6	Ato Lishan Kassa	Hadiya and Wolayta	14	May 22- June 4/16
7	Ato Bekalu Yahey* (consultant)	Amhara	9	May 22-30

Ato Zewdu remained in the Addis Office the first week of data collection to be able to monitor data as it came in to the tangerine cloud based control. Dr. Solomon remained in the Addis office the entire data collection period taking calls from assessors in the field. This role was vital as trouble with the tablets (once), car accidents (two) and schools replacements occurred.

3.4 Administrative Challenges

Although the EGRA assessment went well and no significant issues disturbed the success of the data collection, there was major challenges none-the-less. READ M&E deployed 140 data collectors and our staff to supervise the data collection. The mobilization required 41 vehicles to be deployed simultaneously. The RFQ bidding process was won by a consolidator who subcontracted to individual owner/operators. As such, not all of the vehicles were of the best condition. Fortunately, only two severe accidents occurred and no one was severely hurt. In each case, the assessors chose to continue working and reached their destinations via public transportation. The consolidator was prompt in sending replacement vehicles and the teams were able to continue on schedule. One of the drivers was

hospitalized with malaria and unable to complete the route. He allowed another driver to take over his route without incidence.

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. Fortunately, this did not happen. One tablet did malfunction but it was on the first day of the data collection and READ M&E were able to send a replacement tablet immediately. Data collectors quickly learned to upload data daily using their personal cell phones as internet 'hot spots' and all data were successfully uploaded.

Other challenges mentioned above include the need for 'replacement schools' (section 2.5), dismissing assessors who did not make the cut (section 3.1), and the difficulty of low attendance in the Somali region (section 2.5).

4 Results from Surveys

This section presents the findings from the three questionnaires administered with the midterm EGRA:

- School director questionnaire:
  - characteristics of the school director,
  - characteristics of the schools
- Teacher questionnaire:
  - characteristics of teachers and their classroom,
  - teaching methodologies.
- Student questionnaire:
  - family background of students,
  - reading practices at home
  - availability of support for reading.

In the following sections, we present descriptive statistics of the data collected on several of these surveys. At present, AIR researchers are working on further analyses to determine what factors may be associated with achievement outcomes.

4.1 Director's Questionnaire

Three hundred and fifty school directors responded to the questionnaire answering about themselves and their school. Questions focus on activities related to mother tongue education. The survey questions are in the annex.

Slightly over half of the directors have bachelor's degree (53.7%). The remainder has only a high school diploma (43.7%). In the past year, about twenty-two percent received training on implementing a program on reading leaving 78% without training in reading. However, ninety percent said they support teachers on how to teach reading. Sixty-three percent (63.4%) said they are satisfied with the performance of Grade 2 and 3 students in reading.

Thirty-two percent of the directors said that they are responsible for reviewing the lesson plan and most review the lesson plans every week. 44.3% said they conduct classroom observations while 40.3% said the deputy director carries it out.

When asked how they would know if students are progressing in reading, 78.3% said by classroom observation, 62% said teachers provide them with progress reports, and 57.4% said they evaluate children orally.

Table 8: How do you know if your students are progressing?

Classroom Observation	78.3%
Teachers provide me progress reports	62.0%
Evaluate children myself	57.4%
Monitor student's results on tests given by teachers	50.9%
Review children's assignments or homework	30.6%

4.2 Teacher's Questionnaire

Five hundred eighty-four teachers, with about equal number of female and male, responded to questionnaires. The highest level of education for 81% is a diploma and on average, they have five years

teaching experience with a range of 1 to 35 years. Seventy-five percent (75.3%) attended training related to the newly developed mother tongue and the training lasted on average 10 days.

More than half (68.5%) said their school has a functioning library or reading room and 77.9% said they supervise their students while using the library. Less than half (44.6%) said they have sufficient reading materials to support teaching reading.

Almost all of them (96.7%) have the newly developed mother tongue textbook and 52.9% said they use it on daily basis. Seventy-three percent (73.1%) said it is very useful. Eighty-three percent (83.4%) have the corresponding newly developed mother tongue teacher's guide. Seventy-seven percent (77.6%) said it is very useful.

In addition to the background information, teachers were asked about how frequent they perform different activities with their students during mother tongue class. Their responses were based on the past five school days. Copying text from the board was done more often than the other activities while sounding out unfamiliar words was the least frequent.

Table 9: Teaching activities by frequency

No.	Activity	Never	1 day a week	2 days a week	3 days a week	4 days a week	5 days a week
1	The whole class repeated sentences that you said first.	3.1%	10.1%	27.1%	28.8%	10.8%	20.2%
2	Students copied down text from the chalkboard.	0.9%	5.0%	18.7%	24.7%	17.3%	33.6%
3	Students retold a story that they read.	2.1%	18.3%	36.3%	24.7%	5.5%	13.2%
4	Students sounded out unfamiliar words.	10.8%	17.5%	30.0%	24.0%	7.5%	10.3%
5	Students learned meanings of new words.	1.0%	12.5%	32.2%	25.7%	9.4%	19.2%
6	Students read aloud to teacher or to other pupils.	0.9%	14.7%	29.6%	21.6%	11.1%	22.1%
7	Students were assigned reading to do on their own during school time.	2.6%	16.1%	31.2%	23.6%	8.7%	17.8%

Teachers were asked about the different methods used to monitor student's reading progress. Oral evaluation was found to be used most frequently, followed by review of students' work and checking exercise books.

Table 10: Methods of assessment

No		Never	1 day a week	2 days a week	3 days a week	4 days a week	5 days a week
1	Written evaluations	0.5%	43.8%	26.7%	16.4%	4.8%	7.7%
2	Oral evaluations	3.3%	15.6%	20.4%	14.0%	9.9%	36.8%
3	Review of pupil work	1.2%	21.4%	20.9%	16.6%	8.4%	31.5%
4	Checking of exercise books	0.5%	13.7%	22.3%	21.2%	10.8%	31.5%
5	Checking of homework	1.0%	11.1%	25.9%	29.8%	11.5%	20.7%

Teachers expectation of students' performance at different grade levels was asked by series of questions listed in Table 11. Most teachers expect their students to understand stories they read at Grade 2 (48.3%) and Grade 3 (40.8%). About 48% also expect students to read aloud a short passage with few mistakes at Grade 2. For most teachers Grade 1 students should recognize letter names (57.4%), write their names (50.2%) and recite alphabet (50.2).

Table 11: Grade level expectations of reading skills

In what grade level should pupils FIRST be able to demonstrate each of the following reading skills?						
No.	**Skill**	**Before G 1**	**G 1**	**G 2**	**G 3**	**Not important**
1	Read aloud a short passage with few mistakes	1.0%	24.5%	48.1%	24.8%	1.5%
2	Write name	3.3%	54.5%	32.4%	9.2%	0.7%
3	Understand stories they read	0.5%	9.2%	48.3%	40.8%	1.2%
4	Recognize letters and say letter names	5.8%	57.4%	30.3%	6.5%	0.0%
5	Sound out unfamiliar words	0.7%	17.8%	53.6%	26.0%	1.9%
6	Understand stories they hear	0.9%	20.2%	46.7%	31.0%	1.2%
7	Recite alphabet	2.4%	50.2%	37.8%	8.7%	0.9%

4.3 Student's Questionnaire

READ M&E asked 12,142 students, about 51.2% boys and 48.8% girls, from Grade 2 and 3 about themselves and issues related to mother tongue education. Approximately 95% use the same language at

home and school. Over half of the students (52.1%) attended preprimary education. In Grade 2, 4.7% repeated the year. For Grade 3, repeaters were less at 2.6%. Slightly over a quarter of the students (25.1%) were absent from class for more than five days in the Academic Year.

In relation to the availability of materials 73.4% said they have the mother tongue text, 52.6% said they have additional books and 43.4% said they read additional books. In regard to parental education: 70.3% said their father and 46.6% said their mother could read and write. 68.8% said there is at least one person at home who helps them in their study.

Table 12: Student survey responses

	Yes
Do you have the mother tongue textbook?	73.4%
Do you have books or reading materials at home other than the mother tongue textbook?	52.6%
Do you read books or reading materials other than your mother tongue textbook?	43.4%
Can your mother read and write?	46.6%
Can your father read and write?	70.3%
Is there any one at home who helps you in your studies?	68.8%

5 Results: 2016 data analysis

This section provides insights into major results of EGRA 2016 midterm administration. When working with seven languages, each must receive its due attention. As best as possible, READ M&E has attempted to delineate each type of data analysis and to provide a brief introduction on how to read the data. To keep the report clean and clear, all analytic methods and detailed statistics are provided in the annexes.

5.1 Proportions Meeting Benchmarks

READ M&E uses the benchmarks validated in the 2015 January workshop held by USAID and the MoE, facilitated by RTI. Based on this, the three different levels of reading ability with the proposed oral reading fluency (ORF) benchmarks that would define each level are:

- **Reading fluently with full comprehension(RWFF)** – students achieving the level of reading fluency that the data indicate corresponds with full or almost full comprehension;
- **Reading with increasing fluency and comprehension (RWICF)**– students who have some reading fluency, but have not yet reached the above mentioned levels of fluency and comprehension;
- **Reading slowly and with limited comprehension** (**RWLC**) – students scoring above zero, but at the lower end of the reading fluency score distribution.

Table 13: Proposed correct words per minute by language and grade

Language	Grade	Students who are reading:		
		With limited fluency and comprehension	With increasing fluency and comprehension	Fluently and with full comprehension
Afan Oromo	Grade 2	1 to 19	20 to 47	≥ 48
	Grade 3	1 to 29	30 to 57	≥ 58
Af-Somali	Grade 2	1 to 24	25 to 49	≥ 50
	Grade 3	1 to 24	25 to 54	≥ 55
Amharic	Grade 2	1 to 29	30 to 49	≥ 50
	Grade 3	1 to 34	35 to 59	≥ 60
Hadiyissa	Grade 2	1 to 24	25 to 39	≥ 40
	Grade 3	1 to 24	25 to 49	≥ 50
Sidamu Afoo	Grade 2	1 to 19	20 to 44	≥ 45
	Grade 3	1 to 24	25 to 52	≥ 53
Wolayttatto	Grade 2	1 to 19	20 to 42	≥ 43
	Grade 3	1 to 24	25 to 51	≥ 52
Tigrigna	Grade 2	1 to 19	20 to 54	≥ 55
	Grade 3	1 to 24	25 to 61	≥ 62

Based on the cut scores delineating reading proficiency levels for each language that were established in the above mentioned benchmarking study in 2015, we classified student performance in ORF tasks in 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 old EGRA paper-based form on data collected by Tablets was enabled through the comparability study that was carried out as a part of the 2016 mid-term EGRA data collection. Another important note to keep in mind that all statistical analysis of midterm data is conducted using the enrolment based weights (more details provided in section 7.2).

Percentages of students falling in each reading fluency level (benchmark level) based on ORF scores are computed and presented in this section. Figure 6 shows that in Grade 2 there are large differences between languages in percentages of students achieving reading benchmarks. The highest ORF level 'Reading fluently with full comprehension' was achieved from 0.8% of students in Aff Somali schools to 36.7% of students in Wolayttatto speaking schools. In two languages (Aff Somali and Tigrigna) there are less than 1% of students who achieved that desirable reading fluency level, and in three languages (Afan Oromo, Hadiyissa, and Sidamu Afoo) there are less than 10% of students who can read at this benchmark level.

However, if the level 'Reading with increasing fluency and comprehension' can be also considered as relatively functional reading proficiency level, then percentages of students who achieved this benchmark or above varies from 9.9% in Somali to 63.1% in Wolayttatto speaking schools. Across all languages, in average, about 33% of students can read at functional level (fluently or with increasing fluency).

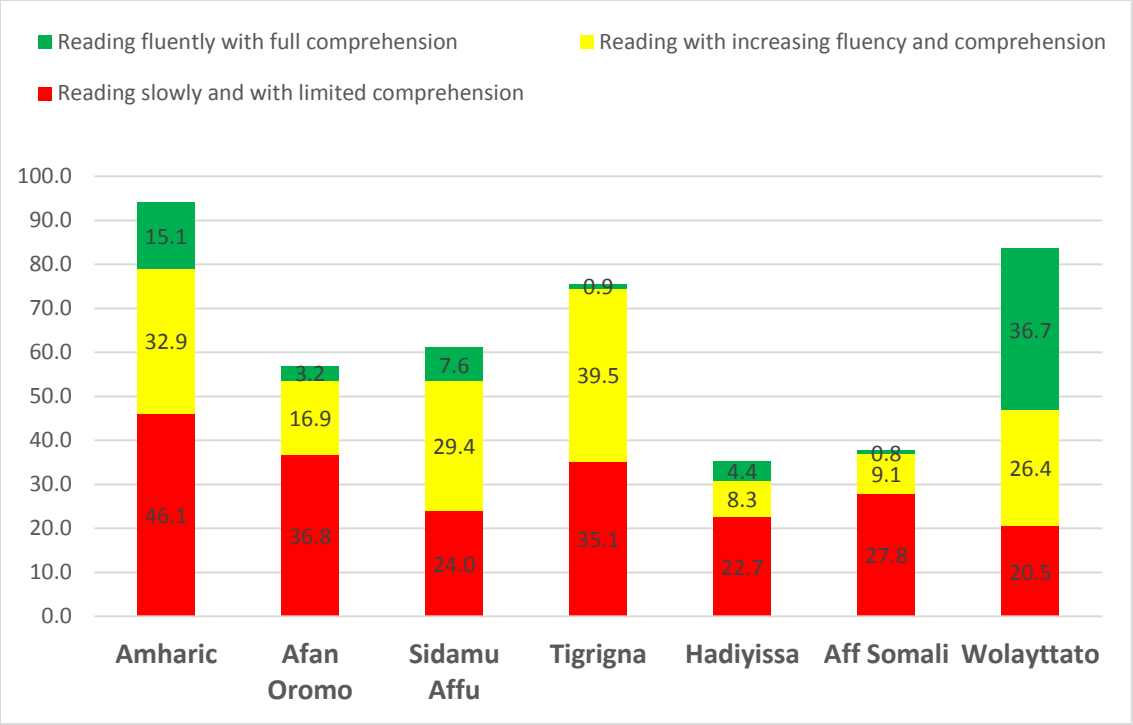

Figure 6: Percentage of students at levels of oral reading fluency in grade 2 by language

In Grade 3, improvements are observed in each language but the pattern remains more or less the same. Figure 7 shows that the percentage of students that achieved the highest level varies from 3.6% in Tigrigna to 36.5% in Wolayttatto speaking schools. In three languages, less than 10% of students achieved that highest benchmark, while in Amharic and Sidamu that percentage is 16.8% and 10.7%, respectively. Considering the two top fluency levels combined, an average of 45% students across all languages are showing that relatively functional reading fluency level. This represents an increase of 12% compared with grade 2, which is pretty encouraging evidence, especially taking into account that benchmarks are designed to be grade level aligned. In other words, even if the percentage of students in benchmark levels would be the same it would still indicate the growth between grades as defined by benchmark setters. From that perspective, the evidence that in grade 3 there are 12% more students than in grade 2that are reaching the benchmark, characterizes this growth as larger than expected by grade level standards. This is an interesting finding suggesting a positive influence of reading instruction, and on the other hand, that the differences between grade level expectations need to be increased.

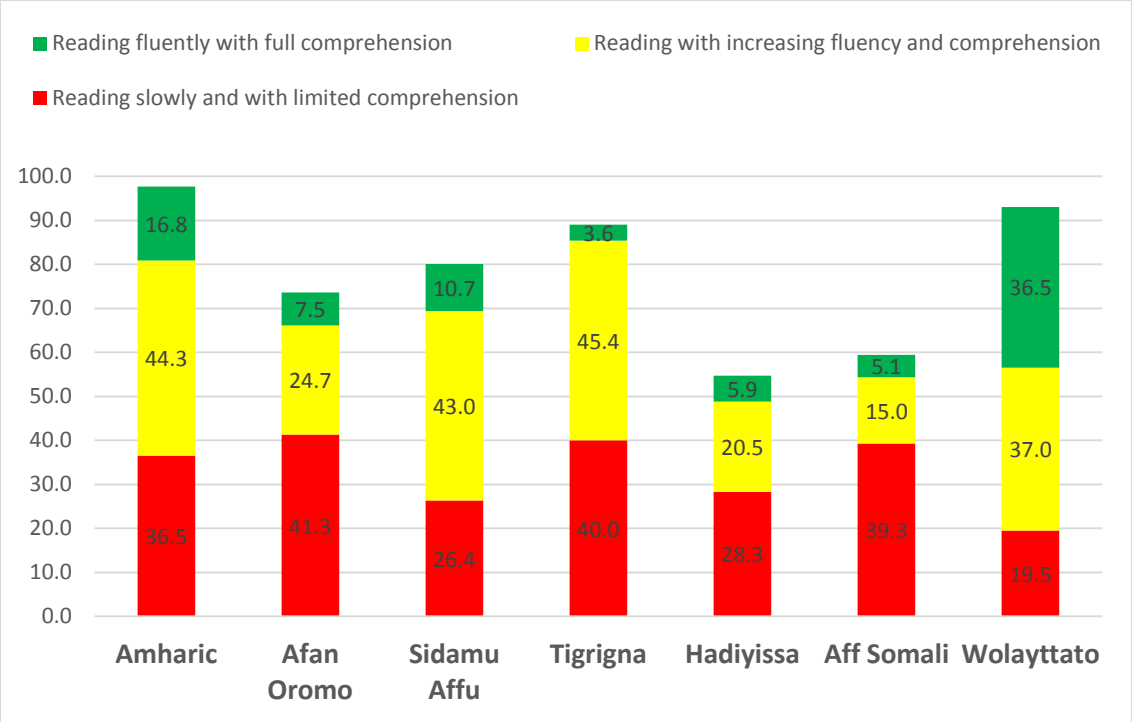

Figure 7: Percentage of students at levels of oral reading fluency in grade 3 by language

To summarize, looking at both grade 2 and 3 combined, 16% percent of Amharic speaking students and 37% of Wolayttatto students achieved the top level of reading proficiency – reading fluently with full comprehension. In other five of the seven languages, less than 10% of the students in assessed grades achieved that desired level of reading proficiency. However, if we consider the top two reading levels combined (reading fluently with full comprehension and reading with increasing fluency and comprehension), it can be observed that about 40% of students in assessed grades and languages are exhibiting relatively satisfactory reading proficiency levels (ranging from 15% in Aff Somali to 68% in Wolayttatto).

5.2 Percentage of Non-Readers

For the timed EGRA components (letter name knowledge, familiar words reading, invented words reading and passage words reading), an auto stop rule was implemented to discontinue the test if students could not correctly respond to a certain number of items within 60 seconds. This rule was established to relieve stress among students and to avoid that students who could not read be tested and frustrated by the task. Students to which auto stop rule was applied receive zero scores and are considered as non-readers.

The passage words reading (oral reading fluency) results reported in the section above are complemented with the information about non-reader rates in this section. Figures 6 and 7 above are based on the students who could read at least one word, thus, the total height of each bar represents the percentage of students who have non-zero scores.

Figure 8below shows the percentages of students that discontinued oral reading fluency task, thus, those that received zero scores and therefore are considered as non-readers. The largest proportions of non-readers are observed in Hadiyissa and Aff Somali languages. It is worrying that the rates of non-readers in those languages are over 60% in grade 2 and still over 40% in grade 3. The lowest rates of non-readers are observed in Amharic (5.9% in grade 2 and 2.4% in grade 3) and Wolayttatto (16.5% in grade 2 and 7.0% in grade 3). In all cases, the percentages of non-readers are larger in Grade 2 indicating that certain improvement of reading skills happens from grade 2 to 3.

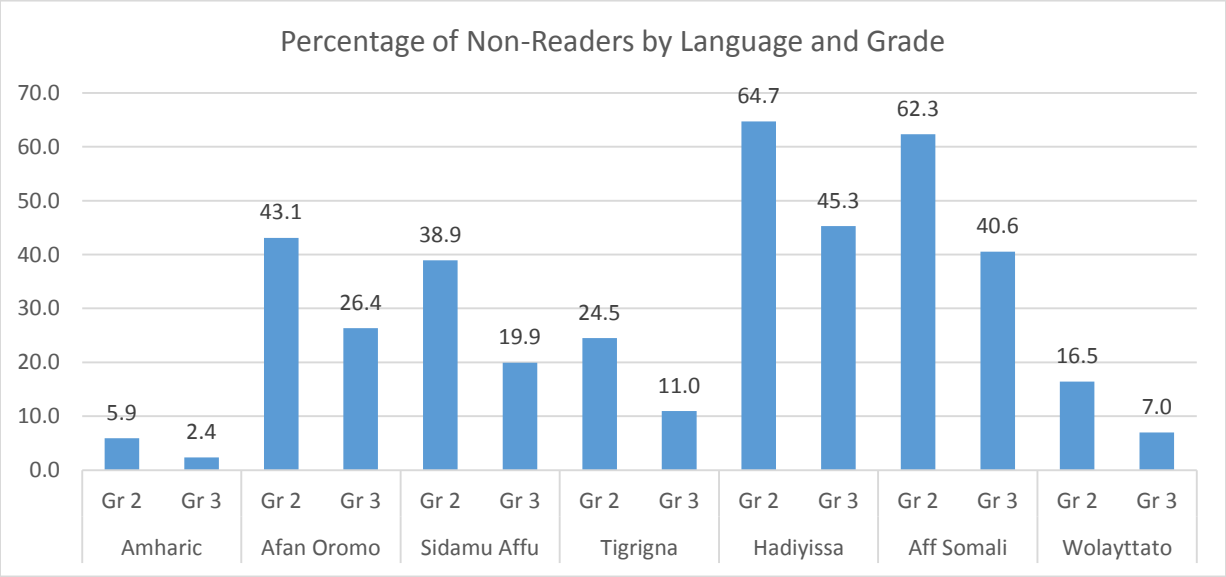

Figure 8: Percentage of Non-Readers in ORF by Language and Grade

5.3 Mean Fluency Scores

This section presents the mean fluency scores of the timed tasks: Letter Naming Correct Per Minute (LSCPM), Familiar Words Correct Per Minute (FWCPM), Invented Words Correct Per Minute (IWCPM) and Oral Reading Correct Per Minute (ORCPM) by grade and gender for each of the seven languages. Note:

Oral Reading Correct Per Minute and Correct Passage Words per minute are the same. In this section, we use the ORCPM abbreviation.

5.3.1 Mean Fluency Scores by Grade and Language

Table 14 and Figure9 show the mean fluency scores of the timed tasks for the seven languages by grade. The results of testing statistical difference between grades (by independent sample T-tests) and practical differences between grades (by Cohen's D) are presented in Appendix 7.3

In each language, the mean differences between the two grade levels in all EGRA sub-tasks 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 rather positive grade gain. The sizes of grade differences for ORF across all languages measured by Cohen's D ranges from 0.40 for Hadiyissa to 0.64 for Aff Somali, all being categorized as educationally significant effect sizes indicating that a substantial reading gains are happening between grades 2 and 3.

Table 14: Mean fluency scores by grade and language

Language	Grade	LSCPM	FWCPM	IWCPM	ORCPM	Cohen's D for ORCPM
Afan Oromo	Two	39.3	10.0	5.0	9.8	0.61**
	Three	53.9	19.0	9.8	21.2	
Aff Somali	Two	23.6	6.4	6.4	6.4	0.64**
	Three	42.4	14.3	14.1	16.5	
Amharic	Two	49.0	30.1	21.4	28.7	0.61**
	Three	60.7	41.0	27.7	40.5	
Haddiysa	Two	34.0	8.8	7.4	7.5	0.40*
	Three	48.9	15.6	12.7	14.4	
Sidamu Affo	Two	54.8	16.2	13.5	16.3	0.53**
	Three	70.0	25.6	21.8	27.1	
Tigrigna	Two	33.9	23.5	14.8	16.4	0.55**
	Three	42.8	37.2	19.4	26.2	
Wolayttotta	Two	63.8	27.7	24.8	30.8	0.41*
	Three	72.3	35.7	33.0	40.8	

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

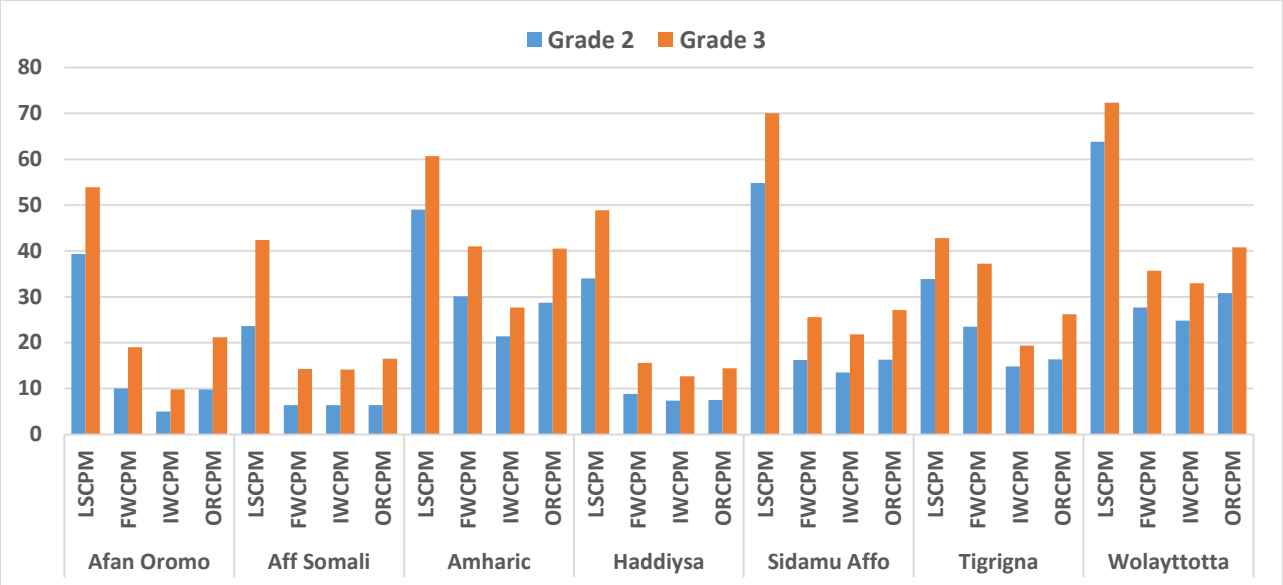

Figure 9: Mean fluency scores by grade and language

5.3.2 Mean Fluency Scores by Gender

Table 15 to 16 and Figure 10below 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-test), as well as measures of the corresponding sizes of differences (by Cohen's D), for all grades and languages are presented in Appendix 7.5.

Looking at the oral reading fluency (passage words) scores in Grade 2 (Table 15),boys performed better than girls in Aff Somali (4.3 words per minute) and Hadiyissa (3.9 wpm), and girls performed substantially better in Sidamu Affo (7.0 wpm) and slightly better in Afan Oromo (2.2 wpm). In other languages differences in ORF were very small or negligible (Cohen's D less than 0.15).

In oral reading fluency (passage reading) in Grade 3 (Table 16), boys performed substantially better than girls only in Hadiyissa (6.6 wpm) and marginally better in Aff Somali (2.7 wpm). However, girls performed substantially better than boys in three languages (Afan Oromo 8.7 wpm, Sidamu Affo 7.5 wpm, and Amharic 5.4 wpm). In the remaining two languages the gender difference were very small or negligible (Cohen's D less than 0.15). Figure 10 shows graphs depicting gender performance in all timed tasks for both grades and all 7 languages.

Table 15: Mean fluency scores by gender and language in grade 2

Language	Gender	LSCPM	FWCPM	IWCPM	ORCPM	Cohen's D for ORCPM
Afan Oromo	Male	35.9	8.7	4.0	8.7	0.15
	Female	42.7	11.2	6.0	10.9	
Aff Somali	Male	26.6	8.0	8.2	8.2	-0.36*
	Female	19.3	4.1	4.0	3.9	
Amharic	Male	48.4	30.6	21.6	29.1	-0.04
	Female	49.7	29.6	21.2	28.3	
Hadiyissa	Male	38.6	11.0	9.5	9.5	-0.26*
	Female	29.3	6.6	5.2	5.6	
Sidamu Affo	Male	49.0	12.9	10.6	12.7	0.36*
	Female	60.5	19.4	16.4	19.7	
Tigrigna	Male	34.6	23.7	15.1	16.4	0.01
	Female	33.1	23.3	14.5	16.5	
Wolayttatto	Male	62.0	26.1	23.0	28.7	0.17
	Female	65.6	29.1	26.5	32.8	

Note: (-) Negative sign of Cohen's D indicates difference in favor of boys
* … indicates that the size of difference is educationally significant (Wolf, 1986)

Table 16: Mean fluency scores by gender and language in grade 3

Language	Gender	LSCPM	FWCPM	IWCPM	ORCPM	Cohen's D for ORCPM
Afan Oromo	Male	49.1	15.3	7.6	16.8	0.40*
	Female	58.5	22.7	12.0	25.5	
Aff Somali	Male	47.2	15.7	15.8	17.7	-0.14
	Female	36.0	12.3	11.8	15.0	
Amharic	Male	58.0	38.9	26.4	37.8	0.26*
	Female	63.3	43.1	29.0	43.2	
Hadiyissa	Male	56.2	19.4	15.6	17.7	-0.35*
	Female	41.7	11.8	9.8	11.1	
Sidamu Affo	Male	65.8	21.8	19.0	23.3	0.37*
	Female	74.0	29.3	24.4	30.8	
Tigrigna	Male	44.5	38.5	19.9	27.0	-0.08
	Female	41.1	35.9	19.0	25.4	
Wolayttatto	Male	72.1	33.5	31.3	39.2	0.13
	Female	72.6	37.8	34.7	42.4	

Note: (-) Negative sign of Cohen's D indicates difference in favor of boys
* … indicates that the size of difference is educationally significant (Wolf, 1986)

Figure 10: Gender Differences in Timed Subtasks by Language and Grade (with Cohen's D)

5.4 Mean Scores of Untimed Tasks by Language and Grade

This section shows the mean scores of the untimed tasks disaggregated by grade for each of the seven languages. The untimed subtasks are Phonemic Awareness (PA), Reading Comprehension (RC) and Listening Comprehension (LC). The 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 7.4.

In all languages, the mean differences between the two grade levels are statistically significant in favor of Grade 3.The strongest grade gains averaged across all languages are observed in reading comprehension (13.4%) followed by Phonemic Awareness (9.1%).In reading comprehension the grade gains vary among languages from 8.6% (Hadiyissa) to 17% (Afan Oromo).The increases of grade means in phonemic awareness range from 1% (Sidamu Afoo) to 23.6% (Af- Somali). The observed grade gains in listening comprehension are relatively small (4.7%).

To summarize, in the context of strong gains in reading comprehension, and gains in oral reading fluency (reported in previous section), indicates that growth in students' comprehension, and ultimately students' learning growth, strongly depends on reading skills. It is also worthwhile to mention that reading gains from grade 2 to 3 were apparent in 2014, but still smaller than in 2016 (the average Cohen's D across languages in 2014 was 0.51 for ORF and 0.43 for RC, whereas in 2016 it was 0.54 for ORF and 0.52 for RC), which suggests that educational effects on reading growth from grade 2 to grade 3 are increasing.

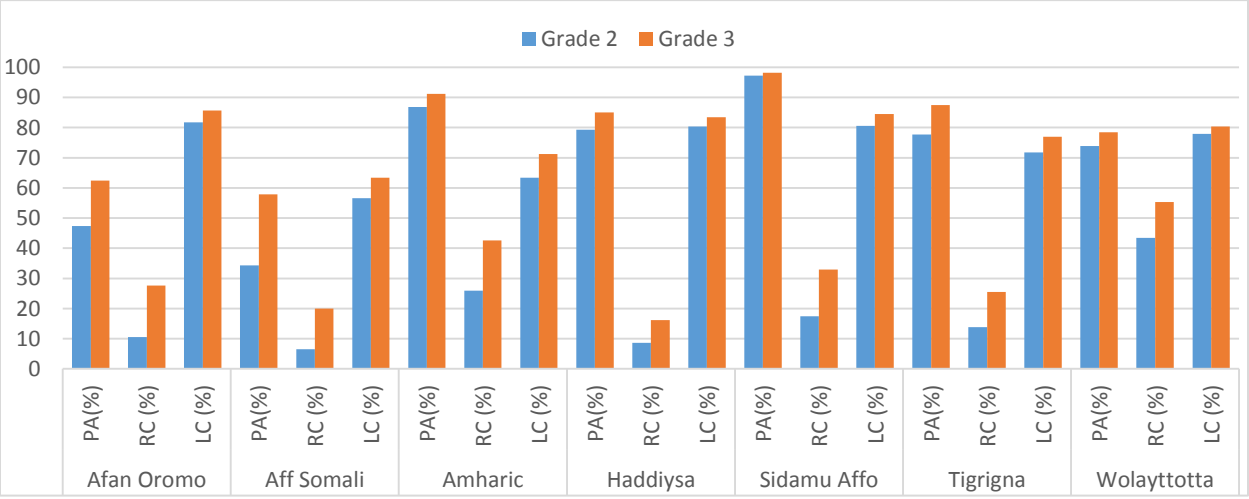

Figure 11: Mean scores of untimed tasks by grade and language

Table 17: Mean scores of untimed tasks by grade and language

Language	Grade	PA(%)	LC (%)	RC (%)	Cohen's D for RC
Afan Oromo	Two	47.3	81.7	10.5	0.65**
	Three	62.4	85.6	27.6	
Aff Somali	Two	34.3	56.6	6.5	0.61**
	Three	57.9	63.4	20.0	
Amharic	Two	86.8	63.4	25.9	0.57**
	Three	91.2	71.2	42.6	
Haddiyssa	Two	79.3	80.3	8.6	0.32*
	Three	85.0	83.4	16.2	
Sidamu Affo	Two	97.2	80.6	17.4	0.59**
	Three	98.2	84.5	32.9	
Tigrigna	Two	77.7	71.7	13.8	0.53**
	Three	87.5	76.9	25.5	
Wolayttatto	Two	73.9	77.9	43.4	0.34*
	Three	78.4	80.3	55.3	

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

5.5 Mean Scores of Untimed Tasks by Language and Gender

This section shows the mean scores of the untimed tasks disaggregated by gender for each of the seven languages. The untimed subtasks are Phonemic Awareness (PA), Reading Comprehension (RC) and Listening Comprehension (LC). The results of statistical significance testing between performance of boys and girls (using T-test) along with corresponding sizes of difference (by Cohen's D) are presented in Appendix 7.6.

Table 18 shows the mean scores of the untimed tasks for boys and girls in Grade 2. Considering reading comprehension (RC) task, girls performed substantially better than boys in Sidamu Affo (for 8.6% points) and Afan Oromo (4.1%). In Aff Somali and Hadiyissa boys performed slightly better than girls (for 2.8% and 3.5%, respectively), but the size of difference (Cohen's D) is below of what is considered practically significant (less than 0.20).

Table 18: Mean scores of untimed tasks by gender and language in grade 2

Language	Gender	PA(%)	LC (%)	RC (%)	Cohen's D for RC
Afan Oromo	Male	43.4	81.6	8.4	0.20
	Female	51.1	81.7	12.5	
Aff Somali	Male	44.3	57.5	7.7	-0.18
	Female	20.3	55.4	4.9	
Amharic	Male	87.4	63.8	26.0	0.00
	Female	86.2	62.9	25.9	
Haddiysa	Male	81.5	81.4	10.3	-0.17
	Female	77.0	79.2	6.8	
Sidamu Affo	Male	97.0	81.4	13.1	0.38*
	Female	97.4	79.9	21.7	
Tigrigna	Male	80.1	75.1	13.0	0.09
	Female	75.2	68.3	14.6	
Wolayttatto	Male	73.7	77.9	40.5	0.16
	Female	74.0	77.9	46.3	

Note: (-) Negative sign of Cohen's D indicates difference in favor of boys
* … indicates that the size of difference is educationally significant (Wolf, 1986)

Table 19: Mean scores of untimed tasks by gender and language grade 3

Language	Gender	PA(%)	LC (%)	RC (%)	Cohen's D for RC
Afan Oromo	Male	54.9	84.6	21.3	0.39*
	Female	69.7	86.5	33.6	
Aff Somali	Male	68.7	64.8	21.2	-0.10
	Female	43.7	61.6	18.5	
Amharic	Male	90.2	73.3	39.8	0.18
	Female	92.2	69.1	45.5	
Hadiyissa	Male	87.3	85.3	19.8	-0.28*
	Female	82.7	81.4	12.5	
Sidamu Affo	Male	98.3	85.9	27.1	0.40*
	Female	98.1	83.2	38.4	
Tigrigna	Male	90.2	79.5	25.8	-0.03
	Female	84.8	74.2	25.1	
Wolayttatto	Male	78.9	79.2	52.7	0.15

Language	Gender	PA(%)	LC (%)	RC (%)	Cohen's D for RC
	Female	77.9	81.4	58.0	

Note: (-) Negative sign of Cohen's D indicates difference in favor of boys
* … indicates that the size of difference is educationally significant (Wolf, 1986)

Table 19 above shows the mean scores of the untimed tasks for boys and girls in Grade 3. Looking at the reading comprehension scores (RC), girls substantially outperformed boys in two languages (in Afan Oromo for 12% points and in Sidamu Affo for 11.3% points). Girls were also slightly better in reading comprehension in Amharic (5.7%) and Wolayttatto (5.3%), whereas boys showed substantially better reading comprehension than girls only in Hadiyissa (7.3%). Thus, it can be concluded that in average across all languages girls in grade 3 showed better performance in reading comprehension than boys did.

Figure 12 below shows graphs depicting gender performance in all untimed tasks for both grades and all 7 languages.

Figure 12: Gender Differences in Un-Timed Subtasks by Language and Grade (with Cohen's D)

5.6 Relationship between Oral Reading Fluency and Reading Comprehension Scores

Table 20and 21 show the relationship of the oral reading fluency scores to the reading comprehension fluency score by language and grade. The oral reading fluency is measured as correct words read per minute while the reading comprehension is percent score. In each language, the reading comprehension score increases as the fluency score increases.

Table 20: Relationship between oral reading fluency and reading comprehension scores grade 2

RC (%)	Afan Oromo	Aff Somali	Amharic	Hadiyissa	Sidamu Affo	Tigrigna	Wolayttatto
0%	3.1	2.2	11.3	1.7	4.8	7.0	3.6
20%	17.3	19.3	27.9	17.6	19.8	25.0	22.6
40%	26.8	23.8	40.2	35.8	31.4	28.8	26.4
60%	44.9	41.9	48.0	35.7	40.9	44.8	47.0
80%	49.0	64.2	55.4	50.2	51.5	65.7	50.2
100%	63.8	----	66.7	70.7	76.7	58.0	69.8

Table 21: Relationship between oral reading fluency and reading comprehension scores grade 3

RC (%)	Afan Oromo	Aff Somali	Amharic	Hadiyissa	Sidamu Affo	Tigrigna	Wolayttatto
0%	3.6	4.2	14.4	3.9	6.7	9.8	6.1
20%	17.1	22.4	30.3	17.8	20.6	27.8	24.7
40%	29.6	25.5	42.2	36.1	32.1	33.8	30.6
60%	46.0	40.5	48.8	38.2	43.0	45.3	49.0
80%	51.0	55.0	60.9	49.7	64.5	60.3	52.5
100%	68.5	73.2	71.0	77.1	63.2	63.0	68.3

6 Comparison of 2014 and 2016 results

Table 22 below shows comparisons of the oral reading fluency scores of the baseline and the midterm by grade. The 2014 scores were based on the report of the benchmarking workshop held in Addis Ababa in January 2015[1]. In both grade levels, major score increases are observed in Wolayttatto, Sidamu Afoo, and Amharic, while a major declines are observed in Aff Somali. In other three languages, the sizes of differences were too small to be practically significant.

Looking at Grade 2 scores, it can be observed that Wolayttatto students were able to read 19.6 words more in 2016 than in 2014, whereas in Sidamu Affo and Amharic these gains were 9.2 and 9.5 words, respectively. On the other hand, in Aff Somali, grade 2 students in 2016 are able to read 14 words less than in 2014.

When comparing Grade 3 students, it can be seen that ORF gains were substantially large in three languages: in Wolayttatto, students were able to read additional 20.7 words, in Sidamu Afo 12.7, and in Amharic 11.5 words per minute. However, in Aff Somali it was observed that students' oral reading fluency scores declined for 15.5 words per minute.

Table 22: Comparison of 2014 and 2016 oral reading fluency

Language	Grade 2			Grade 3		
	2014	2016	Cohen's D	2014	2016	Cohen's D
Aff Somali	20.4	6.4	-.89**	32	16.5	-.98**
Afan Oromo	12.1	9.8	-.12	23.9	21.2	-.14
Amharic	19.2	28.7	.49*	30	40.5	.54**
Hadiyissa	6.5	7.5	.06	11.5	14.4	.17
Sidamu Afoo	7.1	16.3	.45*	14.4	27.1	.62**
Tigrigna	13.3	16.4	.17	24.2	26.2	.11
Wolayttatto	11.2	30.8	.85**	20.1	40.8	.81**

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

[1] "Results of the Early Grade Reading Benchmarking Workshop in Ethiopia" USAID Ethiopia & MoE

Table 23below shows comparisons of the mean reading comprehension scores of the 2014 baseline and the 2016 midterm by grade. Similarly as in oral reading fluency, educationally significant score gains were observed in Wolayttatto and Amharic in both grades, and in Sidamu Affo in grade 3. However, a major decline was observed in Aff Somali in both grade levels. In other three languages, the differences were negligible or too small to bear practical significance.

Table 23: Comparison of 2014 and 2016 reading comprehension

Language	Grade 2			Grade 3		
	2014	2016	Cohen's D	2014	2016	Cohen's D
Aff Somali	1.4	0.3	-1.33**	2.2	1.0	-0.87**
Afan Oromo	0.7	0.5	-0.17	1.4	1.4	-0.01
Amharic	0.9	1.3	0.30*	1.6	2.1	0.33*
Hadiyissa	0.6	0.4	-0.17	1.1	0.8	-0.22
Sidamu Afoo	1	0.9	-0.11	1.0	1.6	0.45*
Tigrigna	0.7	0.7	-0.01	1.1	1.3	0.14
Wolayttatto	1.2	2.2	0.54**	2.0	2.8	0.44*

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

Overall conclusion that can be made from the comparisons between 2016 and 2014 student performance in two major reading proficiency subtasks (ORF and RC) is that, in spite of methodological delimitations, there is a compelling evidence of improvements in three languages, no major changes in three languages, and a major decline just in one language. This score, although positive in overall, leaves lots of space for further investigation and improvements.

7 Discussion and Policy Considerations

The purpose of the 2016 EGRA is to provide a litmus score for the seven languages and five regions. Reading improvement is generational and cannot be expected to increase within a short period. The improvement in reading scores will be gradual and perhaps not a straight trajectory.

From the EGRA 2016 results and the results of the surveys conducted simultaneously, several issues emerge for future discussion:

1. **EGRA scores:** Not enough time has gone by for the EGRA scores to reflect the impact of the READ programs. After only a few years of intervention and not all of the textbooks and teachers guides in the hands of early grade readers, with only a few schools having fully equipped supplemental reading programs and after school reading programs, it is simply too early to show any correlation between improved or static EGRA results. Not all of the teachers have been trained in the revised methodology and few principals know how to support the new pedagogy. Furthermore, not all of the systemic issues previously identified in the IQPEP 2014 EGRA report have been addressed such as time on learning, length of school year, and increased time for Mother Tongue instruction. The READ programs, with the support of the MoE and RSEBs, have been successful in providing revised Mother Tongue textbooks and training teachers. READ M&E will address the lack of continuous assessment in the classroom and READ CO is working hard to put engaging supplemental reading material into every child's hand. However, other issues are beyond the control of interventions and must be improved at the core educational level.

Research based recommendations to improve students' reading scores include:

1.1. Increase the time schools are open and operating. The educational system, particularly the RSEBs, have a role in ensuring that schools are open and operating every day of the academic calendar. (Also cited in 2014 IQPEP study). A time on task study conducted in 24 schools in Ethiopia in 2012 found that approximately 69 days of the 203-day school year were lost due to school closings, absenteeism, late starts, early closings, and time-off-task in the classroom. That's approximately 34% of the school year that is unavailable for teaching and learning due mostly to systemic issues (http://www.equip123.net/docs/e2-School_Effectiveness-Synthesis.pdf)

1.2. **Increase time for reading in non-language classes.** Switching from teacher centered reading instruction to student centered reading instruction and putting good reading material into the children's hands will increase the desire to read but the amount of time spent in the direct teaching of reading must be increased. All teachers, math, science, and social studies must devote part of their class time to explicit teaching of reading skills. This explicit teaching can be done with content specific text. All teachers should be reading teachers. Without reading, there is no learning. This is particularly important in the early grades in order to combat the Matthew Effect (Stanovich, 1986) which describes how children who start with a strong foundation in literacy continue to do well while those who do not get this early foundation, continue to get weaker and are unable to ever catch up to their peers. It will continue to be important in later grades as well but the focus should shift more to developing higher level (grade appropriate) vocabulary in the various subjects.

1.3. **Increase time for reading classes.** Currently most regions allow forty-five minutes a day for Mother Tongue instruction. Reading is seen as one of the four language skills. We know that the time it takes to learn to read varies greatly across languages depending on several factors:

- Number of symbols in the alphabet
- Degree of visual similarities in the alphabet (i.e., "d" and "b" and "p" in English)
- Degree of phonological similarities (i.e., "c" in city and "s" in sip)
- Relative orthographic consistency of the language (i.e., English is very opaque - the sound for /k/ is represented 10 different ways, many of the Ethiopian languages are more transparent (i.e. each symbol represents a single sound)
- Whether children speak and understand the language they are learning to read

Many contextual factors also influence the time necessary for children to become fluent readers such as:

- Language level of the teacher
- Motivation of young students
- Perceived value of the language in which literacy is being taught

- Methodology used to teach literacy
- Hours per week spent on language teaching and learning
- Exposure to language and literacy practice outside the classroom

Knowing that literacy forms the foundation upon which children can build knowledge in all other content areas, we must prioritize, particularly in the early grades, time spent on the teaching and learning of discrete literacy skills, as well as on opportunities to see and hear teachers modeling fluent reading, and for children to participate in reading in groups, pairs, individually, both in and out of school. 45 minutes may sound like a lot but in many cases, only 15 or 20 minutes of that time is actually dedicated to the teaching and learning of reading, if that.

2. **Cross-language comparisons:** Although it is tempting, the EGRA results of one language cannot be compared to the results of another language. Scores from one year (2014) to another year (2016) within a language group are a better measure of progress.

While all learners move through the same stages in reading acquisition, the pace at which the learner moves from stage to stage differs by language. Difference in orthography, syllabic complexity, and word length all influence the rate of acquisition.

There has been little research on the standard pace of acquisition in languages in Ethiopia. Although it may not be the most pressing issue in teaching reading, having a clear idea of the standard pace, may help identify children struggling to learn. Understanding the pace of learning, may also assist in pacing the newly revised Mother Tongue textbooks so that children are not pushed or held back in their progress.

3. The 17% decrease in Af-Somali ORF scores and 3.5% drop in Afan Oromo ORF scores from 2014 to 2016, may be an indication of **the loss of time on learning** because of the effect of El Nino and/or continued local strife. Although we cannot attribute in any direct way the small decrease in scores to particular circumstances, we may say that as the issues with the El Nino effect, local strife, and corresponding migration are most severe in the Somali and Oromia region it may point to a need for prompt attention to education for relief.

4. **Gender gap** in Af-Somali in all areas of the EGRA is concerning. For example with correct letter names per minute (clnpm) boys outperformed girls by 11.5 letter names per minute (male 28.4 and female

16.9). In Afan-Oromo and Hadiyissa the difference between male scores and female scores is greater (for Afan Oromo =14.2 and Hadiyissa=13.5). It is not that males outscoring females is in and of itself alarming but worldwide EGRA scores show girls as higher performers than boys. Therefore, it is curious that three out of seven language groups in Ethiopia show the opposite trend. Further research may show that the El Nino effect and the ongoing strike may have affected girls more significantly than boys. This may be due to the need for extra time to carry water, a task traditionally carried out by girls or it may be that girls are not attending school due to security concerns. Further investigation is warranted to ensure that these results are not due to systemic school based issues.

5. Perhaps the most interesting survey question results concerned the **frequency of teaching activities.** Although this data is self-reported and thus not verified, it is none-the-less the teachers' own perception of their teaching activities. As such, the most frequent activity done in classrooms five days a week is "Students copied down text from the chalkboard." Copying from the blackboard may be a simple word or two or it may be a whole lesson- our data does not make this clear. Certainly, in the revised MT textbooks, there are times when the teacher is instructed to have students copy from the blackboard. This is also a factor highly associated with time loss in the classroom as teachers spent inordinate amounts of time copying text onto the board (while students wait) and then students recopy that text into their notebooks. It takes away from more active learning methodologies that can engage students in the learning process.

However, it may indicate that despite the training and availability of MT textbooks, teachers are still teaching in the old methodology (at least 33% of the time). In READ M&E's limited (N=30 in Amhara, Oromia, and Tigray) classroom observations, done for the Formative Continuous Assessment study, the majority of the teachers were actively using the new "I do-you do-we do" methodology. This is confirmed by the findings of the READ TA mid-term evaluation, which found that teachers are highly satisfied with the contents, embedded methodology, and physical qualities of the revised mother tongue materials. From this limited data, we cannot confirm one way or another but the discussion is important. Unfortunately, research confirms that changing teachers teaching practice is an intensive and long-term task. The evidence indicates that teacher attitudes toward educational change are "extremely influential in either facilitating or hindering the installation of a change relative to that issue" (Stern et al., 1975, p. 1). According to Karavas-Doukas (1991), introducing a new program creates competition with "well-established theories of language teaching and learning which are the product of previous teaching and learning experiences, prejudices, and beliefs" (p.188). She goes on to note that teacher attitudes are often

unconsciously held and have a direct effect on what happens in the classroom, including teaching styles (1991). Attitude change is thus a critical component of any pedagogical innovation.

Because attitudes and beliefs are deeply ingrained and often unconsciously held, changing these beliefs can be very difficult. According to Hunzicker (2004), "Changing a teacher's beliefs requires that new information be presented repeatedly over time, to the point that the person begins to feel disequilibrium between current beliefs and new information" (p. 45). This is where professional development programs enter the picture. Through professional development, we have the opportunity to present information about a new innovation, model it in action, and give teachers the opportunity to practice using it, hopefully resulting in changed attitudes and beliefs about that innovation. Professional development is particularly effective when it does not focus solely on implementation of a new program, but also focuses on teachers' attitudes and beliefs around that particular innovation (Hunzicker, 2004, p. 45).

6. Discussion of results

- Creating a culture of reading is a challenging goal. USAID and the Ethiopian Ministry of Education, RSEBs and CTEs have made impressive steps toward a holistic approach to integrating reading into every child's life and every parent's priorities. In February 2011, USAID proposed the five t's as the key to reading success and the READ suite of projects address all but increased time devoted to teaching reading: More **time** devoted to teaching reading
- Better **techniques** for teaching reading (READ TA)
- More **texts** in the hands of children (READ TA and READ CO)
- Teaching children in the **mother tongue** (a language they speak and understand) (READ TA, READ CO and READ M&E
- **Testing** children's reading progress (READ M&E)

It could be said that READ CO, through the reading clubs and libraries, is increasing time spent reading which is extremely important, but it is also critical to increase time devoted to reading in the classroom. The current 45 minute period once a day is not enough time to improve reading as that period is also split with listening, speaking and writing. Furthermore, the length of the school year and the time spent at school on task may also be factors. Certainly, in some regions, schooling in the 2015-2016 school year was

shortened by factors well beyond the control of the Ministry of Education or Regional State Education Bureaus.

Although the 'I do, we do, you do" methodology is well liked and understood by teachers, in the Formative Continuous Assessment Material Development workshop, READ M&E found that teachers and other educators were not fluent in the scope and sequence of reading skills. In other words, more training is needed for early grade educators to internalize the skills necessary for reading acquisition. The second 't' focuses on better techniques for teaching reading, and although the revised textbooks and methodology are certainly helpful, teachers need to be able to teach and assess the five components of beginning reading (phonological awareness, phonics, vocabulary, fluency, and comprehension) without following a script.

Changing teaching methodologies is not an easy task. The revised curriculum, textbooks, and teachers guides provide a well thought-out and accessible means of changing teaching styles. However, teachers tend to teach the way they were taught and it may take more than a few years, and continued training and pedagogical support to create movement toward newer teaching methodologies.

In conclusion, although the 2016 EGRA scores remain quite low despite ambitious interventions, these scores do not indicate a failure of the interventions themselves. It is simply too early to be able to attribute these scores to the interventions which are still in the process of going to scale.

It is useful to remember that there are few demands for literacy in many rural settings and relatively few even in urban areas. Improving national reading scores is a generation goal that cannot be accomplished through improved curriculum, new reading materials, and school based learning alone, but must be thought of as a nation-wide process that increases from one generation to the next.

8 Annexes

8.1 Data collection methods and processes

8.1.1 Roles and responsibilities of assessors, team leaders, and supervisors

There were three groups of assessors. All assessors worked in teams of four. Each group had a supervisor. The supervisor coordinated with the schools, interview the principals and teachers, and test children for part of the day.

One hundred forty assessors were used for this data collection. READ M&E had teams of four assessors responsible for data collection. Three of the assessors were supervised by the team leader. One team-two assessors tested eleven children a day; one assessor tested twelve children a day for ten days. The team leader assessed the remaining students. In this way, a team of four assessors assessed one school/forty children a day.

However, the team leader was the point of contact for the schools.

Team members	Primary responsibilities
Data collectors (3 per team)	Attend training; collect data in a kind careful manner
Data collection team leader (1 per team)	Attend training; collect data in a kind careful manner; supervise other collectors and ensuring the delivery of high quality data; interview principal; organize children for selection; and be the primary contact for schools. If needed, team leader will consult with Dr. Solomon in Addis on swapping schools or other technical/logistical issues.
READ M&E Supervisor	Provide support and supervision to teams collecting data

8.1.2 Types of assessors

Regular assessors

The first group of assessors were called Regular Assessors. Sixteen regular assessors in four teams per language gave the EGRA for 10 days in 10 schools (112 assessors in total). One team-two assessors tested eleven children a day; one assessor tested twelve children a day for ten days. In this way, a team of four assessors assessed one school/forty children a day. After two weeks, each team had tested 400 children in one language. Each language group was assessed at the same time.

READ M&E trained 3 additional regular assessors per language to ensure that only the highest quality data collectors are deployed. Through the Inter-rater reliability (IRR) process, we eliminated three data collectors per language. In case one of the accepted assessors was for some reason unable to perform their duties, READ M&E would have worked with the highest scoring dropped assessor to help them improve and then deploy them to a team. This did not happen.

Plus assessors

The second group of assessors were called "Plus Assessors". To minimize costs, one team from each language group worked an additional week. This group tested an additional 5 school or 200 children from each language group bringing the total number of children by language group to 1,800 and the total number of children accessed to 12,600.

8.1.3 Special Research assessors

The third group of assessors were called "Special Research Assessors". There was one team per language. These assessors conducted a special study using the Common-Persons design (explained in section 2.2.1). They administered the 2014 EGRA paper exam and the 2016 EGRA tablet exam to the same children. Administering the exam this way, each assessors tested 5 children per day instead of 10,leaving each team with 20 students a day rather than 40. Special research teams took 2 days per school or 10 days to access 5 schools.

Week	Regular	Plus	Special Research Assessors	Total
1	**Team #1:** 5 schools/ 200 children **Team #2:** 5 schools/200 children **Team #3:** 5 schools/200 children	**Team #4:** 5 schools / 200 children	**Team # 5**: 2.5 schools/ 100 children	22. 5 schools/ 900 children
2	**Team #1**: 5 schools/200 children **Team #2**: 5 schools/200 children **Team #3:** 5 schools/200 children	**Team #4:** 5 schools /200 children	**Team #5:** 2.5 schools /100 children	22.5 schools/ 900 children
3		**Team #4:** 5 schools/200 children		5 schools/ 200 children
Total # schools	50 schools in each language (*7 languages= 350)			
Total # children	2000 children in each language (* 7 languages= 14,000 children)			

8.1.4 Data Collection tools and end of the day procedures

Data collectors were provided with the required materials; the tangerine tablet, pupil stimuli, sampling sheets, school visit summary sheets, EGRA protocol, pencils, envelopes, clipboards, and folders and a vehicle will be assigned for each team. Special research data collectors will have the 2014 EGRA paper version and data collection sheets.

At the end of each school visit, teams met and discussed issues that were noted during the data collection with their team leaders. Issues noted during the meetings were addressed immediately, calling READ M&E staff as necessary. The daily review meetings provided a forum for sharing the day's experiences. This process proved very useful by providing the assessors with opportunities to learn from each other and correct any data collection misunderstandings they may have had.

8.2 Weighting details

In the context of sampling a fixed number of students from each school, it is important to use weights because the data from a group of 25 cases are representing different number of students depending on the school enrolment in a particular grade. What can happen if we do not apply weighting? We would actually underrepresent large schools, and large schools are mainly in cities, and school in cities usually perform higher, thus, we would underrepresent higher performing schools and consequently the overall student performance in the country would be underestimated. By weighting, we correct this issue.

However, when weighting is applied, it virtually increases the sample size, which causes the probability of increasing Type I errors. The weights can be either designed in such a way that they do not increase the actual number of observed data and don't affect Type I errors, or that the interpretation of any significance testing is done keeping in mind an increased Type I error. In other words, if we run any weighted data analysis checking for significance of differences, because of increased Ns, we will more frequently reject null-hypothesis (expose to Type I error) than it would be the case with real Ncounts.

The 2014 EGRA data was not weighed. Thus, it remains as a delimitation point when we compare 2014 unweighted with 2016 weighted results. Furthermore, the 2014 data was disaggregated into intervention and comparison group and the 2014 EGRA assessed students for whom the test was not their mother tongue

8.3 Independent Sample T-test of the Fluency Tasks by Grade and Language

Language = Afan Oromo

	Grade	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	**Cohen's D**
LSCPM	Three	900	53.9	26.2	12.0	1796	0.000	14.6	0.56
	Two	898	39.3	25.4					
FWCPM	Three	900	19.0	18.3	11.9	1796	0.000	9.1	0.57
	Two	898	10.0	13.6					
IWCPM	Three	900	9.8	11.7	10.3	1796	0.000	4.8	0.49
	Two	898	5.0	7.7					
ORCPM	Three	900	21.2	22.5	12.7	1796	0.000	11.4	0.61
	Two	898	9.8	14.8					

Language = Aff Somali

	Grade	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	**Cohen's D**
LSCPM	Three	662	42.4	28.4	12.7	1369	0.000	18.8	0.69
	Two	709	23.6	26.3					
FWCPM	Three	662	14.3	16.9	10.0	1369	0.000	7.9	0.54
	Two	709	6.4	12.0					
IWCPM	Three	662	14.1	16.4	9.9	1369	0.000	7.7	0.54
	Two	709	6.4	12.0					
ORCPM	Three	662	16.5	18.6	11.8	1369	0.000	10.1	0.64
	Two	709	6.4	12.9					

Language = Amharic

	Grade	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	**Cohen's D**
LSCPM	Three	881	60.7	29.4	8.4	1782	0.000	11.6	0.40
	Two	903	49.0	28.7					
FWCPM	Three	881	41.0	20.0	11.9	1782	0.000	10.9	0.57
	Two	903	30.1	18.5					
IWCPM	Three	881	27.7	14.2	9.7	1782	0.000	6.3	0.46
	Two	903	21.4	13.5					
ORCPM	Three	881	40.5	20.6	12.8	1782	0.000	11.8	0.61
	Two	903	28.7	18.4					

Language = Hadiyissa

	Grade	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	**Cohen's D**
LSCPM	Three	895	48.9	35.07	9.5	1791	0.000	15.0	0.45
	Two	898	34.0	31.35					

FWCPM	Three	895	15.6	19.28	8.2	1791	0.000	6.8	0.39
	Two	898	8.8	15.56					
IWCPM	Three	895	12.7	15.93	6.9	1791	0.000	5.3	0.33
	Two	898	7.4	16.50					
ORCPM	Three	895	14.4	19.16	8.4	1791	0.000	6.9	0.40
	Two	898	7.5	15.27					

Language = Sidamu Affo

	Grade	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	**Cohen's D**
LSCPM	Three	819	70.0	29.65	10.5	1790	0.000	15.1	0.50
	Two	973	54.8	31.03					
FWCPM	Three	819	25.6	18.21	11.6	1790	0.000	9.4	0.55
	Two	973	16.2	15.95					
IWCPM	Three	819	21.8	18.09	10.5	1790	0.000	8.2	0.50
	Two	973	13.5	15.15					
ORCPM	Three	819	27.1	20.84	11.2	1790	0.000	10.8	0.53
	Two	973	16.3	19.98					

Language = Tigrigna

	Grade	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	**Cohen's D**
LSCPM	Three	900	42.8	31.44	6.2	1784	0.000	8.9	0.29
	Two	886	33.9	29.59					
FWCPM	Three	900	37.2	25.00	12.4	1784	0.000	13.8	0.59
	Two	886	23.5	21.58					
IWCPM	Three	900	19.4	13.94	7.4	1784	0.000	4.6	0.35
	Two	886	14.8	12.34					
ORCPM	Three	900	26.2	18.94	11.5	1784	0.000	9.8	0.55
	Two	886	16.4	16.77					

Language = Wolayttatto

	Grade	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	**Cohen's D**
LSCPM	Three	848	72.3	26.24	6.6	1798	0.000	8.5	0.31
	Two	952	63.8	28.31					
FWCPM	Three	848	35.7	20.02	8.4	1798	0.000	8.0	0.40
	Two	952	27.7	20.36					
IWCPM	Three	848	33.0	18.70	9.2	1798	0.000	8.2	0.44
	Two	952	24.8	18.83					
ORCPM	Three	848	40.8	24.39	8.8	1798	0.000	10.0	0.41

	Two	952	30.8	24.05					

8.4 Independent Sample T-test of the Untimed Tasks by Grade and Language

Language = Afan Oromo

	Grade	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
PA(%)	Three	900	62.4	40.57	7.73	1796	0.000	15.1	0.36
	Two	898	47.3	42.40					
RC (%)	Three	900	27.6	32.41	13.325	1796	0.000	17.1	0.64
	Two	898	10.5	20.59					
LC (%)	Three	900	85.6	23.24	3.345	1796	0.001	3.9	0.16
	Two	898	81.7	25.94					

Language = Aff Somali

	Grade	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
PA(%)	Three	662	57.9	41.60	10.517	1369	0.000	23.7	0.57
	Two	709	34.3	41.63					
RC (%)	Three	662	20.0	27.74	11.077	1369	0.000	13.5	0.61
	Two	709	6.5	16.22					
LC (%)	Three	662	63.4	25.29	4.78	1369	0.000	6.8	0.26
	Two	709	56.6	27.09					

Language = Amharic

	Grade	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	
PA(%)	Three	881	91.2	15.62	4.958	1782	0.000	4.4	0.24
	Two	903	86.8	21.22					
RC (%)	Three	881	42.6	31.88	12.084	1782	0.000	16.7	0.57
	Two	903	25.9	26.27					
LC (%)	Three	881	71.2	24.89	6.437	1782	0.000	7.8	0.31
	Two	903	63.4	26.36					

Language = Hadiyissa

	Grade	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	
PA(%)	Three	895	85.0	25.57	4.351	1791	0	5.7	0.21
	Two	898	79.3	29.94					
RC (%)	Three	895	16.2	26.36	6.822	1791	0	7.6	0.32
	Two	898	8.6	20.47					
LC (%)	Three	895	83.4	23.43	2.549	1791	0.011	3.0	0.12
	Two	898	80.3	26.93					

Language = Sidamu Affo

	Grade	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	
PA(%)	Three	819	98.2	10.30	1.941	1790	0.052	1.0	0.09
	Two	973	97.2	11.63					
RC (%)	Three	819	32.9	28.98	12.508	1790	0	15.4	0.59
	Two	973	17.4	23.24					
LC (%)	Three	819	84.5	19.98	3.847	1790	0	3.9	0.18
	Two	973	80.6	22.53					

Language = Tigrigna

	Grade	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	
PA(%)	Three	900	87.5	26.56	6.767	1784	0	9.9	0.32
	Two	886	77.7	34.73					
RC (%)	Three	900	25.5	25.57	11.076	1784	0	11.7	0.53
	Two	886	13.8	18.28					
LC (%)	Three	900	76.9	24.62	4.355	1784	0	5.2	0.21
	Two	886	71.7	26.08					

Language = Wolayttatto

	Grade	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	
PA(%)	Three	848	78.4	29.97	3.125	1798	0.002	4.6	0.15
	Two	952	73.9	31.71					
RC (%)	Three	848	55.3	34.54	7.17	1798	0	11.9	0.34
	Two	952	43.4	35.82					
LC (%)	Three	848	80.3	22.64	2.237	1798	0.025	2.4	0.11
	Two	952	77.9	23.09					

8.5 Independent Sample T-test of the Fluency Tasks by Gender and Language

Language = Afan Oromo, Grade = 2

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
LSCPM	Female	456	42.7	24.30	4.047	896	0	6.8	0.27
	Male	442	35.9	26.10					
FWCPM	Female	456	11.2	14.08	2.736	896	0.006	2.5	0.18
	Male	442	8.7	12.94					
IWCPM	Female	456	6.0	8.47	3.906	896	0	2.0	0.26
	Male	442	4.0	6.76					
ORCPM	Female	456	10.9	15.47	2.239	896	0.025	2.2	0.15

	Male	442	8.7	14.06					

Language = Afan Oromo, Grade = 3

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
LSCPM	Female	458	58.5	25.32	5.516	898	0	9.5	0.37
	Male	441	49.1	26.24					
FWCPM	Female	458	22.7	19.57	6.196	898	0	7.4	0.42
	Male	441	15.3	16.15					
IWCPM	Female	458	12.0	12.96	5.735	898	0	4.4	0.39
	Male	441	7.6	9.84					
ORCPM	Female	458	25.5	23.95	5.948	898	0	8.8	0.40
	Male	441	16.8	19.98					

Language = Aff Somali, Grade = 2

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
LSCPM	Female	295	19.3	24.42	-3.714	707	0	-7.4	-0.29
	Male	414	26.6	27.19					
FWCPM	Female	295	4.1	9.22	-4.326	707	0	-3.9	-0.35
	Male	414	8.0	13.46					
IWCPM	Female	295	4.0	8.84	-4.6	707	0	-4.2	-0.37
	Male	414	8.2	13.62					
ORCPM	Female	295	3.9	9.37	-4.474	707	0	-4.3	-0.36
	Male	414	8.2	14.67					

Language = Aff Somali, Grade = 3

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
LSCPM	Female	285	36.0	30.648	-5.125	660	0	-11.2	-0.40
	Male	377	47.2	25.623					
FWCPM	Female	285	12.3	18.428	-2.607	660	0.009	-3.4	-0.20
	Male	377	15.7	15.474					
IWCPM	Female	285	11.8	17.832	-3.115	660	0.002	-4.0	-0.24
	Male	377	15.8	14.981					
ORCPM	Female	285	15.0	20.254	-1.808	660	0.071	-2.6	-0.14
	Male	377	17.7	17.277					

Language = Amharic, Grade = 2

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
LSCPM	Female	455	49.7	29.82	0.719	901	0.473	1.4	0.05
	Male	448	48.4	27.57					
FWCPM	Female	455	29.6	18.03	-0.75	901	0.453	-0.9	-0.05
	Male	448	30.6	18.89					

IWCPM	Female	455	21.2	13.80	-0.424	901	0.672	-0.4	-0.03
	Male	448	21.6	13.22					
ORCPM	Female	455	28.3	17.89	-0.672	901	0.502	-0.8	-0.04
	Male	448	29.1	18.83					

Language = Amharic, Grade = 3

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
LSCPM	Female	438	63.3	30.10	2.686	879	0.007	5.3	0.18
	Male	443	58.0	28.47					
FWCPM	Female	438	43.1	20.37	3.17	879	0.002	4.3	0.21
	Male	443	38.9	19.43					
IWCPM	Female	438	29.0	14.58	2.804	879	0.005	2.7	0.19
	Male	443	26.4	13.66					
ORCPM	Female	438	43.2	21.11	3.917	879	0	5.4	0.26
	Male	443	37.8	19.71					

Language = Hadiyissa, Grade = 2

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
LSCPM	Female	444	29.3	28.60	-4.483	896	0	-9.3	-0.30
	Male	454	38.6	33.22					
FWCPM	Female	444	6.6	13.30	-4.232	896	0	-4.4	-0.29
	Male	454	11.0	17.22					
IWCPM	Female	444	5.2	11.29	-3.864	896	0	-4.2	-0.27
	Male	454	9.5	20.14					
ORCPM	Female	444	5.6	12.98	-3.833	896	0	-3.9	-0.26
	Male	454	9.5	17.01					

Language = Hadiyissa, Grade = 3

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
LSCPM	Female	446	41.7	33.19	-6.315	893	0	-14.5	-0.42
	Male	448	56.2	35.43					
FWCPM	Female	446	11.8	16.19	-5.978	893	0	-7.6	-0.40
	Male	448	19.4	21.29					
IWCPM	Female	446	9.8	13.56	-5.549	893	0	-5.8	-0.37
	Male	448	15.6	17.53					
ORCPM	Female	446	11.1	16.00	-5.174	893	0	-6.5	-0.35
	Male	448	17.7	21.37					

Language = Sidamu Affo, Grade = 2

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
LSCPM	Female	493	60.5	30.77	5.893	971	0	11.5	0.38

	Male	480	49.0	30.24					
FWCPM	Female	493	19.4	16.24	6.462	971	0	6.5	0.41
	Male	480	12.9	14.95					
IWCPM	Female	493	16.4	15.73	6.155	971	0	5.9	0.40
	Male	480	10.6	13.93					
ORCPM	Female	493	19.7	19.04	5.53	971	0	7.0	0.35
	Male	480	12.7	20.33					

Language = Sidamu Affo, Grade = 3

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
LSCPM	Female	417	74.0	28.10	4.008	817	0	8.2	0.28
	Male	402	65.8	30.65					
FWCPM	Female	417	29.3	18.51	6.019	817	0	7.5	0.42
	Male	402	21.8	17.11					
IWCPM	Female	417	24.4	18.37	4.283	817	0	5.4	0.30
	Male	402	19.0	17.40					
ORCPM	Female	417	30.8	19.70	5.223	817	0	7.5	0.37
	Male	402	23.3	21.33					

Language = Tigrigna, Grade = 2

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
LSCPM	Female	445	33.1	31.66	-0.747	884	0.455	-1.5	-0.05
	Male	441	34.6	27.37					
FWCPM	Female	445	23.3	23.51	-0.286	884	0.775	-0.4	-0.02
	Male	441	23.7	19.47					
IWCPM	Female	445	14.5	12.77	-0.726	884	0.468	-0.6	-0.05
	Male	441	15.1	11.90					
ORCPM	Female	445	16.5	18.78	0.04	884	0.968	0.0	0.00
	Male	441	16.4	14.50					

Language = Tigrigna, Grade = 3

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
LSCPM	Female	448	41.1	31.64	-1.642	898	0.101	-3.4	-0.11
	Male	452	44.5	31.19					
FWCPM	Female	448	35.9	24.51	-1.517	898	0.13	-2.5	-0.10
	Male	452	38.5	25.45					
IWCPM	Female	448	19.0	14.11	-0.943	898	0.346	-0.9	-0.06
	Male	452	19.9	13.77					
ORCPM	Female	448	25.4	19.11	-1.254	898	0.21	-1.6	-0.08
	Male	452	27.0	18.76					

Language = Wolayttatto, Grade = 2

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
LSCPM	Female	481	65.6	28.76	1.936	950	0.053	3.5	0.13
	Male	471	62.0	27.75					
FWCPM	Female	481	29.1	20.09	2.298	950	0.022	3.0	0.15
	Male	471	26.1	20.55					
IWCPM	Female	481	26.5	19.73	2.854	950	0.004	3.5	0.19
	Male	471	23.0	17.71					
ORCPM	Female	481	32.8	25.18	2.592	950	0.01	4.0	0.17
	Male	471	28.7	22.67					

Language = Wolayttatto, Grade = 3

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
LSCPM	Female	423	72.6	25.27	0.243	846	0.808	0.4	0.02
	Male	425	72.1	27.20					
FWCPM	Female	423	37.8	20.25	3.145	846	0.002	4.3	0.22
	Male	425	33.5	19.57					
IWCPM	Female	423	34.7	18.08	2.63	846	0.009	3.4	0.18
	Male	425	31.3	19.17					
ORCPM	Female	423	42.4	22.17	1.893	846	0.059	3.2	0.13
	Male	425	39.2	26.34					

8.6 Independent Sample T-test of the Untimed Tasks by Gender and Language

a Language = Afan Oromo, Grade = 2

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
PA(%)	Female	456	51.1	42.76	2.744	896	0.006	7.7	0.18
	Male	442	43.4	41.71					
RC (%)	Female	456	12.5	22.66	3.026	896	0.003	4.1	0.20
	Male	442	8.4	17.99					
LC (%)	Female	456	81.7	26.03	0.058	896	0.954	0.1	0.00
	Male	442	81.6	25.88					

a Language = Afan Oromo, Grade = 3

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
PA(%)	Female	458	69.7	38.75	5.54	898	0.000	14.7	0.37
	Male	441	54.9	41.08					
RC (%)	Female	458	33.6	34.13	5.77	898	0.000	12.3	0.39
	Male	441	21.3	29.29					
LC (%)	Female	458	86.5	22.88	1.228	898	0.220	1.9	0.08

	Male	441	84.6	23.60					

a Language = Aff Somali, Grade = 2 t-test for Equality of Means

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
PA(%)	Female	295	20.3	34.74	-7.869	707	0.000	-23.9	-0.61
	Male	414	44.3	43.28					
RC (%)	Female	295	4.9	13.61	-2.338	707	0.020	-2.9	-0.18
	Male	414	7.7	17.77					
LC (%)	Female	295	55.4	27.43	-1.041	707	0.298	-2.1	-0.08
	Male	414	57.5	26.83					

a Language = Aff Somali, Grade = 3

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
PA(%)	Female	285	43.7	42.46	-7.998	660	0.000	-24.9	-0.62
	Male	377	68.7	37.58					
RC (%)	Female	285	18.5	29.23	-1.242	660	0.215	-2.7	-0.10
	Male	377	21.2	26.54					
LC (%)	Female	285	61.6	26.12	-1.632	660	0.103	-3.2	-0.13
	Male	377	64.8	24.59					

a Language = Amharic, Grade = 2

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
PA(%)	Female	455	86.2	22.30	-0.815	901	0.415	-1.2	-0.05
	Male	448	87.4	20.08					
RC (%)	Female	455	25.9	25.67	-0.087	901	0.931	-0.2	-0.01
	Male	448	26.0	26.88					
LC (%)	Female	455	62.9	26.77	-0.502	901	0.616	-0.9	-0.03
	Male	448	63.8	25.97					

a Language = Amharic, Grade = 3

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
PA(%)	Female	438	92.2	13.06	1.85	879	0.065	1.9	0.13
	Male	443	90.2	17.76					
RC (%)	Female	438	45.5	32.41	2.694	879	0.007	5.8	0.18
	Male	443	39.8	31.11					
LC (%)	Female	438	69.1	24.70	-2.511	879	0.012	-4.2	-0.17
	Male	443	73.3	24.93					

a Language = Hadiyissa, Grade = 2

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
PA(%)	Female	444	77.0	31.42	-2.235	896	0.026	-4.5	-0.15

	Male	454	81.5	28.28					
RC (%)	Female	444	6.8	17.81	-2.567	896	0.01	-3.5	-0.17
	Male	454	10.3	22.67					
LC (%)	Female	444	79.2	27.52	-1.254	896	0.21	-2.3	-0.08
	Male	454	81.4	26.32					

a Language = Hadiyissa, Grade = 3

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
PA(%)	Female	446	82.7	28.04	-2.71	893	0.007	-4.6	-0.18
	Male	448	87.3	22.64					
RC (%)	Female	446	12.5	23.57	-4.198	893	0.000	-7.3	-0.28
	Male	448	19.8	28.43					
LC (%)	Female	446	81.4	24.13	-2.465	893	0.014	-3.9	-0.16
	Male	448	85.3	22.59					

a Language = Sidamu Affo, Grade = 2

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
PA(%)	Female	493	97.4	11.39	0.519	971	0.604	0.4	0.03
	Male	480	97.0	11.88					
RC (%)	Female	493	21.7	24.94	5.908	971	0.000	8.7	0.38
	Male	480	13.1	20.47					
LC (%)	Female	493	79.9	23.96	-1.054	971	0.292	-1.5	-0.07
	Male	480	81.4	20.95					

a Language = Sidamu Affo, Grade = 3

	Gender	**N**	**Mean**	**Std. Deviation**	**t**	**df**	**Sig. (2-tailed)**	**Mean Difference**	**Cohen's D**
PA(%)	Female	417	98.1	11.12	-0.315	817	0.752	-0.2	-0.02
	Male	402	98.3	9.39					
RC (%)	Female	417	38.4	28.56	5.695	817	0	11.3	0.40
	Male	402	27.1	28.32					
LC (%)	Female	417	83.2	20.85	-1.949	817	0.052	-2.7	-0.14
	Male	402	85.9	18.97					

a Language = Tigrigna, Grade = 2

	Gender	**N**	**Mean**	**Std. Deviation**	**t**	**df**	**Sig. (2-tailed)**	**Mean Difference**	**Cohen's D**
PA(%)	Female	445	75.2	35.34	-2.116	884	0.035	-4.9	-0.14
	Male	441	80.1	33.96					
RC (%)	Female	445	14.6	19.26	1.325	884	0.186	1.6	0.09
	Male	441	13.0	17.21					
LC (%)	Female	445	68.3	26.60	-3.899	884	0.000	-6.8	-0.26
	Male	441	75.1	25.11					

a Language = Tigrigna, Grade = 3

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
PA(%)	Female	448	84.8	29.57	-3.061	898	0.002	-5.4	-0.21
	Male	452	90.2	22.91					
RC (%)	Female	448	25.1	25.06	-0.42	898	0.674	-0.7	-0.03
	Male	452	25.8	26.08					
LC (%)	Female	448	74.2	25.71	-3.241	898	0.001	-5.3	-0.22
	Male	452	79.5	23.22					

a Language = Wolayttatto, Grade = 2

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
PA(%)	Female	481	74.0	32.03	0.119	950	0.905	0.2	0.01
	Male	471	73.7	31.40					
RC (%)	Female	481	46.3	36.55	2.492	950	0.013	5.8	0.16
	Male	471	40.5	34.86					
LC (%)	Female	481	77.9	23.68	0.008	950	0.994	0.0	0.00
	Male	471	77.9	22.49					

a Language = Wolayttatto, Grade = 3

	Gender	N	Mean	Std. Deviation	t	df	Sig. (2-tailed)	Mean Difference	Cohen's D
PA(%)	Female	423	77.9	30.46	-0.481	846	0.631	-1.0	-0.03
	Male	425	78.9	29.50					
RC (%)	Female	423	58.0	33.24	2.269	846	0.023	5.4	0.16
	Male	425	52.7	35.63					
LC (%)	Female	423	81.4	21.67	1.416	846	0.157	2.2	0.10
	Male	425	79.2	23.54					

References

Cohen, J. (1977). Statistical power analysis for behavioral sciences (revised ed.). New York: 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.* Volume 5, Issue 3, 2001. pg. 239-256.

Hirsch Jr., E. D. (2003). Reading comprehension requires knowledge of words and the world: Scientific insights into the fourth-grade slump and the nation's stagnant comprehension scores. American Educator (Spring), 10–44.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. Reading and Writing: An Interdisciplinary Journal, 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. & Ho, C. S.-H. (2000). Naming speed and phonological awareness in Chinese children: Relations to reading skills. Journal of Psychology in Chinese Societies, 1, 93-108.

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

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.

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

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. British Journal of Psychology, 94, 143–174.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children.* Prepared on behalf of the Committee on the Prevention of Reading Difficulties in Young Children under Grant No. H023S50001 of the National Academy of Sciences and the U.S. Department of Education. Washington, DC: National Academy Press.

Wolf, F.M. (1986). Meta-analysis: Quantitative Methods for Research Synthesis. Beverly Hills, CA: Sage.