AFGHAN CHILDREN READ

EARLY GRADE READING ASSESSMENT BASELINE

DARI & PASHTO LANGUAGES

CONDUCTED IN HERAT, NANGARHAR & LAGHMAN, AFGHANISTAN

APRIL AND SEPTEMBER 2017

FINAL BASELINE REPORT

Submitted to:

United States Agency International Development/Afghanistan

COR: Abdul Alim Ghafary

Contract No. 306-TO-16-00003

Submitted by:

Creative Associates International

Submission Date: January 12, 2019

TABLE OF CONTENTS

List of Figures

List of Tables

Acronyms

AIR	American Institutes for Research
BRAC	Building Resources Across Communities
CARE	Cooperative for Assistance and Relief Everywhere
CBE	Community Based Education
CRS	Catholic Relief Services
DED	District Education Directorate
EGR	Early Grade Reading
EGRA	Early Grade Reading Assessment
EMIS	Education Management Information System
FWR	Familiar Word Reading
GPS	Global Positioning System
IWR	Invented Word Reading
LC	Listening Comprehension
MoE	Ministry of Education
NGO	Non-Governmental Organization
ORF	Oral Reading Fluency
PA	Phonemic awareness
PED	Provincial Education Directorate
PPS	Probability Proportionate Sampling
RC	Reading Comprehension
SSI	Syllable Sound Identification
UNICEF	United National International Children's Emergency Fund
USAID	United States Agency for International Development

1. INTRODUCTION

1.1. Afghanistan Country Context

Afghanistan has made significant progress in creating sustainable governing institutions over the past 15 years thanks to the efforts of the Afghan people working in collaboration with their international partners. However, while much has been accomplished, many challenges remain. According to the World Bank, almost 40 percent of Afghans still live in poverty. Nearly one-third of the population is food insecure and many families lack access to basic services according to data from the United Nations World Food Program. Thousands of children suffer from chronic malnutrition and lack access to quality education. The United States Agency for International Development (USAID) seeks to promote stable and inclusive institutions and an economically prosperous country. Going forward, USAID's strategy for meeting development goals calls for sustaining achievements in education, health, and for women; continuing to stimulate agriculture-led economic growth and fiscal sustainability; and supporting legitimate and effective Afghan governance (USAID, https://www.usaid.gov/afghanistan/education).

1.2. Afghanistan Education Context

Improving the quality of education and increasing basic literacy remains a priority for the Afghan government. Overall, literacy rates (defined as those aged 15 and above who can read and write) is just 38.2%, with males at 52% and females at 24.2% (2015 estimations, CIA Fact Book). Complicating the challenge to improving literacy rates is the fact that some students at primary levels are studying in languages of instruction that are not their home language. To address these challenges, USAID is working with the Ministry of Education (MoE) to improve access to education and reading in the early grades through *Afghan Children Read.*

1.3. Objectives

Funded by USAID, the objectives of *Afghan Children Read* are to improve access to education and to improve reading outcomes for primary students in Afghanistan. *Afghan Children Read* supports schools in the formal education system as well as Community Based Education (CBE) schools. The project will achieve its objectives by bolstering the capacity of the government to provide and sustain an evidence-based early grade reading (EGR) curriculum in Dari and Pashto for Grades 1 to 3. An analysis of the results of *Afghan Children Read*'s EGR model for interventions will enable sound policy and targeted support to sustain growth in literacy levels in both public schools and communities where CBE predominates.

Afghan Children Read conducted an Early Grade Reading Assessment (EGRA) baseline in three pilot provinces (Herat for the Dari language, and Nangarhar and Laghman for the Pashto language) in order to generate actionable data on the state of reading outcomes. The report that follows presents an analysis of the results of the EGRA baseline conducted in April 2017 in Herat province, and in September 2017 in Nangarhar and Laghman provinces. The data for Dari and Pashto are not representative of all early grade students of Afghanistan.

1.4. Rationale

To strengthen teacher and administrator skills, teacher preparation systems, and reading materials, reforms and policy initiatives must be undertaken on the basis of quality data on reading outcomes. *Afghan Children Read's* EGRA results will provide essential information to the ministry and donor

community that will enable them to refine and focus their endeavors to promote better reading instruction and materials. The EGRA also provides an opportunity for policy makers and educators to measure and monitor learning progress over the next five years as several EGRA administrations are currently planned.

This report, prepared by American Institutes for Research (AIR) in collaboration with Creative Associates, presents learner results obtained from the baseline data collected in Herat, Nangarhar and Laghman provinces. The purpose of the analyses reported on in this document is to obtain an understanding of the quality of the EGRA instrument, gain insight into student performance in critical reading skills, establish baseline for *Afghan Children Read's* performance indicators, and understand differences between certain student groups within the languages (Dari and Pashto). This report presents evidence about the quality of the EGRA assessment (item and scale levels), EGRA results by subtask, results by key groups of interest (e.g. gender, school type, etc.), and an analysis of the school and home background factors related to reading outcomes.

1.5. Limitations

Afghan Children Read generated reading outcomes data that allow plausible inferences that reflect regional realities on the state of literacy in the two languages. However, due to security concerns and limited mobility, data collection in all the districts of Herat, Nangarhar and Laghman was not possible.

Sampling for the EGRA baseline has some limitations, especially in terms of CBEs. The same sampling approach could not be applied for both public schools and CBEs because the number and locations of CBEs frequently change due to factors which include their funding status from the CBE providers.

Comparison of results between schools and CBEs is another limitation. They are not comparable because of the sampling limitations mentioned above, thus rendering any discussion of which schooling type (public school or CBEs) is performing better than other as unsubstantiated.

There were also limitations in terms of what analytic methods could be used to examine the data. Because of the positive skew in EGRA results (primarily due to the prevalence of a large number of "zero scores"), some types of parametric methods have questionable utility as a primary means to analyze the data. More on the relationship between the reading outcomes and the analytic methods employed to analyze the results can be found in the methods section that follows.

An important caution about the use of EGRA results is also in order. These EGRA results should not be used as a high stakes evaluation of individual schools or students. The purpose of the *Afghan Children Read* data collection was to understand the situation in early grade reading and to employ the results to find ways to improve early grade reading outcomes in Afghanistan.

The Dari baseline results are generalizable only at the provincial level in Herat. because the sampling was not intended to give district level statistical significance. *Afghan Children Read*'s original intent was to establish the baseline for the Dari language based on the entire province. However, due to security concerns and inaccessibility to some parts of certain districts as advised by the Provincial Education Directorate (PED), EGRA was administered in 10 accessible districts of Herat (Herat City, Injil, Guzara, Ghoryan, Kohsan, Karukh, Zenda Jan, Kushk Rubar Sangi, Pashtoon Zarghoon and Obe),

The Pashto baseline results are not generalizable at the provincial level. Due to security concerns and limited mobility, data collection was not possible in all the districts of Nangarhar and Laghman. Only nine out of 28 districts in Nangarhar (Jalalabad, Surkhrud, Goshta, Kama, Dara I Noor, Koz Kunar (Khiwa), Behsud) and Laghman (Mehtarlam, and Qarghayee) were considered viable by the PED for the implementation of EGRA.

2. THEORETICAL PERSPECTIVES

2.1. Early Grade Reading

Research demonstrates that in languages with alphabetic scripts, most children do not learn to read simply from exposure to written words and their pronunciation (Gough et al., 1996). Instead, children need to be explicitly taught the phonological structure of words (the sounds that letters make in words) and then map the sounds they learn to words in print (Hattie, 2014). Most children need to go through an explicit process where they learn to read words first by sounding them out, grapheme by grapheme, or by grapheme cluster, and then learning to recognize patterns of graphemes within words so that recognition becomes automatic. By teaching children to sound out word's grapheme by grapheme or by clusters of graphemes – the decoding process required before automaticity is acquired - they can begin to figure out recently learned and new words independently.

This focus on learning to read words and understanding what the words mean underpins a set of hypotheses that form the *Simple View of Reading* (Hoover & Gough, 1990). In this view, reading is comprised of two constituent parts: learning to read words and understanding the meaning of the words read. In other words, reading consists of the two constructs of decoding skill and language comprehension skill. Hoover and Gough (1990) define decoding as "efficient word recognition; the ability to rapidly derive a representation from printed input that allows access to the appropriate entry in the mental lexicon and, thus, the retrieval at the word level" (ibid, p. 130). They define language comprehension as "the ability to take the lexical information (i.e., semantic information at the word level) and derive sentence and discourse interpretations" (ibid, p. 131). To become a proficient reader, both decoding and language comprehension skills are necessary.

The *Simple View of Reading* has been validated in several languages and writing systems (Joshi, Tao, Aaron, & Quiroz, 2012) and in second language (L2) reading as well (Lervåg & Aukrust, 2010; Proctor, Carlo, August, & Snow, 2005; Verhoeven, van Leeuwe, & Vermeer, 2011). The *Simple View of Reading* with its focus on both decoding skills and language comprehension thus provides a foundational framework to view reading proficiency regardless of the language. In a report produced by the National Reading Panel and the National Institute of Child Health and Human Development (2000) the *Simple View of Reading* identifies 5 core components of reading considered as important for the development of decoding and comprehension skills; they are: Phonological awareness; Alphabetic principle, Fluency, Vocabulary, and Comprehension, the first 3 of which relate to decoding skills, the latter 2 to comprehension. It is important that all reading programs in Afghanistan ensure that children learn about the sounds of the letters in words, how to map the sounds to print, and to learn the meaning of new words (Gove & Cvelich, 2011).

2.2. Factors That Influence Early Grade Reading

Many factors influence achievement in reading. The *Afghan Children Read* project focuses on the five fundamental factors that are amenable to control by system administrators, teachers, teacher trainers, and curriculum and assessment specialists as they develop grade appropriate programming to successfully teach reading. These factors, which are slightly adapted to capture *Afghan Children Read*'s focus, include (Gove & Cvelich, 2011):

- The allocation of sufficient **time** to teach reading
- Access to quality **texts** at the appropriate reading level
- Provision of effective **teaching**
- Appropriate classroom-based **testing**
- Use of the child's mother **tongue**

2.3. Early Grade Reading Assessment in Multilingual Settings

Research supports the idea that for alphabetic scripts there is somewhat of a "universal" framework for teaching children to read (Geva & Genesee, 2006). Instruction in decoding skills and spelling, for example, can be similar across languages and contexts (Lesaux et al., 2006). There is also evidence of a positive relationship between the native language and a second language (frequently a language of instruction) in learning to decode, read sight words, and spell (Comeau et al., 1999). Nonetheless, the EGRA instruments used in this study were designed to meet the needs of mother tongue speakers of specific languages. The Dari instruments differ in some ways from the Pashto EGRA instruments as linguistic nuance, curricula, and culture were taken into consideration during instrument development. As will be highlighted in later sections, it should also be noted that some students sitting for the EGRA in Dari and Pashto were not taking the assessment in their mother tongue.

2.4. Why EGRA?

The EGRA is a battery of early grade reading skills assessments (subtasks) that assess learning outcomes in the core skills essential for the development of reading fluency with comprehension. The EGRA subtasks include assessments of the five core skills: Phonological awareness, alphabetic principle, fluency, vocabulary, and comprehension (Vaughn & Linan-Thompson, 2004; Gove & Cvelich, 2011). We provide a brief definition of each of these 5 core skills in the following table:

Table 1: The Five Core Skills of Reading

Core Skill	Definition
Phonological awareness	• Ability to hear and manipulate the sounds in spoken words • Prepares students to link sounds with print • Precedes decoding (sounding out words) and encoding (spelling) General acquisition progression (from beginning to more advanced): Hear – Identify – Manipulate (from phoneme (sound) to syllable to word)
Alphabetic principle	• Ability to match letters (symbols) to phonemes (sounds) • Basis for an approach (phonics) to teaching reading that stresses letter-sound correspondence and its use in reading and spelling words • Enables students to decode, which is an essential means for reading new words General acquisition progression (from beginning to more advanced):

Core Skill	Definition
	Print awareness – Letter/grapheme awareness – Letter/grapheme-sound correspondence – Decoding – Spelling – Word attack skills
Fluency	• Ability to read words with no noticeable cognitive effort • Must be developed to an 'automatic' level • Automaticity allows student to focus attention on understanding and on content • Research shows that students need to reach a fluency threshold before they can comprehend reliably • Fluency thresholds vary by language General acquisition progression (from beginning to more advanced): Decode accurately – Increase speed – Self-correct – Read with expression
Vocabulary	• Recognition and acquisition of knowledge of words and the meaning of words • Students will only comprehend text if they understand the words they are reading • Familiarity/recognition of words aids in decoding
Comprehension	• Active process of constructing meaning from text • Ability to actively engage with, and derive meaning from, text being read • The goal of reading

The appropriateness of the EGRA for the Afghanistan context stems from three primary rationales:

1. More information is needed about what reading and pre-reading skills need to be enhanced in Afghanistan;
2. EGRA is research-based and has been proven effective in similar contexts (RTI International, 2016). In its initial analyses of the baseline data discussed in this report, AIR conducted what are known as "factor analyses" to show that EGRA data are indeed based on a theoretically valid foundation. We describe the results of these validity analyses in the section "Construct Validity of EGRA Measures" to be found on pages 20-24 of this report.
3. EGRA is efficient to administer, analyze, and score.

While the EGRA does not serve as a diagnostic tool for individual students or as a high-stakes examination for teachers, schools, or regions, it does provide an efficient means to check overall reading progress in early grades and determine where improvement is necessary. A thorough analysis of EGRA results can provide evidence to support improvements in ministerial and local instructional policy, teacher training programs, and classroom-level approaches to improving reading outcomes. Wider dissemination of EGRA results can also encourage greater community and parental involvement in early childhood education and literacy initiatives (Dubeck & Gove, 2015).

3. RESEARCH METHODOLOGY

3.1. Research Design

Over the course of the project, *Afghan Children Read* is piloting an early grade reading program in Dari and Pashto languages in the Herat, Nangarhar, and Laghman provinces of Afghanistan. The project is also working in two districts of Kabul to showcase the implementation of the reading program for

stakeholders. To establish the reading outcomes baseline for the pilot program, *Afghan Children Read* conducted the EGRA in Herat Province for the Dari language (April 2017), and in Nangarhar and Laghman Provinces for the Pashto language (September 2017). The baseline was conducted for completers of completers of Grade 2 and Grade 3[1] in public schools and CBE settings. CBEs are community-based classes established and managed by NGOs/CBE providers such as UNICEF, BRAC, CRS, CARE, and the Swedish Committee Afghanistan. CBEs are associated with the nearest public school that acts as hub school for CBE classes.

3.2. Research Questions

Data from *Afghan Children Read's* 2017 baseline EGRA will be used to answer the following research questions:

(1) What is the overall level of reading at the end of Grades 2 and 3 in Dari and Pashto languages in *Afghan Children Read*'s pilot provinces? In what subskill areas are the students performing well? Where is improvement needed?

(2) How schools and non-school background factors are associated with reading outcomes?

3.3. EGRA Instruments Adaption

Afghan Children Read engaged the MoE in adaptation of the EGRA instruments. Its team, comprised of personnel from the Departments of Academic Supervision, Teacher Education, General Education and early grade teachers, participated in the adaptation process. The EGRA instruments developed under USAID's national EGRA (2016) were adapted for the *Afghan Children Read* baseline. EGRA instrument adaptation took place in a five-day workshop for each of Dari and Pashto languages and was facilitated by international reading experts. Adapted instruments were field tested and finalized based on results from the field test. The following instruments were adapted in both languages:

- Student Test
- Student Questionnaire
- Teacher Questionnaire
- Principal Questionnaire (for public schools only as CBEs do not have principals)

Following the finalization, the instruments were digitized in Tangerine software for use on tablets.

3.4. Selection of Subtasks

The MoE team, facilitated by international experts, adapted the subtasks for the country context using a protocol for their localization. After reviewing primary grade reading standards, textbooks, and other related materials, item writers participated in item and subtask development workshops in which subtasks were developed for all three administrations at the same time (baseline, midline, and endline). The team adapted EGRA subtasks to two main national languages and cultural contexts – Dari in the Herat province and Pashto in the Nangarhar and Laghman provinces – for the 2017 baseline. The baseline consisted of a total of seven subtasks for both grades (2 and 3) and languages (Dari and Pashto). The subtasks measured skills in: Listening Comprehension (LC); Phonemic Awareness (PA); Syllable Sound Identification (SSI); Invented Word Reading (IWR); Familiar Word Reading (FWR); Oral Reading Fluency (ORF); and Reading Comprehension (RC). All EGRA subtasks were administered orally in one-

[1] Due to logistical reasons, the EGRA was administered to these students at the start of Grade 3 (for Grade 2 completers) and at the start of Grade 4 (for Grade 3 completers).

on-one sessions with a test administrator and a single student over a 20 to 25-minute period. During this time, a student questionnaire was also administered. A full description of each subtask can be found in Appendix 1. Below is a brief description of each of the subtasks with a description of what the subtask demands of each student.

Table 2: Summary of EGRA Baseline Subtasks

Subtask	Pre/Reading Skills	Students were asked to:
1. Listening Comprehension (LC)	Oral language comprehension, vocabulary knowledge	Demonstrate listening comprehension of grade-appropriate text by answering 5[2] questions about the text
2. Phonemic Awareness (PA)	Phonemic awareness	Identify the first phoneme from 10 commonly used words by isolating and sounding out just the first sound (phoneme) from a whole word read by the administrator
3. Syllable Sound Identification (SSI)	Decoding, produce sound	Produce the sounds of the written form of a series of syllables (TIMED)
4. Invented Word Reading (IWR)	Letter–sound correspondence, decoding	Read aloud 50 grade-appropriate one to three syllable pseudo words (TIMED)
5. Familiar Word Reading (FWR)	Word recognition and decoding	Read aloud 50 familiar, grade-appropriate unrelated words (TIMED)
6. Oral Reading Fluency (ORF)	Reading with fluency, accuracy, and speed	Demonstrate oral reading of grade-appropriate passage with 66 (TIMED)
7. Reading Comprehension (RC)	Comprehension of text read	Demonstrate comprehension of a passage by answering 5 oral questions, including two inferential questions

3.5. Scoring Process

Timed tasks in EGRA are syllable sound identification, invented word reading, familiar word reading, and oral reading fluency. The scores for these tasks were calculated as the number of correct syllables, invented words, or words, read per minute. Three data points are needed to calculate the total score with the following formula applied:

$$Syllables/words\ \ per\ minute = \frac{total\ read - total\ incorrect}{(60 - time\ remaining)/60}$$

EGRA implements an early stop rule where, if the student does not provide a correct response for the first 10 items, the subtask is discontinued, and the student gets a zero score on the task. Untimed tasks in EGRA are listening comprehension, phonemic awareness, and reading comprehension. The scores for these tasks were calculated as the proportion of correct responses based on the number of questions attempted by the learner.

$$subtask\ score = \frac{total\ correct}{total\ attempted} * 100$$

[2] Numbers of items for each of the subtasks have been recommended in the EGRA toolkit (RTI, 2016) and were observed in the construction and localization of the instruments developed for this baseline.

After administration, each item was reviewed and analyzed to ensure fairness and balance based on gender. Item p-values (a p-value is an estimate of the difficulty level of an item obtained from the number of students who got the item right over the total number of students who answered the item) were evaluated along with item-test correlation coefficients (this is a measure of how well an item discriminates between good and bad students by correlating item score with test score) and differential item functioning (analyses that are used to determine if an item favors a particular sub-group of students) by gender. The full results of these item level analyses can be found in Appendix 3.

3.6. Sampling Method

The purpose of this section is to explain the sampling approach and methods used for collecting baseline data. The student population was divided into two major population groups by school type, i.e. public schools and CBEs for each of the three provinces targeted.

The MoE's Education Management Information System (EMIS) data was used for sampling of schools. Since CBEs were not part of EMIS, CBE data was therefore obtained from the PEDs and CBE providers.

The student population was divided into two major population groups by type of schooling - public or CBE. A multi-stage sampling approach was employed for the EGRA Pashto baseline. Cluster sampling was applied to the total number of students in each relevant grade from the public schools. The cluster samples of schools in the province were proportionately allocated to the districts based on the district's proportional representation in the total population. This allowed for appropriate representation in each of the districts in the sample and reduced skewing towards a district or set of districts. Schools were selected using a probability proportional technique in each district as per proportionate sample size. Ten students per grade per school was the sample size assessed in public schools.

CBEs are viewed as offering alternate, temporary arrangements for education and usually after some time these classes are merged into hub schools. CBEs are unlike schools and typically consist of isolated classes for one grade. In Herat, Nangarhar and Laghman, CBEs with required grades were fewer compared to public schools. Thus, for CBEs in Herat, all CBEs with both grades in the accessible districts were assessed. In Nangarhar's and Laghman's CBEs, there were a significant number of Grade 2 completers, which allowed for the application of applying the same sampling methods as used for public schools. However, the CBEs with completers of Grade 3 were very few in these same two provinces so for Grade 3 completers all CBEs were assessed. In CBEs, 15 students per grade were assessed to get a reasonable student sample.

Enrollment for the two targeted grades was summed up for each school to get the total. Probability proportionate sampling (PPS) was used for random selection of the schools using the cumulative school population. Ten students per grade per school was the sample size assessed for the EGRA. A complete explanation of each step in the derivation of the sample can be found in Appendix 2.

The corresponding numbers of students, teachers and principals by sex for the provinces are provided in Table 2 and Table 3 (after data cleaning).

Table 3: Student by School Type

	School	CBE
Grade 2 completers – Dari	1,440	390
Grade 3 completers – Dari	1,413	312
Total – Dari	2,853	702
Grade 2 completers – Pashto	1,331	988
Grade 3 completers – Pashto	1,316	102
Total – Pashto	2,647	1,090
Grand Total	**5,500**	**1,792**

Table 4: Teachers and Principals by School Type

	School	CBE
Principals – Dari	139	-
Teachers – Dari	285	46
Principals – Pashto	131	-
Teachers – Pashto	266	72

3.7. Data Collection

3.7.1. Selection and Training of Assessors

Afghan Children Read, in consultation with the MoE and USAID, selected two cohorts of assessors from within the MoE for administering the EGRA based on criteria that included: ability to read and speak the necessary languages (Dari in the case of Herat, and Pashto in the case of Nangarhar and Laghman); previous experience administering assessments or with data collection; experience working with primary-age children; availability during the data collection phase; ability to work in target areas; and, experience and proficiency using a computer or hand-held electronic device (tablet or smartphone). *Afghan Children Read* trained 197 assessors (93 for Dari and 94 for Pashto) in two 5-day workshops to administer EGRA in paper and electronic formats from the April 2-6, 2017 for Herat assessment, and September 10-14, 2017 for Nangarhar and Laghman assessment. Training included in-class, as well as, field-based test administration exercises to observe the assessors were trained to administer a quality test. To ensure quality of data, 129 assessors scoring high in the Gold Standard were selected for administration of EGRA.

3.7.2. Data Collection and Cleaning

EGRA data were collected using tablets (100% digitized data collection) by the trained and qualified MoE assessors. A senior official of the MoE Academic Supervision department, with the *Afghan Children Read* M&E team, led the data collection process. To ensure data quality, the data collection process was spot checked 100% by a team of PEDs, District Education Directorates (DEDs) and *Afghan Children Read*. The quality was also monitored through Global Positioning System (GPS) enabled digitized instruments and location data of the sample schools was used to ensure that data collection was done properly in the field.

Using EGRA guidelines, *Afghan Children Read* performed thorough data cleaning before beginning the analysis. Various filters were used to clean the data keeping intact, inter alia, the validity, accuracy, relevance, completeness and reliability of the data.

3.8. Data Analyses and Interpretation

The AIR team analyzed EGRA data at the item level to determine the reliability of EGRA subtasks, sought evidence to ensure the validity of our inferences about reading skills based on those subtasks, and examined the empirical evidence that there is a relationship between Oral Reading Fluency and Reading Comprehension. We present correlation coefficients for reliability, the results of factor analyses to support validity claims, as well as correlation coefficients to demonstrate the relationship between Oral Reading Fluency and Comprehension.[3] In the data tables, we present the number of students in the sample and the mean scores, while the standard deviations (amount of spread across the score distribution) can be found in the appendices for each of the subtasks. The percentage of students that scored zero (also called the zero scores) for each timed subtask is also presented since in many areas of the pilot provinces, there was a relatively high number of students of the sampled schools and CBEs who could not perform on the subtasks.

Mean scores and standard deviations were estimated with sampling weights applied[4]. When comparing performance by groups (e.g. gender, school type), we have provided the results of tests of the statistical significance of mean score differences in the appendices for the benefit of those who wish to see more technical details. The purpose of testing for the statistical significance of differences in outcomes is to determine whether the differences were an effect of chance due to sampling error or were the result of some systematic factor present in the population.

The sample sizes shown in Table 2 and Table 3 contain unweighted sample totals. In the tables and graphs that summarize student achievement results in the following sections of this document, the assumed total samples (Ns) are weighted and do not match unweighted Ns. In some cases, the Ns appear in the tables and should be understood as weighted Ns. The following table compares unweighted and weighted Ns.

Table 5: Comparison of Unweighted and Weighted Sample Sizes

Dari (Herat)								
	Public Schools				CBEs			
	Male		Female		Male		Female	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
Grade-2	710	841	730	916	218	39	172	33
Grade-3	572	639	841	1018	142	32	170	36
Pashto (Nangarhar and Laghman)								
Grade-2	735	1064	596	1011	489	125	499	119
Grade-3	723	739	593	656	27	7	75	15

As data sets with large Ns frequently produce statistically significant results, it is important to estimate an effect size to gain a fuller understanding of the practical effect of any differences in mean scores. In

[3] Each individual subtask item was also analyzed in detail. The full item statistics with difficulty, discrimination, and differential item functioning indices can be found in Appendix 3.

[4] Sampled districts and schools should be proportionally represented in analyses because they are of different sizes and different enrolment rates. For this reason, weights are applied to account for relative differences in sampled units.

the section titled Factors Associated with Reading Achievement (see page 46 of this report) we present the results of analyses conducted with the survey and outcome data. One challenge to analyzing the data with parametric models (e.g. regression modeling) is that an assumption that must be met is that the outcome or dependent variable must be normally distributed. Unfortunately, the baseline data for both grades indicate considerable positive skew due to the large number of "zero" and low scores on the subtasks. Because of this, emphasis in our analyses was placed on bivariate analyses (rather than parametric analyses) where chi-squared statistics were employed to test the relationship between subtask performance (high vs. low scorers) and the categorical background variables of interest. All statistical analyses were conducted using SPSS software.

4. OVERALL FINDINGS

4.1. Construct Validity of EGRA Measures

As noted in the section above "Why EGRA?", an assumption of *Afghan Children Read* was that this combination of subtasks provided a complete assessment of how well students were performing on key pre-reading and reading skills that constitute the two latent constructs of decoding and comprehension (Hoover & Gough, 1990). The purpose of factor analysis is to obtain empirical evidence that the assumption that EGRA does indeed measure the two latent constructs of decoding and comprehension is met. Factor analysis will enable us to determine the number of distinct dimensions or constructs that underlie a theoretical domain of knowledge, trait, or ability measured by an assessment or survey instrument (Kim & Mueller, 1978). Using subtask scores from the Dari and Pashto assessments, the data were analyzed employing factor analysis methodology.

Factor analyses allowed us to confirm that there were indeed 2 dimensions (latent factors, which we confirmed as decoding and comprehension) around which the subtask variables (the manifest variables) were loading. The strength of the loading is calculated as a correlation between the variables.

In Table 6 below we present the results for Dari and Pashto for completers of grades 2 and 3. The indication of two factors can be seen by the clustering of subtasks. For factor 1, Familiar Word Reading, Invented Word Reading, and Oral Reading Fluency were dominant markers showing high factor loading values of over .9, while for factor 2, Phonemic Awareness and Listening Comprehension are clustering together indicating that they are tapping into the same underlying construct. The correlation between the factors is moderate (.53 for Dari, and .45 for Pashto) indicating that we identified two relatively similar, but still different constructs.

In the table below, we use a color-coded scheme for ease of interpretation: Orange for factor 1 (interpreted as decoding), green for factor 2 (interpreted as comprehension). In interpretation of the meaning of those factors we drew on theoretical propositions (e.g., Simple View of Reading). These results help us to understand the construct validity of the EGRA instrument as representing a combination of subtasks that serve for the assessment of reading, assuming that the propositions of the Simple View hold true.

Table 6: Factor Analysis Results (Dari and Pashto, Grades 2 and 3 combined)

Dari			Pashto		
Subtask	Grades 2 and 3		Subtask	Grades 2 and 3	
	Factor 1	Factor 2		Factor 1	Factor 2
Familiar Word Reading (FWR)	1.00		Oral Reading Fluency (ORF)	0.99	
Oral Reading Fluency (ORF)	0.98		Familiar Word Reading (FWR)	0.99	
Invented Word Reading (IWR)	0.96		Invented Word Reading (IWR)	0.83	
Reading Comprehension (RC)	0.81		Reading Comprehension (RC)	0.77	
Syllable Sound Identification (SSI)	0.67		Syllable Sound Identification (SSI)	0.47	
Phonemic Awareness (PA)		0.66	Phonemic Awareness (PA)		0.86
Listening Comprehension (LC)		0.48	Listening Comprehension (LC)		0.42
Factor correlation	0.53		Factor correlation	0.45	

In brief, Table 6 indicates that factor analysis of all EGRA subtasks for Dari and Pashto revealed two major latent traits that determine performance in these subtasks. Factor 1 can be interpreted as decoding or the oral reading fluency dimension, which was strongly marked with the subtasks ORF, FWR, and IWR. It is interesting that the RC subtask was also highly loaded with the decoding dimension, pointing to the fact that ORF represents a strong prerequisite for comprehension. As a matter of fact, considering the nature of the computation of the RC score, this variable could hardly justify the label 'Comprehension' because a very large number of students who cannot read, or read poorly, were not asked any comprehension questions. From that perspective, it may be worthwhile to consider some other label for the RC subtask, such as 'Information processing', which pragmatically reflects both ***ability to decode scripts into information*** and ***ability to understand the information*** that was successfully decoded. With respect to the second factor, tentatively called comprehension, this factor has much weaker associations with the analyzed EGRA subtasks; however, it clearly explains the PA and LC subtasks. Both of these subtasks are characterized by orally provided stimuli, whereas Factor 1 explains the subtasks related to graphically presented stimuli. In other words, Factor I can be viewed as the ***ability to process written information***, and Factor II can be interpreted as the ***ability to process auditory information***. To conclude, factor analysis provides important information that helps in the understanding of the constructs being measured, i.e., the number and nature of latent traits tapped by EGRA subtasks.

4.2. Reliability of Subtasks

The reliability, or internal consistency, of the EGRA subtasks was examined. The term internal consistency refers to the extent to which the items in the instrument measure the same construct consistently. As the reliability coefficient increases, the portion of a score that can be attributed to error will decrease; thus, higher values (generally above .80) are desirable, although a high coefficient does not always indicate quality. Several of the subtasks (e.g., Listening Comprehension, Reading Comprehension, and Phonemic Awareness) have a very low number of items, which can result in low reliability coefficients.

For this EGRA administration, the Listening Comprehension subtask (5 items) indeed showed weak reliability, while Phonemic Awareness (10 items) and Reading Comprehension (5 items) showed

acceptable reliability levels (above 0.80; see tables 7-10 below). Regarding the difference in reliability index between LC and RC, both of which have a total of 5 test items on which reliability is calculated, we do not consider 5 items representing a single construct (or variable) as being too few. Thus, the high reliability index for RC (0.88 for Dari and 0.87 for Pashto) indicates that these 5 items (5 for Dari and 5 for Pashto) are consistently measuring the same variable (or construct), and that therefore the items are well constructed. In the case of LC, reliability is much lower (0.58 for Dari and 0.61 for Pashto), arguably below the limit of acceptability. There may be several reasons for this: some items are not well constructed relative to the text; some items (or the listening text itself) may contain language that confuses students; and students may be forced to guess their responses. To raise the reliability index for LC one would have to evaluate the reliability values for each of the 5 items (5 for Dari and 5 for Pashto) and make judicious changes to those items that appear to be causing "noise".

Table 7: Reliability Coefficient for Listening Comprehension

N of Items	Cronbach's Alpha (Dari)	Cronbach's Alpha (Pashto)
5	.585	.619

Table 8: Reliability Coefficient for Phonemic Awareness

N of Items	Cronbach's Alpha (Dari)	Cronbach's Alpha (Pashto)
10	.926	.889

Table 9: Reliability Coefficient for Reading Comprehension

N of Items	Cronbach's Alpha (Dari)	Cronbach's Alpha (Pashto)
5	.883	.872

Table 10: Reliability Coefficient for EGRA Timed Subtasks

N of Subtasks	Cronbach's Alpha (Dari)	Cronbach's Alpha (Pashto)
4	.827	.706

As standard reliability estimation approaches (such as Chronbach's alpha) are not appropriate for the reliability estimation of timed tests, summary scores of the timed subtasks were used to calculate the overall Alpha for those subtasks (RTI International, 2016). Table 10 contains the reliability coefficients for all 4 combined subtasks that are timed, viz., FWR, IWR ORF, and SSI, providing an overall estimate of their reliability. We see that the reliability for Dari timed tests is reasonably high at 0.82, while for Pashto somewhat low at 0.70. With a view to improving the reliability of the untimed tests for Pashto, it would be appropriate to examine which timed subtask(s) is/are lowering the reliability of the untimed tests.

4.3. Item Quality on The EGRA Administration

Item level evaluation of the EGRA items on each of the subtasks suggested a sound design with a homogeneous distribution of item difficulty and discrimination indices. Discrimination indices refer to the extent to which an item can distinguish between strong and weaker examinees. Differential item

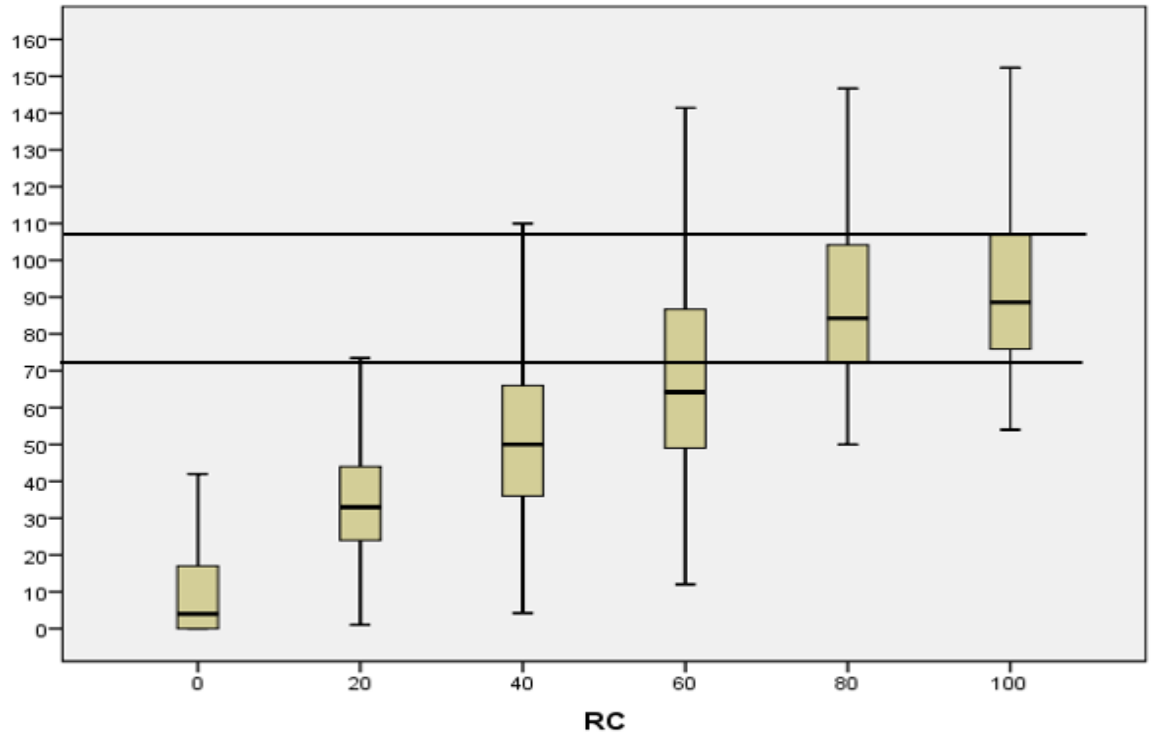
statistics (to check for bias) were also estimated and complete item level statistics can be found in Appendix 3 for all the subtasks.

4.4. Oral Reading Fluency and Reading Comprehension

Before presenting the EGRA results by subtask, an analysis of the relationship between Oral Reading Fluency and Reading Comprehension is in order. That is because it is possible that a student (in some instances) can simply read a list of words (or coherent text as in the ORF subtask) and not fully comprehend what he or she has read. As attention to EGRA results frequently focuses on the number of words per minute read (ORF subtask results), the relationship between this important subtask and Reading Comprehension needs to be established. Many studies indicate that ORF is positively correlated with Reading Comprehension (Dubeck & Gove, 2015; RTI International, 2015).

On the 2017 *Afghan Children Read* baseline EGRA, students responded to up to five reading comprehension questions upon completion of the ORF subtask to ensure they are not simply reciting words in a list they do not understand (comprehension questions were only asked for the portion of the text read by each examinee). The baseline data demonstrated that Reading Fluency and Reading Comprehension were indeed positively correlated. The correlation coefficient between the two subtasks for Dari (both grades estimated together) was 0.57 and for Pashto was 0.72.

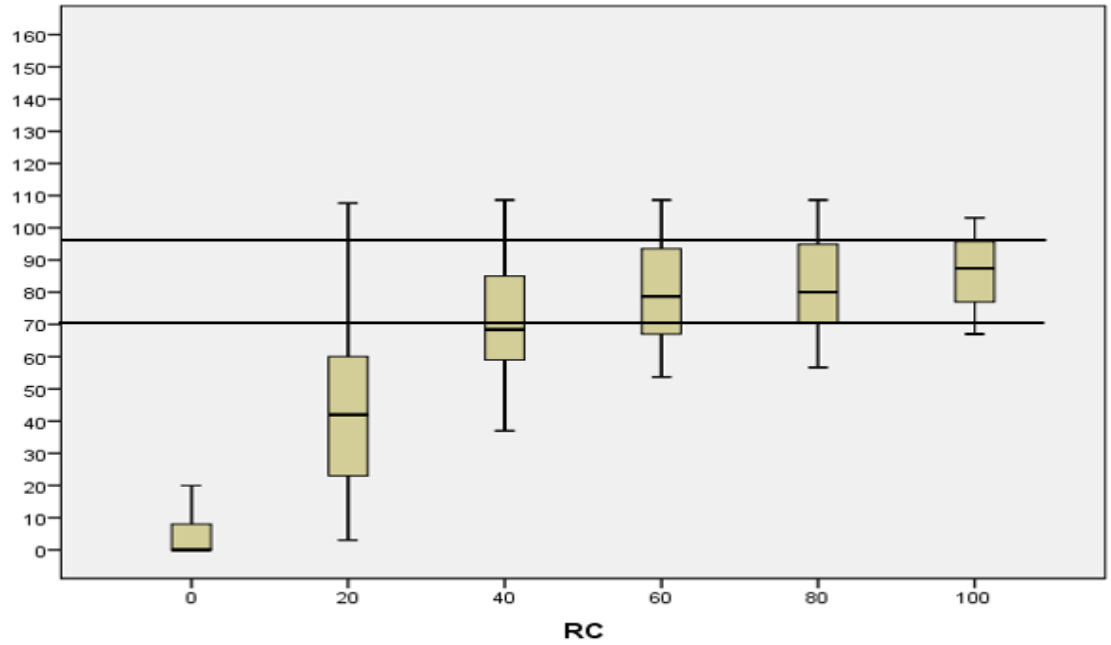
Figure 1: Oral Reading Fluency and Reading Comprehension Grades 2 and 3 (Dari)



This correlation can be visually demonstrated through a "box and whisker plot." Note in Figure 1 above the positive relationship between the number of words read per minute and the number of reading comprehension questions answered correctly. The X axis represents the percent correct scores for Reading Comprehension while the Y axis represents the number of correct words read per minute. As scores rise on one subtask, they rise on the other as well. The edges of the boxes within the plot represent the 25th through 75th percentile range of scores, and the line through the box indicates the median score on ORF. The lines, or "whiskers," indicate the extent of the dispersion of

scores.[5] To a certain extent, the ORF subtask can be considered a proxy for Reading Comprehension and hence give meaning to the metric of "words-read-per-minute."

Figure 2: Oral Reading Fluency and Reading Comprehension Grades 2 and 3 (Pashto)



Note that the two black horizontal lines in Figures 1 and 2 show the range of oral fluency scores coinciding with 4 and 5 comprehension questions correctly responded to (i.e., 80-100% comprehension). In the case of Dari, the range is approximately between 72 words-per-minute to 108 words-per-minute, while in the case of Pashto the range is approximately between 71 words-per-minute to 96 words-per-minute. These ranges are perfectly reasonable and in line with estimates for other languages in other countries.

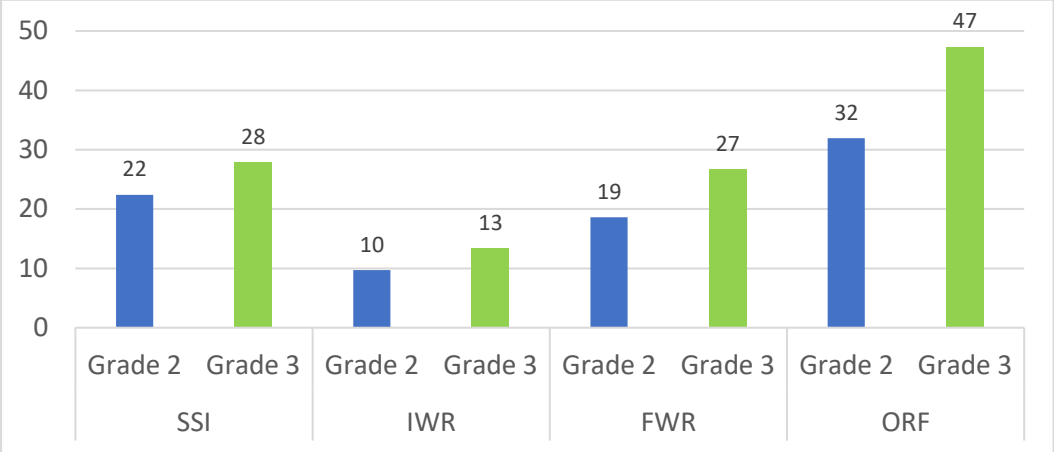
5. EGRA RESULTS BY SUBTASK FOR DARI LANGUAGE IN HERAT

5.1. Subtask Results by Grade Level

We present the mean scores for the subtasks in graphs so as to simplify understanding of the main outcomes and trends. Complete tables with the number of students assessed, standard deviations, and results of statistical tests (coefficients and effect size measures) are to be found in Appendices 4 through 10. Mean scores presented in the text for each subtask were computed by disaggregating the sample by grade level. Subtasks are labeled with their subtask abbreviations where appropriate. For efficiency's sake, the timed and untimed subtasks are normally presented separately.

[5] In several places, scores appear to extend to the extremes where a few learners have read very few words-per-minute yet answer several questions correctly. These extreme data points represent outliers in the data. These "whiskers", or scores beyond the 25th – 75th percentiles, have been removed from the figure in order to make the information easier to understand.

Figure 3: Timed Subtasks Mean Scores (Syllables/Words Per Minute) by Grade (Dari)



In Figure 3 above, there is a perceivable growth trend in subtask outcomes across all four timed subtasks from Grade 2 to Grade 3 in Dari. On the ORF subtask, for example, students at Grade 2 are reading an average of 32 words-per-minute while by Grade 3 they are reading at an average of 47 words-per-minute. For all four of these timed subtasks the differences from Grade 2 to Grade 3 were statistically significant at the .05 level (Appendix 6 for t-test statistics). When ORF values of 32 to 47 words-per-minute are viewed on the Box and Whisker Plot in Figure 1 above, they are associated with a comprehension level of about 20%, in other words Oral Fluency still needs to be improved substantially to reach an acceptable comprehension level of at least 80%.

ORF scores can be further broken down into score ranges. In Table 11 below, we see that 36% of students scored between 0-10 words-per-minute in Grade 2 but this percentage dropped to 21% in Grade 3. However, as Table 11 presents, for both grades the number of examinees concentrated in the upper ORF score ranges (only 14% of students scored above 70 words-per-minute in Grade 2; the percentage rises to 28% in Grade 3) was quite low.

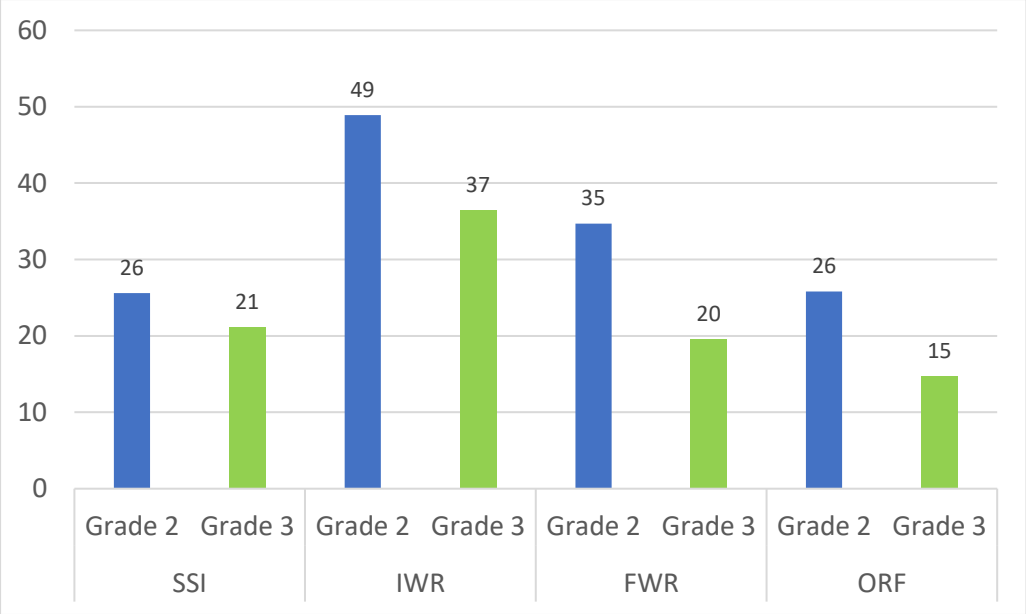
Table 11: Oral Reading Fluency Score Distribution (Dari)

Level	Words Read per Minute	Grade 2		Grade 3	
		Frequency	Percent	Frequency	Percent
1	0-10	659	36.2	356	20.6
2	11-20	236	12.9	186	10.8
3	21-30	186	10.2	184	10.7
4	31-40	133	7.3	120	6.9
5	41-50	109	6.0	159	9.2
6	51-60	121	6.6	92	5.3
7	61-70	120	6.6	133	7.7
8	71-80	55	3.0	136	7.9
9	81-90	82	4.5	98	5.7
10	91-100	31	1.7	82	4.7
11	101-110	34	2	70	4
12	111-120	33	2	38	2
13	121-130	12	1	23	1
14	131-140	9	0	27	2
15	141-150	3	0	10	1

Level	Words Read per Minute	Grade 2		Grade 3	
		Frequency	Percent	Frequency	Percent
16	151-160	0	0	5	0
17	161-170	1	0	4	0
18	171-180	---	---	2	0

Another way to look at the results by grade and the trends across grades is to examine the "zero score" rates. Zero scores refer to instances where students were stopped on a subtask when they could not answer a single item correctly from the first ten items and subsequently received a score of "0" for that subtask. The percentages presented in Figure 4 below reflect the percentage of students who were not able to answer a single item correctly on the first ten items. The Invented Word subtask had the highest rate of zero scores at 49% for Grade 2 while the Familiar Word Reading subtask for Grade 2 had the second highest at 35%. Note in the figure below that the zero score rates systematically decrease from Grade 2 to Grade 3, with the largest decline on the Familiar Word Reading subtask at 15%.

Figure 4: Percentage of Zero Scores for Timed Subtasks (Dari)



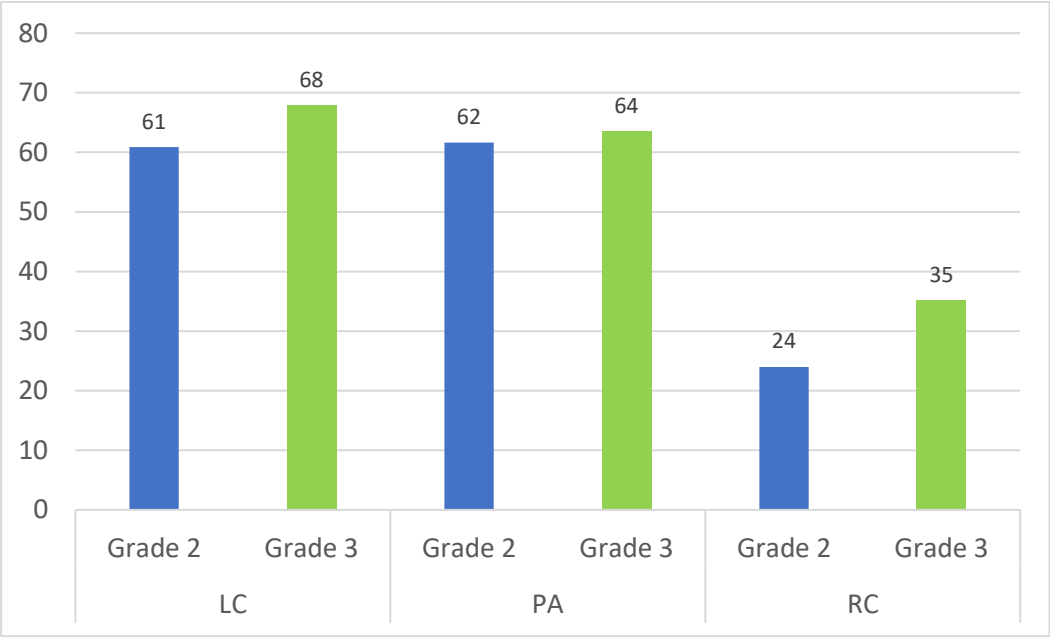
Another way of examining the data is to consider the mean number of words read by students that scored 80% and above in Reading Comprehension (see Table 12 below). Thus, the average number of words-per-minute on Oral Reading Fluency in Grade 2 in Dari was 66 and in Grade 3 was 78 for those students scoring 80% or higher on Reading Comprehension. Note also that there are in the data set, associated with these high levels of Reading Comprehension, some extremely low minimum scores of oral reading fluency. These can only be interpreted as anomalies or outliers.

Table 12: Mean Scores (Correct Syllables/Words Per Minute) of Students Performing at 80% and Above Correct on Reading Comprehension (Dari)

Subtask	Grade 2	Grade 3
	Mean (wpm)	Mean (wpm)
SSI	40	44
IWR	21	24
FWR	39	45
ORF	66	78

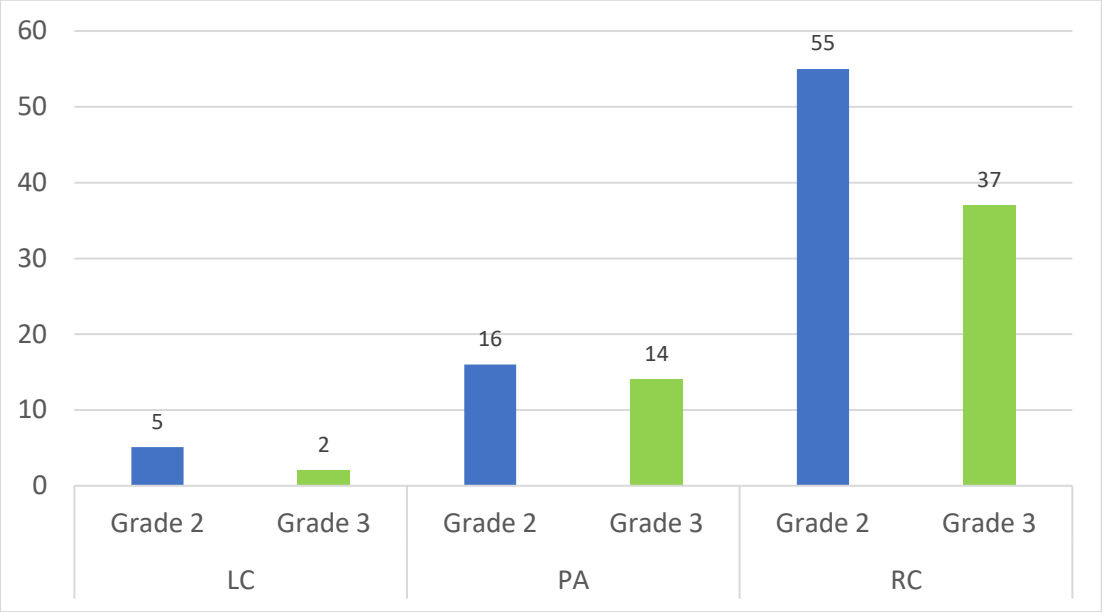
The results of the untimed subtasks - Listening Comprehension, Phonemic Awareness and Reading Comprehension – also indicate increases in performance from Grade 2 to Grade 3, as was observed for the timed subtasks (see Figure 5 below). The largest percentage increase was for RC with an 11% increase. These increases were statistically significant for LC and RC but not for PA. While these increases are important to observe, the levels achieved in Grade 3 for RC for example are still significantly below acceptable reading comprehension levels (established at 80% and above) – in other words, by Grade 3 students in the Province of Herat still need to improve on reading comprehension by about 135%.

Figure 5: Mean Percentage Correct Scores Untimed Subtasks by Grade (Dari)



As was the case with the timed subtasks, the percentage of "zero scores" on the untimed tasks also decreased from Grade 2 to Grade 3 (see Figure 6). The subtask with the largest decrease across grades was Reading Comprehension with an 18 percentage-point decrease in the number of students with zero scores from Grade 2 to Grade 3. However, zero scores in RC in both Grades 2 and 3 are particularly high, falling from 55% to 37%.

Figure 6: Percentage of Zero Scores for Untimed Subtasks (Dari)



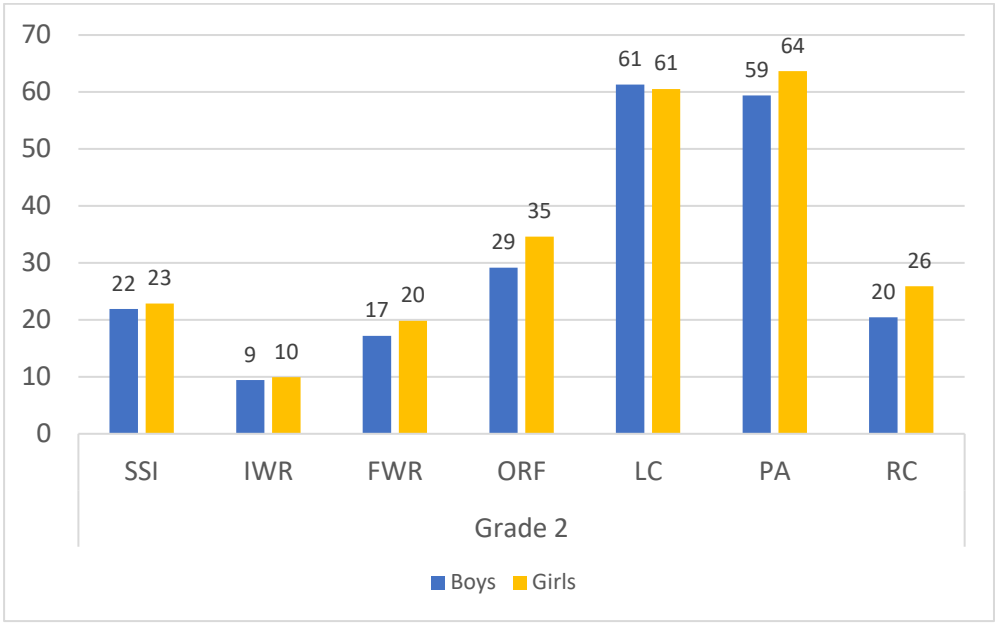
Summary and conclusions regarding overall subtasks scores for Dari

- Student achievement levels significantly improve from Grade 2 to Grade 3 in all subtasks measured in the baseline; however, overall, these improvements still leave students far short of an acceptable level in Reading Comprehension (of 80% and above). In Grade 2, the mean ORF score is 32 words-per-minute (with Reading Comprehension at 24%), in Grade 3 ORF is 47 words-per-minute (with Reading Comprehension at 35%).
- The percentage of students who score zero points on Reading Comprehension (55% in Grade 2 and nearly 40% in Grade 3) is extremely high; the percentage of zero scores on Oral Reading Fluency (the predictor variable for Reading Comprehension) is also high, with a quarter of students in Grade 2 scoring zero, which falls to 15% in Grade 3.
- Listening comprehension mean scores are high (at 61% in Grade 2 and 68% in Grade 3). Apart from indicating that oral-based skills develop easily in Dari, these scores also indicate that comprehension of text as a general construct does not appear to be a problem in learning; what inhibits comprehension of reading text is the still under-developed decoding skills as seen in the low ORF, FWR, IWR and SSI scores.
- For a baseline, these results are hardly surprising and leave room for significant improvement with the application of appropriate pedagogical strategies contemplated under *Afghan Children Read*. This will be ascertained in the midline evaluation.

5.2. Subtask Results by Sex

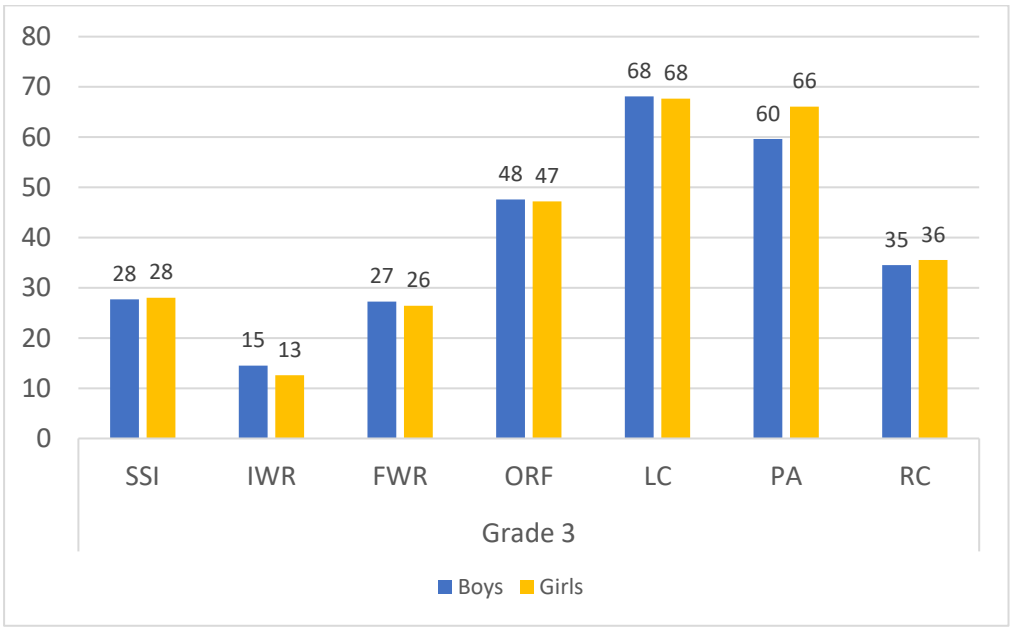
The Grade 2 results by sex (Figure 7 below) indicate that numerical differences between the scores for Dari favored girls over boys on all but the LC subtask. However, on none of the 7 subtasks were these

Figure 7: Grade 2 Mean Scores for Subtasks by Sex (Dari)



differences statistically significant at the .05 level (see Appendix 7 for t-test statistics). We can for all intents and purposes consider that there is no significant difference between girls' and boys' performance on the baseline EGRA in Grade 2.

Figure 8: Grade 3 Mean Scores for Subtasks by Sex (Dari)



For Grade 3 (Figure 8 above), we observe a similar pattern to that of Grade 2, although boys have reduced some of the gaps and even outperformed the girls on 3 of the measures (IWR, FWR, and ORF). Differences are, however, not significant on any of the 7 subtasks at the .05 level, so we can again conclude that there is no statistical difference between girls' and boys' performance on Grade 3.

Recall that mean scores for FWR, IWR, SSI, and ORF are presented in terms of syllables/words-per-minute, while LC, PA, and RC are presented as percentage correct.

Another way of looking at the data is to consider the percentage of boys and girls who scored zero on each of the subtasks by grade. In Table 13 (below), we can see that the lowest percentage of zero scores was for Listening Comprehension (both genders and grades) while the highest percentage of zero scores is on the Reading Comprehension subtask.

Table 13: Percentage of Zero Scores by Grade and Sex (Dari)

Subtask	Grade 2		Grade 3	
	Boys	Girls	Boys	Girls
SSI	27	25	19	22
IWR	50	48	34	38
FWR	38	31	19	20
ORF	28	24	15	15
LC	5	5	3	3
PA	20	15	17	13
RC	60	51	38	38

Summary and conclusions regarding subtasks scores by sex for Dari

- In Grade 2, girls outperform boys on most measures (6 of 7) but by differences that are statistically insignificant. By Grade 3 it appears that these differences are even smaller and in the case of 3 of 7 subskills boys slightly outperform girls, but differences in Grade 3 continue to be statistically insignificant.
- Zero scores on Reading Comprehension in Grade 2 are high, and significantly higher for boys – 60% of boys scored zero versus 51% of girls. These drop to equal percentages by sex in Grade 3 (38%). By contrast, zero scores on Listening Comprehension are very low, ranging from 5% to 3% by sex across grades, reinforcing the notion that the obstacle to Reading Comprehension is lack of appropriate fluency rates.
- By the end of Grade 3 both boys and girls appear to be somewhat distant from developing appropriate levels of Oral Reading Fluency to permit acceptable levels of Reading Comprehension (viz. 48 and 47 words-per-minute on ORF with 35% and 36% RC).

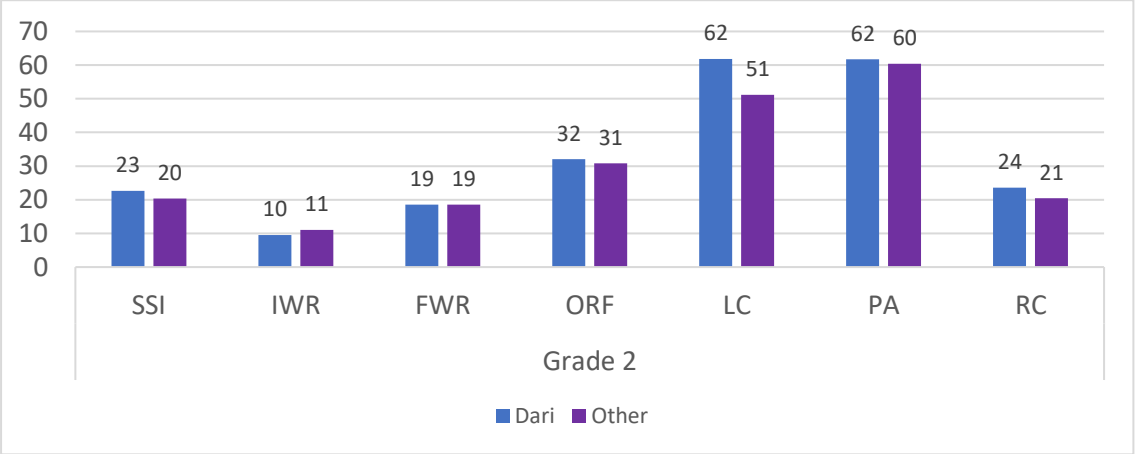
5.3. Subtask Results by Language Spoken at Home

Another variable of interest was whether or not mean scores would be similar across the various language groups in the Province of Herat. Students learning to read in a language other than their mother tongue could be at a disadvantage. While most respondents (some 90% in Grade 2 and 94% in Grade 3) reported that Dari was the language they spoke at home, there were nonetheless pockets of students whose home language was another language such as Pashto (9.7% in Grade 2 and 6.1% in Grade 3), Uzbek, Balochi, Turkmen or other.

For Grade 2 (Figure 9 below), there were statistically significant differences (.05 level) only on Listening Comprehension, favoring the Dari speaking cohorts. For Listening Comprehension, Dari speakers scored 11 percentage points higher than their non-native Dari speaking peers (effect size = .2). This was hardly surprising given that early listening skills are linked to the language of the home environment.

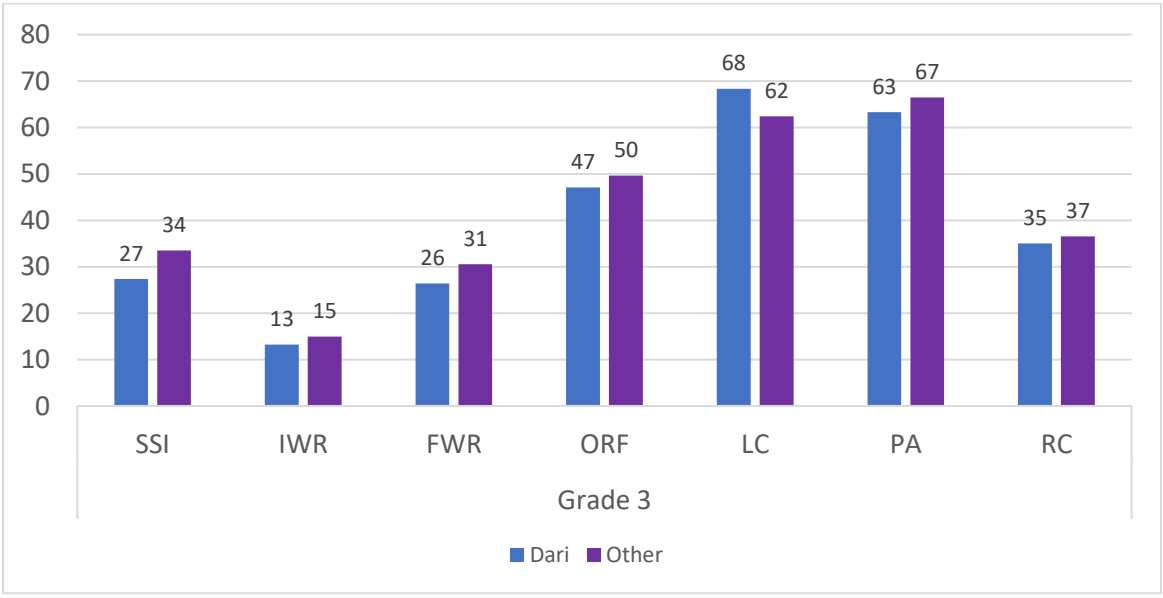
There were no statistically significant differences on any of the other 6 early grade reading measures. For Reading Comprehension, Dari speakers scored 3 percentage points higher than speakers of other languages (effect size = .4) although this was not a significant difference.

Figure 9: Grade 2 Mean Scores by Language Spoken at Home (Dari)



Interestingly, for Grade 3 the results were slightly different (Figure 10 below). This time, on all but one measure (Listening Comprehension) Dari speakers were outperformed by students whose home language was not Dari. However, on only one measure was there a statistically significant difference in favor of non-Dari speakers, on SSI, with a mean score difference of some 6% points. For all remaining 5 EGR measures there was no significant difference in mean scores, although they all showed non-Dari speaker higher scores. Recall that mean scores for FWR, IWR, SSI, and ORF are presented in terms of syllables/words-per-minute, while LC, PA, and RC are presented as percentage correct.

Figure 10: Grade 3 Mean Scores by Language Spoken at Home (Dari)



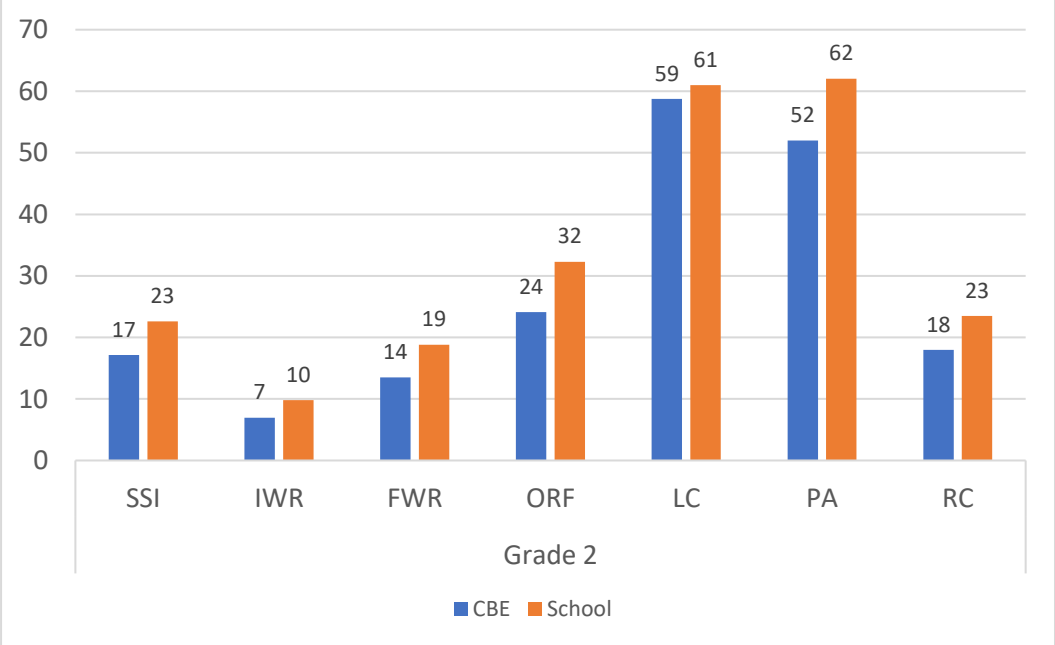
Summary and conclusions regarding language spoken at home for Dari

- These baseline results seem to suggest that from Grade 2 to Grade 3 students who do not speak Dari at home, i.e., they are native speakers of languages other than Dari, significantly improve some of their decoding skills in Dari to the point where they appear to have caught up with or surpassed the level of their Dari peers. Note that the numbers of students who were non-Dari speakers in the home are extremely low in this sample. The issue requires further investigation. Regardless of language background, all students are remote from developing acceptable decoding skills to permit appropriate levels of Reading Comprehension, as we have seen in other sub-group scores.
- Of those students in Grade 2 who scored zero on Reading Comprehension, 20% were speakers of Pashto (i.e., not the language of the test, Dari, represented by 78% of the total of zero scorers). In Grade 3, a similar pattern exists: 16% of students in Grade 3 who scored zero were Pashto speakers while 81% were Dari speakers. It would appear from this information that being a speaker of Pashto in a Dari-speaking instructional environment may continue to be a problem in Grade 3. Again, given the weak representation of native speakers of languages other than Dari in this baseline sample, the issue would merit further research.
- Dari speakers outperform their non-Dari peers by 11% in Listening Comprehension in Grade 2, which is halved by Grade 3 (a difference of 6%).

5.4. Subtask Results by School Type

Figures 11 and 12 below present results for the EGRA subtasks by school type: public school and CBE. For Grade 2, public school students outperformed CBE students on all EGR measures, with 5 of the 7 subtasks showing statistically significant differences. For ORF, there was an 8 words-per-minute advantage for public school students over CBEs. In RC, public school students outperformed CBE students by about 5% (18% vs. 23% comprehension levels) although this was not a significant difference.

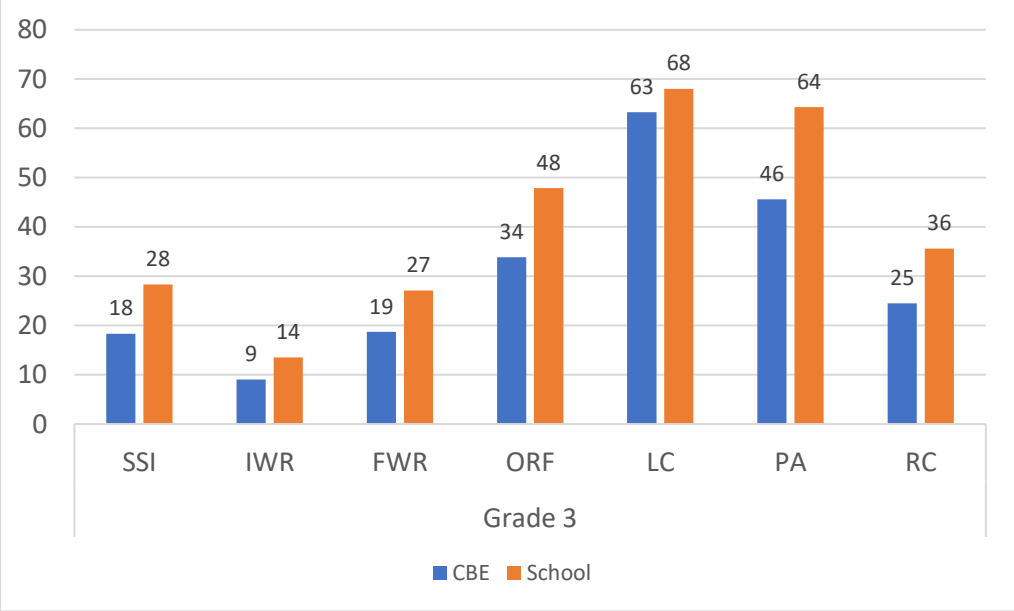
Figure 11: Grade 2 Mean Scores by School Type (Dari)



By Grade 3, the gaps between public schools and CBEs appear to be growing - it is important to point out, however, that the number of students in Grade 3 CBE schools (27 male and 75 female) was too low to provide reliable information.

Public school students outperformed CBE students on all measures, as in Grade 2, there was an increase in the number of measures for which there was a significant difference in score (6 of 7 subtask scores showed significant differences), and the mean scores on each measure got larger. For example, on ORF there was a 14 word-per-minute difference in favor of Public school students (48 vs 34 words-per-minute), while on RC the percentage difference was 12% (36% comprehension vs 24%). As we have seen with all sub-group scores, there are improvements from Grade 2 to Grade 3, but these increases do not reach levels of decoding (fluency) and comprehension that would be considered appropriate. By the end of Grade 3, Public school students can comprehend close to 40% of a text. See Appendix 9 for the results of the t-tests and effect size values.

Figure 12: Grade 3 Mean Scores by School Type (Dari)



CBE classes were established to provide educational resources to children in remote rural areas who are unable to attend formal schools due to insecurity, distance or other constraints. EGRA achievement results for CBEs, which clearly fall significantly below the results of the public-school system, must therefore be viewed in light of the complex educational context in which they are obtained, results which without CBE facilities would inevitably be significantly worse. Note in Table 13 that there are clear significant mean-score differences between urban and rural public schools (in favor of urban public schools, as we discuss in the next section of this report) and significant differences between rural public schools and CBEs (which are only found in rural contexts as explained above).

Table 14: Subtask Scores by Type of School and Location for Grades 2 and 3 (Dari)

Subtask	Grade 2						Grade 3					
	Public Schools				CBEs		Public Schools				CBEs	
	Urban		Rural		Rural[6]		Urban		Rural		Rural	
	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N
Syllable Sound Identification	26	545	21	1209	17	72	32	516	27	1140	18	68
Invented Word Reading	13	545	9	1209	7	72	16	517	12	1137	9	67
Familiar Word Reading	24	542	16	1211	13	72	33	517	25	1138	19	67
Oral Reading Fluency	41	538	29	1212	24	72	56	517	44	1141	34	67
Listening Comprehension	65	545	59	1212	59	72	71	517	66	1141	63	68
Phonemic Awareness	68	545	59	1212	52	72	68	517	62	1141	46	68
Reading Comprehension	30	545	20	1212	18	72	41	517	33	1141	24	68

Summary and conclusions regarding scores by type of school and location for grades 2 and 3 for Dari

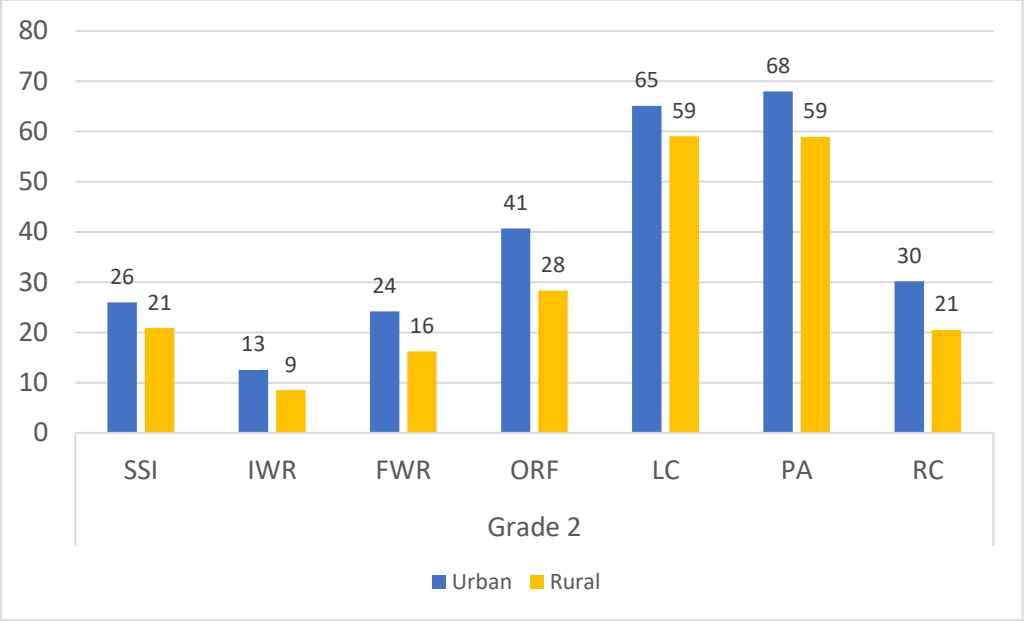
- Public schools outperform CBEs with significant differences on most subskills measured. Differences get larger from Grade 2 to Grade 3. Bear in mind that achievement in both Public schools and CBEs remains well below the targeted achievement level in ORF and RC. Also note that the grade 3 sample is small and does not provide a reliable comparison
- When a comparison is made of urban public schools, rural public schools, and CBEs (which are only found in rural locations), there is a significant decline with CBEs achieving lowest scores on all subskills measured, rural public schools outperforming CBEs, and public schools outperforming all others.
- Note: The Grade 3 sample is too small to provide a reliable comparison.

5.5. Subtask Results by School Location

Subtask results by urban and rural schools present a clear picture of the state of reading outcomes in Herat Province. In Grade 2, there are statistically significant differences (at the .05 level) in all subtask measures results between urban and rural schools, all differences favoring urban schools (see Figure 13). The effects sizes for these differences are in the small to medium range. At Grade 2, urban students are reading 41 words-per-minute while rural students are reading only 28. In Reading Comprehension, urban students outperform rural students by 9% (30% vs. 21%).

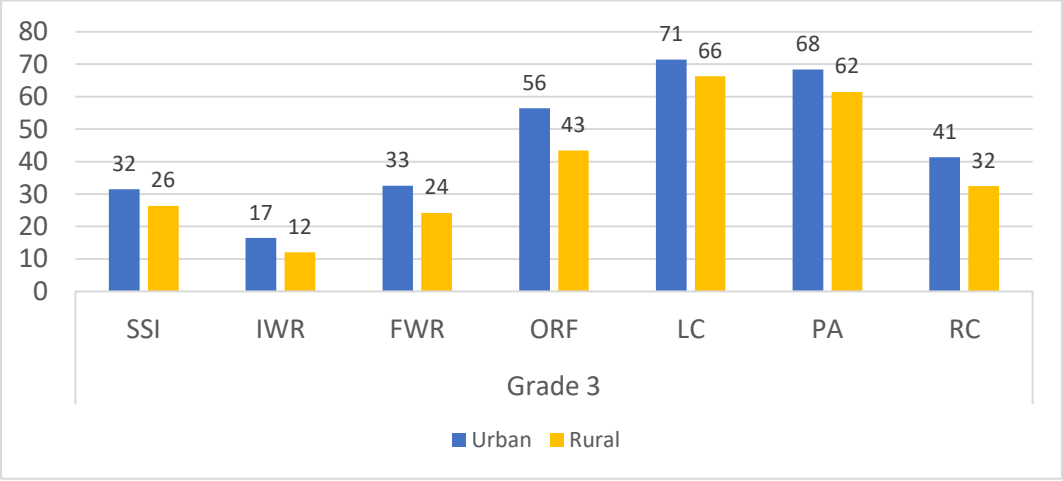
[6] Note that CBE classes are not provided in urban contexts.

Figure 13: Grade 2 Mean Scores by Location (Dari)



At Grade 3 (see Figure 14), the situation is similar across subtasks with significant differences on all measures except on LC and PA. On ORF, urban students' mean score was 56 words-per-minute while rural students' mean score was 43 words-per-minute. In terms of RC, urban students outperformed rural students by 9% points (41% vs 32%).

Figure 14: Grade 3 Mean Scores by Location (Dari)



Summary and conclusions regarding mean scores by location for grades 2 and 3 for Dari

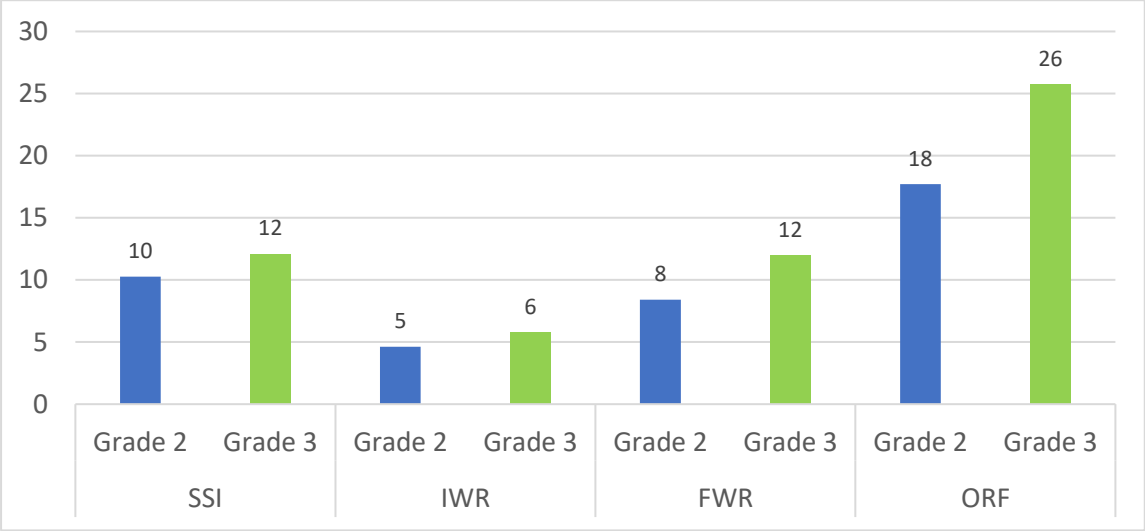
- Urban schools outperform rural schools on all subskills measured in Grade 2 and 3. Improvements on all subskills are evident from Grade 2 to 3 but insufficient to qualify students as good readers, with Oral Reading Fluency falling well short in Grade 3 of a minimum target of 72 words-per-minute (see Figure 1 above) associated with 80% or more comprehension.

6. EGRA RESULTS BY SUBTASK FOR PASHTO LANGUAGE IN NANGARHAR AND LAGHMAN

6.1. Subtask Results by Grade Level

As with the Dari results in the previous section, we present the mean scores for the subtasks for Pashto in graphs to simplify understanding of the main outcomes and trends. Complete tables with the number of students assessed, standard deviations, and results of statistical tests (coefficients and effect size measures) are to be found in Appendices 4 through 10. Mean scores presented in the text for each subtask were computed by disaggregating the sample by grade level. Subtasks are labeled with their subtask abbreviations where appropriate. For efficiency's sake, the timed and untimed subtasks are normally presented separately.

Figure 15: Timed Subtasks Mean Scores (Syllables/Words Per Minute) by Grade (Pashto)



In Figure 15 (above), we observe a growth trend in subtask outcomes across all four timed subtasks from Grade 2 to Grade 3. On the subtask with the largest gains, Oral Reading Fluency, students at Grade 2 are reading an average of 18 words per minute while by Grade 3 they are reading at an average of 26 words per minute. For 2 of these timed subtasks (FWR and ORF), the differences from Grade 2 to Grade 3 were statistically significant at the .05 level but the effect sizes were small (see Appendix 6 for t-test statistics). Clearly ORF levels are in the case of both grades significantly below values associated with acceptable reading comprehension (of 80%+), which for Pashto are calculated to be at 70 words-per-minute or above. Current mean ORF scores in Grade 2 are associated with reading comprehension at 0-20%, with 20% RC in Grade 3. Actual mean RC scores (as shown in Figure 17 below) are slightly higher but still well below desired levels.

Before presenting mean score results for each individual subtask for Pashto, some overall findings will illuminate the state of reading in the Provinces of Nangarhar and Laghman. Note that in Table 15 below, there is a very high concentration of scores at the lower end of the scale. 65% of students read only between 0 and 10 words-per-minute at Grade 2. While the percentage drops to 53% by Grade 3, this number is still over half the total respondents. As shown in the "box and whisker plot" for Pashto (see page 24), acceptable reading comprehension (of 80% and above) is related to an ORF score of at least

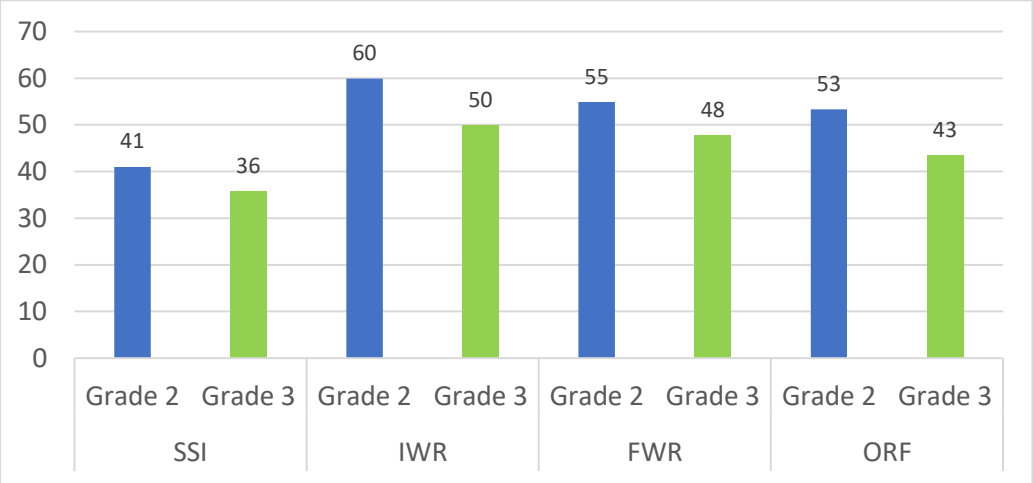
70 words-per-minute, which as shown in Table 15 is achieved by a total of only 8% of students in Grade 2 and 16% in Grade 3.

Table 15: Oral Reading Fluency Score Distribution (Pashto)

Level	Words-read-per-minute	Grade 2		Grade 3	
		Frequency	Percent	Frequency	Percent
1	0-10	1473	64.8	713	53.4
2	11-20	171	7.5	112	8.4
3	21-30	105	4.6	69	5.2
4	31-40	85	3.7	77	5.8
5	41-50	71	3.1	37	2.7
6	51-60	78	3.4	49	3.7
7	61-70	109	4.8	63	4.7
8	71-80	53	2.3	71	5.3
9	81-90	56	2.5	56	4.2
10	91-100	39	1.7	55	4.1
11	101-110	31	1.4	34	2.6

Another way to look at these baseline results by grade and the trend across grades is to examine the rates of zero scores, i.e., the percentage of students who score zero on a subtask. The percentages presented below in Figure 16 reflect the percentage of students who were not able to answer a single question correctly on the first ten items. Their attempt to complete the subtask was thus stopped and they received a zero score for that subtask. The Invented Word subtask had the highest rate at 60% for Grade 2 while the Familiar Word Reading subtask for Grade 2 had the second highest at 55%. Note in the figure below that the rates of zero scores decrease from Grade 2 to Grade 3, with the largest decline on the Invented Words Reading subtask at a 10% decline and Oral Reading Fluency with an 10% decline. Of particular note, especially, is the high level of zero scores in ORF at over 50% of all sampled students in Grade 2 and just over 40% in Grade 3.

Figure 16: Percentage of Zero Scores for Timed Subtasks (Pashto)



Another interesting way to examine the data is to look at the mean scores for SSI, IWR, FWR, and ORF (decoding subtasks, which are all timed tasks) for those students who demonstrate a high level of Reading Comprehension (in this case 80% or higher correct on the RC questions). In Table 16 below

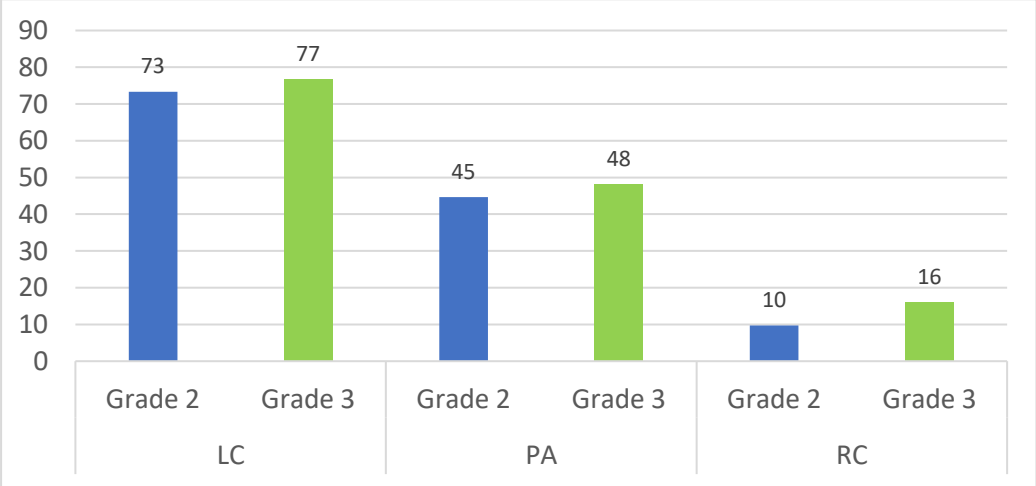
we present these results for Pashto. The table shows that for ORF in Grade 2 the mean number of words-per-minute is 55 for those students who have an acceptable level of comprehension while in Grade 3 the mean number of words-per-minute is 62. These mean ORF scores are slightly below desired levels for 80%+ comprehension, but given that these are mean values (i.e., represent the average of the complete sample) they take into account the large percentage of students who actually scored zero on this subtask. We should also point out that there are also some very low ORF scores associated with high comprehension levels – these scores are anomalies or outliers in the data set, i.e., may be the result of data entry errors produced by Tangerine.

Table 16: Mean Scores (Correct Syllables/Words Per Minute) of Students Performing at 80% and Above Correct on Reading Comprehension (Pashto)

Subtask	Grade 2	Grade 3
	Mean (wpm)	Mean (wpm)
SSI	26	26
IWR	14	14
FWR	26	30
ORF	55	62

The results of the untimed subtasks (see Figure 17) – Listening Comprehension, Phonemic Awareness, and Reading Comprehension – indicate increases in performance from Grade 2 to Grade 3. The largest percentage point increase was for Reading Comprehension at +6%. The changes across years were statistically significant for RC but not for PA and LC. Listening Comprehension scores are unsurprisingly good with over 70% comprehension in both grades evaluated. Contrastingly, Reading Comprehension mean scores are extremely low with 10% comprehension in Grade 2 rising to 16% in Grade 3. These scores are far below desired values, reinforced by the fact that fluency rates are correspondingly low (18 and 26 words-per-minute in Grade 2 and 3 respectively).

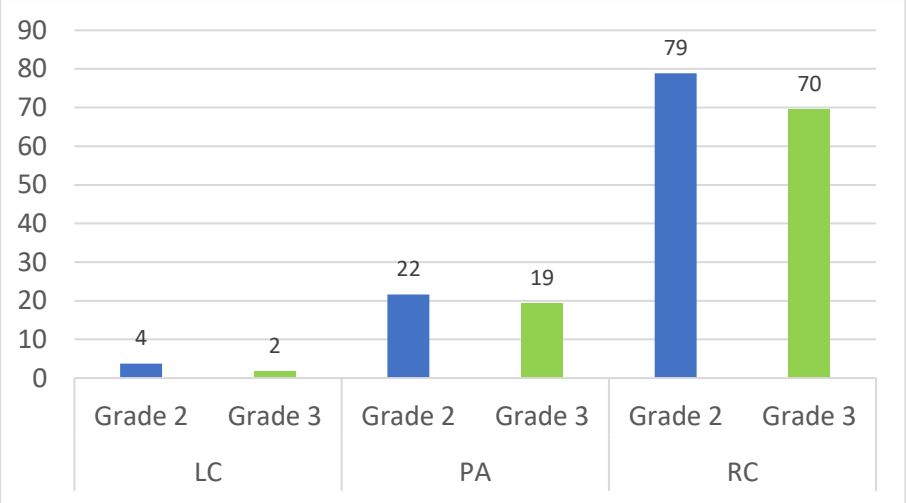
Figure 17: Mean Percentage Correct Scores Untimed Subtasks by Grade (Pashto)



As was the case with the timed subtasks, the percentage of "zero scores" on the untimed tasks also decreased from Grade 2 to Grade 3 (see Figure 18 below). The subtask with the largest decrease across grades was Reading Comprehension with a 9% decrease in the number of students with zero scores from Grade 2 to grade 3. However, zero scores on Reading Comprehension are high, 79%

falling to 70% in Grades 2 to 3 respectively, that is approximately three quarters of the students sampled in both grades fail to score a single point on this subtask.

Figure 18: Percentage of Zero Scores for Untimed Subtasks (Pashto)



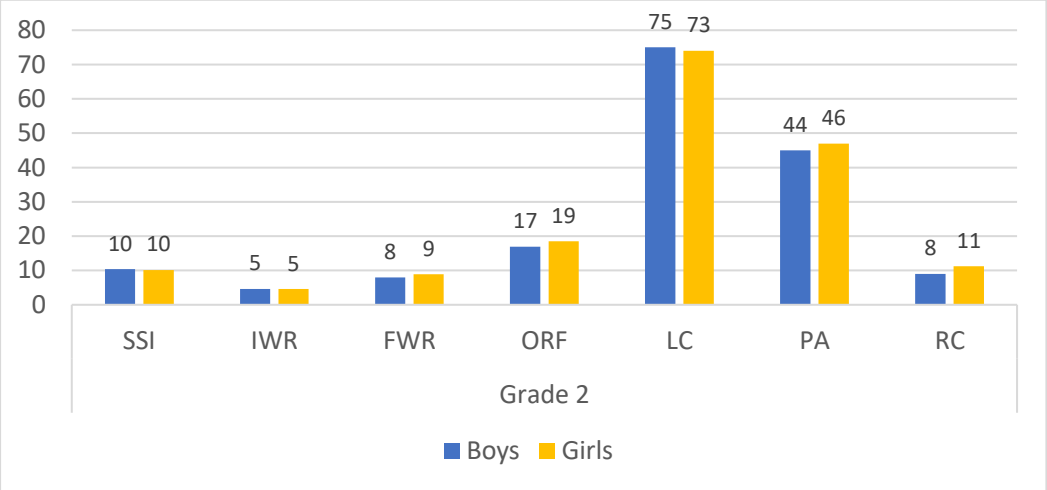
Summary and conclusions regarding overall subtasks scores for Pashto

- Despite significant improvement on all subtasks from Grade 2 to Grade 3, mean achievement scores on all subtasks are extremely low with only 8% of students in Grade 2 reaching an acceptable level of Reading Comprehension (of 80% or above) which rises to 16% in Grade 3.
- In Grade 2, the mean ORF score is 18 words-per-minute and RC at 10% while in Grade 3 mean ORF score is 26 words-per-minute and RC is 16%; lack of comprehension skill is entirely predictable from the low level of fluency.
- In terms of percentage of students scoring zero on subtasks, 53% in Grade 2 and 43% in Grade 3 scored zero on ORF, with 79% in Grade 2 and 70% in Grade 3 scoring zero on RC.
- Students in both grades for Pashto scored high on Listening Comprehension (73% and 77% on Grades 2 and 3 respectively) – indicating that comprehension as a construct is not a problem for acquisition, rather students lack sufficient decoding skills to achieve the same levels of comprehension in reading.

6.2. Subtask Results by Sex

The EGRA 2017 baseline results by Sex (Figure 19 below) indicate that there are statistically no differences between the achievement of boys and girls in Grade 2. Mean scores mirror those overall scores described above – high Listening Comprehension performance, low decoding skills, very

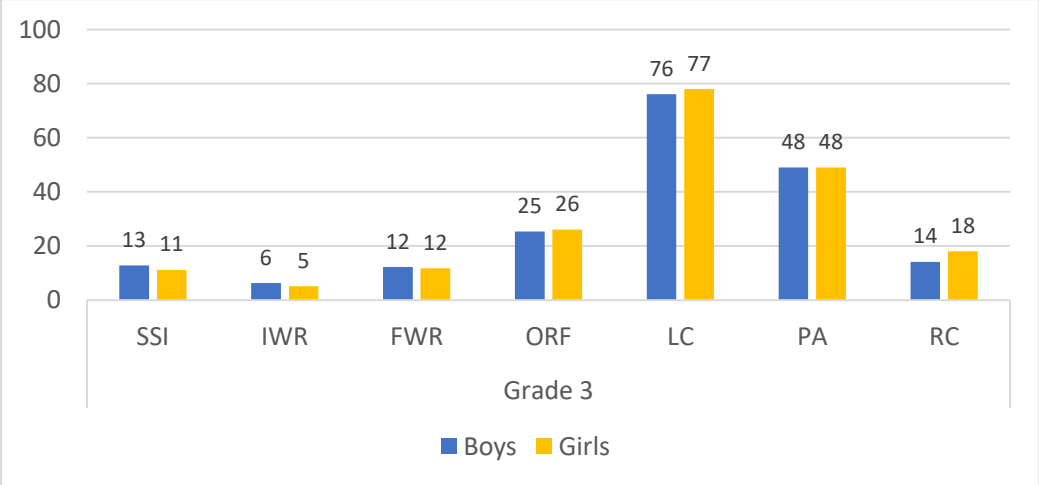
Figure 19: Grade 2 Mean Scores for Subtasks by Sex (Pashto)



low Oral Reading Fluency and Reading Comprehension scores. There are no significant differences on any of the EGR measures for Grade 2, indicating that boys' and girls' scores are about the same, but as indicated above scores that are still very low, especially in ORF and RC.

Mean scores for Grade 3 (Figure 20 below) provide the same picture although there is an increase in performance from Grade 2 to 3. None of the Grade 3 scores are statistically significant across boys and girls, indicating that scores by Sex are the same, and continue to be too low by the end of Grade 3 for acceptable Reading Comprehension. Recall that mean scores for FWR, IWR, SSI, and ORF are presented in terms of the number of syllables/words-correct-per-minute, while LC, PA, and RC are presented as percentage correct.

Figure 20: Grade 3 Mean Scores for Subtasks by Sex (Pashto)



Another way to examine results by Sex is to compare zero scores across these two groups. In the table below (Table 17), we present the total number of students receiving zero scores (in percentages) by Sex. Note that the highest proportion of zero scores are for Reading Comprehension for both boys and girls (from approximately 80% to 70% in Grades 2-3 respectively). Zero scores on all word-level decoding tasks are also high (from approximately 55% to 45% in Grades 2-3 respectively), providing a clear explanation of the reason for low Reading Comprehension. On a note of optimism, there is 10%

improvement from Grade 2 to 3. With appropriate pedagogical support, one would expect to see achievement improve to more acceptable levels of reading ability.

Table 17: Percentage of Zero Scores by Grade and Sex (Pashto)

Subtask	Grade 2		Grade 3	
	Boys	Girls	Boys	Girls
SSI	41	41	33	39
IWR	59	61	46	54
FWR	55	55	46	50
ORF	53	54	41	46
LC	3	4	2	2
PA	22	21	19	20
RC	81	76	70	69

Summary and conclusions regarding subtask scores by sex for Pashto

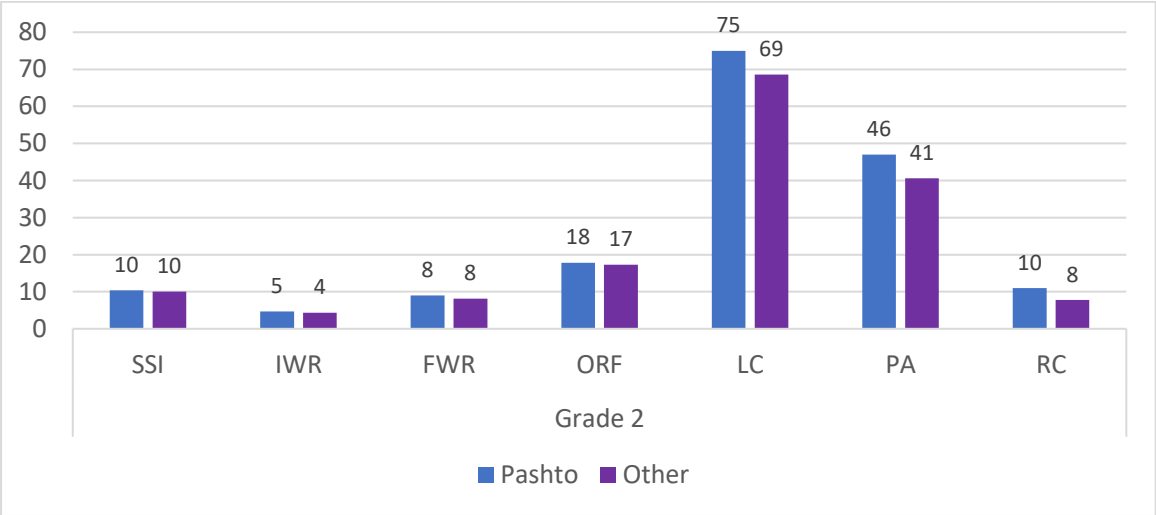
- There is no significant difference between girls' performance and boys' performance on any reading indicator for either grade.
- All reading scores improve by some 10% from Grade 2 to 3 although remain significantly low by the end of Grade 3 (Grade 2 ORF scores are 17 words-per-minute (boys) and 19 words-per-minute (girls) improving to 25 and 26 words-per-minute in Grade 3. These ORF scores correctly predict the low comprehension scores that were achieved.
- Comprehension scores in Grade 2 are at 8% (boys) and 11% (girls) respectively increasing to 14% and 18% in Grade 3 respectively.

6.3. Subtask Results by Language Spoken at Home

Most sampled respondents reported that Pashto was the language they spoke at home; however, there were some students whose home language was identified as Pashai, Balochi, Turkmen or other. Although the numbers of home speakers of languages other than Pashto in the Provinces of Nangarhar and Laghman were very small, making valid comparisons difficult, we expected the data to provide some initial sense of how much of an obstacle to the acquisition of Pashto reading skills it is for non-native speakers of Pashto, i.e., not to be a native speaker of the language of instruction.

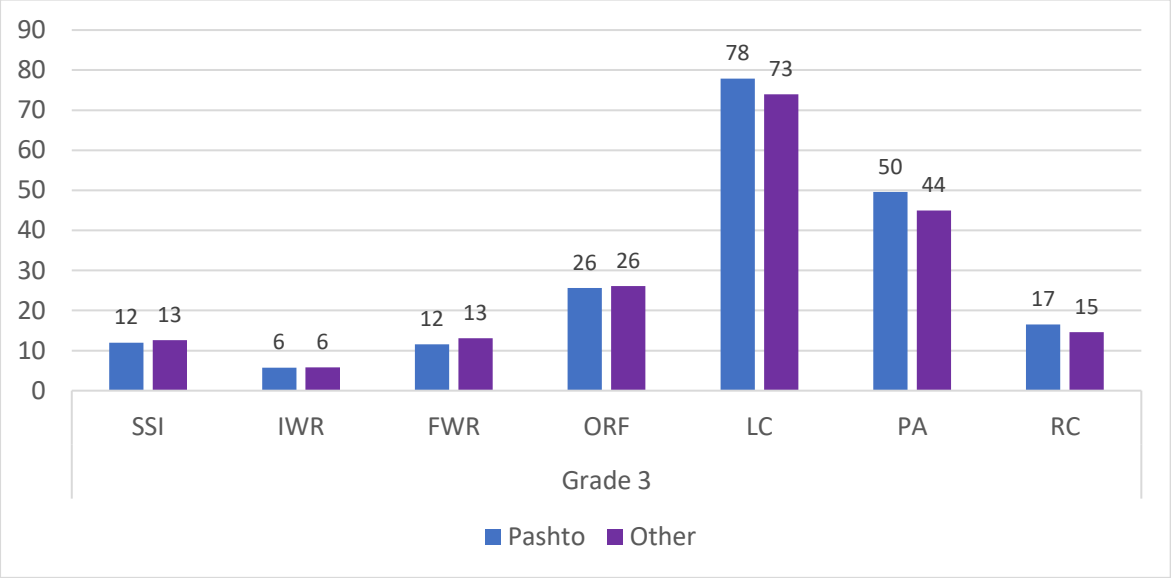
For Grade 2 (Figure 21), Pashto speakers scored higher than speakers of other languages on all EGR measures for the 7 subtasks. For none of these were there statistically significant differences except for Listening Comprehension. For LC, Pashto speakers scored 6 percentage points higher than their non-native Pashto speaking peers (effect size = .3). For Reading Comprehension, they scored 2 percentage points higher (effect size = .2). It is interesting to note that non-native speakers of Pashto scored reasonably high on Listening Comprehension (at 69%) although outperformed by Pashto speakers.

Figure 21: Grade 2 Mean Scores by Language Spoken at Home (Pashto)



Trends were somewhat similar for Grade 3 although the gaps seem to have become smaller between the two groups, with speakers of languages other than Pashto outperforming Pashto speakers on SSI and FWR measures, and scoring lower than Pashto speakers on LC, PA and RC. None of these measures across Home Language showed statistically significant differences except for LC, as one would expect.

Figure 22: Grade 3 Mean Scores by Language Spoken at Home (Pashto)



Improvements from Grade 2 to Grade 3 are evident, at 7% on Reading Comprehension, and by some 8 words-per-minute on ORF. These improvements are small and give grounds for optimism but are certainly still at a very low level with ORF in need of a 100%-200% increase to approach acceptable Reading Comprehension levels. Recall that mean scores for FWR, IWR, SSI, and ORF are presented in terms of syllables/word per minute, while LC, PA, and RC are presented as a percentage correct.

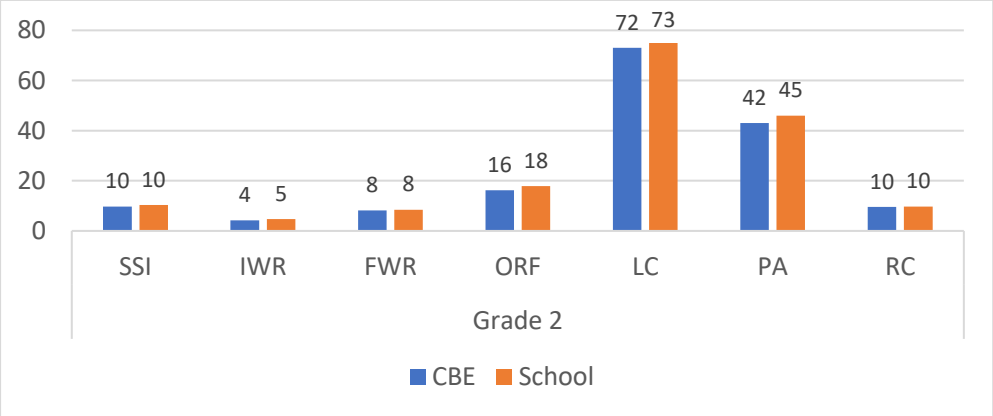
Summary conclusions regarding language spoken at home for Pashto

- While sample sizes for non-native speakers of Pashto in Nangarhar and Laghman are too low to provide valid comparisons, the tentative evidence from these results suggests that non-native speakers of Pashto are not at a significant disadvantage over native Pashto speakers. Note that on Listening Comprehension, non-native speakers of Pashto score reasonably well in both grades (69% and 74% in Grades 2 and 3) although were outperformed by native speakers of Pashto.
- In terms of the key variables of interest, decoding skills improve from Grade 2 to Grade 3 for both groups, but remain very low by the end of Grade 3 (ORF: 18 and 17 words-per-minute to 26 and 26 words-per-minute, with Pashto native speakers cited first; in RC, 10% and 8% in Grade 2 to 17% and 15% in Grade 3, Pashto native speakers are cited first).

6.4. Subtask Results by School Type

Figures 23 and 24 (below) present the results for the Pashto EGRA subtasks by school type, Public school and CBE on each of the 2 grades. For Grade 2, Public School students outperformed CBE students on 4 of 7 EGR measures (IWR, ORF, LC and PA), and on the remaining measures scored exactly the same as CBE students. However, in Grade 2 none of the results show statistically significant differences, indicating that CBE students did no worse, and no better, than Public school students. In Grade 2, ORF word-per-minute scores favored public school students (18 over 16 words-per-minute) while RC scores were the same at 10% correct. As we have seen for all groups measured, decoding levels in Grade 2 are extremely low and predictive of the low Reading Comprehension scores obtained.

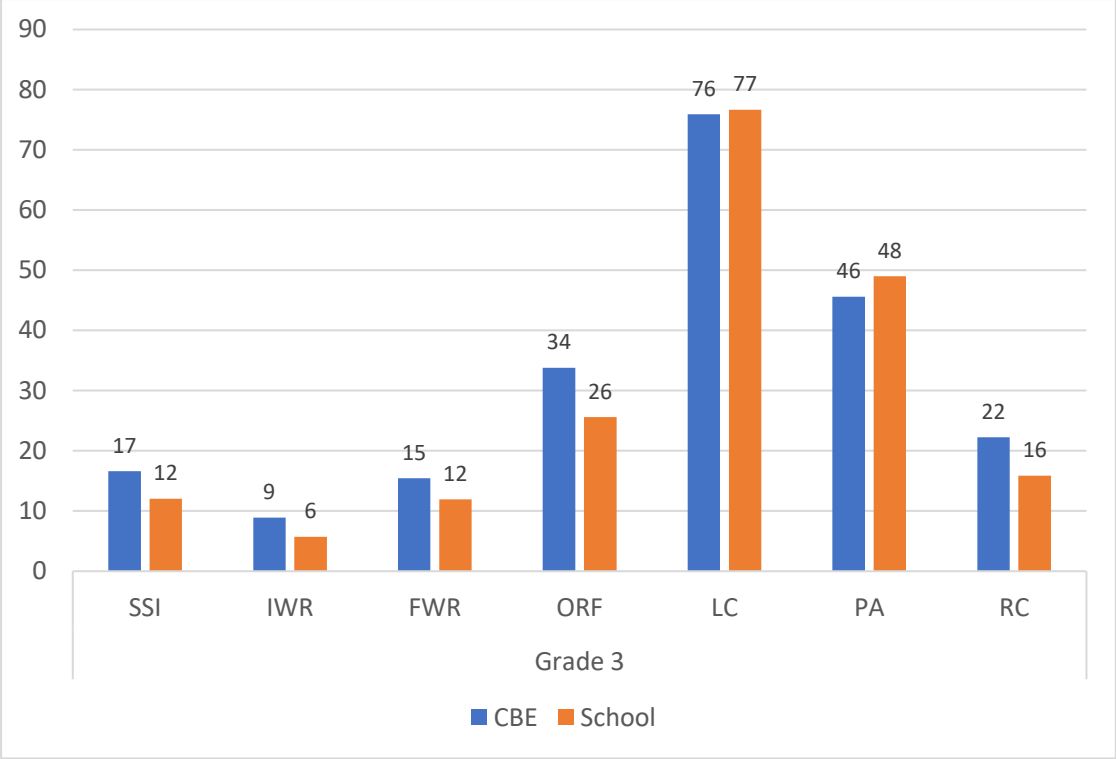
Figure 23: Grade 2 Mean Scores by School Type (Pashto)



In Grade 3, CBE students outperformed Public school students on 5 of the 7 measures (SSI, IWR, FWR, ORF and RC). All of these scores in favor of CBE students were statistically significant. However, it should be noted that due to challenges with sampling, the proportion of students from CBEs in the overall sample at Grade 3 is considerably smaller than the proportion at Grade 2. When the sample size is small, that makes results amenable to fluctuations as a relatively low number of schools with outlier results can dramatically change overall results. For example, if there are only 5 schools in a sample and two of them are outliers (high or low), this can generate results that are not as truly representative of the overall situation. In Grade 3 – and bearing in mind that sample Ns for Grade 3 CBEs are too low to permit valid comparisons – CBE students appear to have improved ORF scores to 34 words-per-minute, some 8 words higher than public school students. A similar level of increase

in favor of CBE students is noticed in Reading Comprehension, but these scores are not to be viewed as significant in any way.

Figure 24: Grade 3 Mean Scores by School Type (Pashto)



CBE classes were established to provide educational resources to children in remote rural areas who are unable to attend formal schools due to insecurity, distance or other constraints. EGRA achievement results for CBEs must therefore be viewed in light of the complex educational context in which they are obtained, results which without CBE facilities would inevitably be significantly worse. Note in Table 18 that urban public schools tend to outperform rural public schools in all the EGRA subscales, however, the differences between rural public schools and CBEs (which are only found in rural contexts) are either negligible or favoring CBEs.

Table 18: Public Schools and CBEs by School Location for Grades 2 and 3 (Pashto)

Subtask	Grade 2						Grade 3					
	Public Schools				CBEs		Public Schools				CBEs	
	Urban		Rural		Rural[7]		Urban		Rural		Rural	
	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N
Syllable Sound Identification	11	867	10	1138	10	236	13	574	11	751	17	20
Invented Word Reading	5	863	4	1129	4	232	6	549	5	734	9	20
Familiar Word Reading	10	878	7	1131	8	234	14	564	10	739	15	20
Oral Reading Fluency	20	895	16	1143	16	236	29	577	23	741	34	20
Listening Comprehension	74	911	73	1164	72	244	78	611	76	785	76	22

[7] Note that CBE classes are not provided in urban contexts.

Subtask	Grade 2						Grade 3					
	Public Schools				CBEs		Public Schools				CBEs	
	Urban		Rural		Rural[7]		Urban		Rural		Rural	
	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N
Phonemic Awareness	47	911	44	1164	42	244	50	611	47	785	46	22
Reading Comprehension	13	911	7	1164	10	244	19	611	13	785	22	22

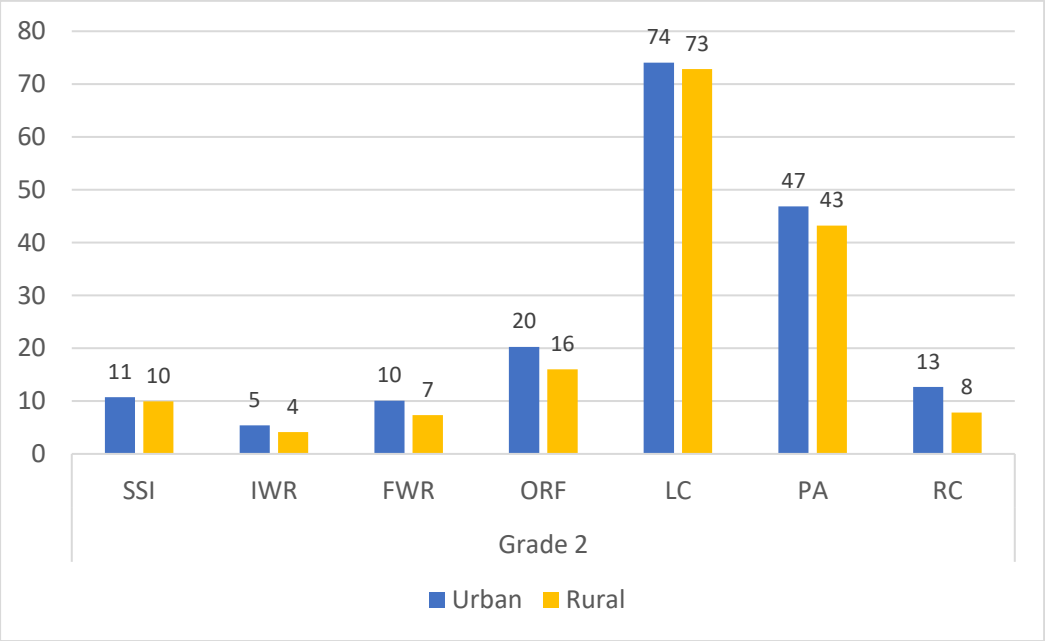
Summary and conclusions regarding scores by type of school and location for grades 2 and 3 for Pashto

- In Grade 2, scores on all reading subtask measures indicate no difference in performance between students from public schools and those from CBEs.
- In Grade 3, superficially, scores in ORF and RC (as well as familiar word and invented word decoding) appear to favor CBE students but valid comparisons are not possible between public schools and CBEs given the low Ns for students in CBEs in this grade.
- When a comparison is made only in rural context, it can be observed that the gap between public schools and CBEs either disappears or that CBEs outperform public schools.

6.5. Subtask Results by School Location

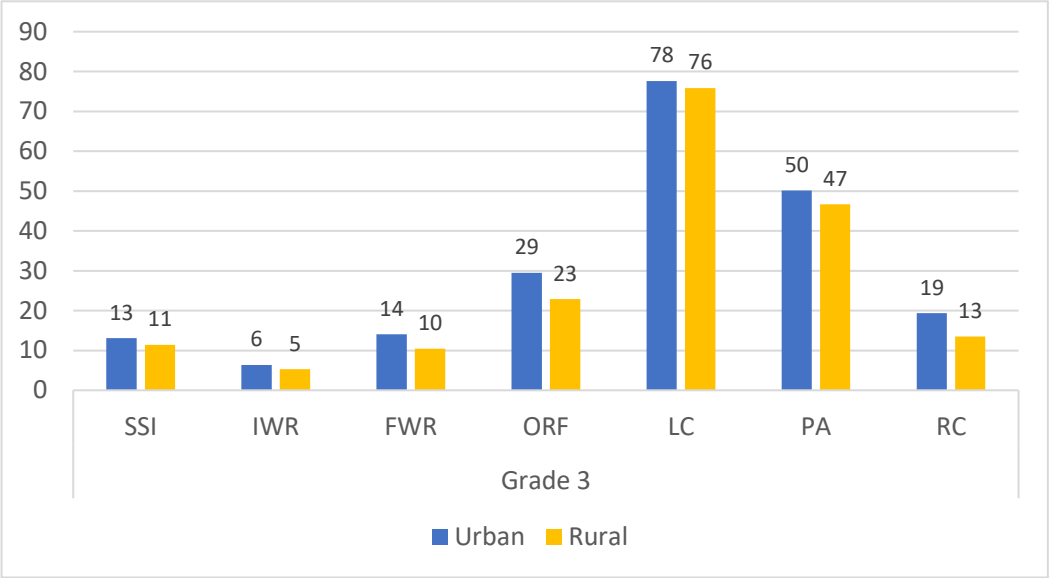
Urban students outperformed rural students in Grade 2 (see Figure 25 below) on all reading subskills measured, two of which were statistically significant (FWR and RC). ORF scores reached 20 words-per-minute for urban students while rural students scored 16 words-per-minute. These differences on ORF in Grade 2 were not statistically significant. In terms of Reading Comprehension, urban students outperformed rural students by 5% (13% vs. 8%) and this difference was statistically significant.

Figure 25: Grade 2 Mean Scores by Location (Pashto)



At Grade 3, urban students continued to outperform their rural peers on all reading subtasks measured, with 2 of the 7 subtasks showing statistically significant differences (as with Grade 2: FWR, and RC). At

Figure 26: Grade 3 Mean Scores by Location (Pashto)



Grade 3, urban students are reading 29 words-per-minute while rural students are reading only 23. In Reading Comprehension, urban students' mean score was 19% compared with rural students' mean score of 13%. As we have seen with overall performance and all sub-group scores, end-of-grade 3 scores do not approach acceptable levels of comprehension, as shown by the low decoding skills.

Summary and conclusions regarding mean scores by location for grades 2 and 3 for Pashto

- In Grade 2, urban students outperformed rural students on all reading measures with statistically significant levels of FWR and RC. As we have seen with other sub-groups, ORF and RC scores were extremely low in the case of both urban and rural students.
- At Grade 3, urban students continue to outperform rural students and there is clear improvement from Grade 2, by some 5% on Reading Comprehension and by some 7-9 words-per-minute on ORF, but levels remain extremely low. By the end of Grade 3, ORF levels need to increase by at least 100% to approach acceptable Reading Comprehension.

7. MULTIPLE REGRESSION OF BACKGROUND FACTORS ASSOCIATED WITH STUDENT PERFORMANCE

Afghan Children Read collected contextual data for the project's baseline Early Grade Reading Assessment (EGRA) study by means of student, teacher and principal questionnaires. Each questionnaire was designed to address the contextual factors potentially associated with learning outcomes of students, particularly reading performance as measured by the Oral Reading Fluency (ORF) task on the EGRA.

The associations between contextual factors and student performance were analyzed by means of two strategies:

1. The Univariate strategy, which focuses on the associations between each contextual variable separately and student performance on EGRA. This approach employs one-way analysis of variance (ANOVA) where responses on questions served as an independent or predictor variable (for classification of respondents into groups) and the ORF task served as a dependent or criterion variable.
2. The Multivariate approach, which brings multiple contextual factors into analysis simultaneously. This approach is based on Multiple Regression (MR) where the responses on questionnaires served as multiple predictor variables, and the ORF task served as a criterion or outcome variable.

Multiple Regression is a statistical procedure that establishes a prediction equation between multiple variables serving as 'predictors' on one side of the equation, and a single criterion variable on the other side. Predictors are variables known to individually predict (correlate with) the criterion, so together they make a more accurate prediction than each variable alone. Note that the label 'predictor' is used tentatively as the MR is not always used to make actual predictions of the criterion, but oftentimes it is used to explore the nature of the relationship between a criterion (also called outcome or dependent variable) and a set of associated contextual variables.

MR relies on several basic assumptions: 1) Linear relationship between the criterion and the predictor variables; 2) Multivariate normality, which assumes that all involved variables are measures of the concepts that are normally distributed in nature; and 3) No multicollinearity, which assumes that the predictor variables are not highly correlated with each other. Although in practice these assumptions may not be always fully met, the MR is commonly used by researchers as it appears to be a valuable tool for explaining the complexity of relationships between multiple contextual variables and a targeted outcome variable. The associations identified by MR bear relevant information useful for making policy and instructional support decisions.

From a technical viewpoint, MR constructs a weighted linear combination of predictors, where the weights are determined in a such a way that yields the maximum possible correlation between that linear combination and the criterion variable. Thus, the major information produced by MR is a set of regression weights indicating the importance of each predictor in explaining the variation of the criterion. The standardized regression weights are comparable by size and they indicate their unique contribution to the prediction that is above and beyond other predictors included in the equation.

In the ACR Baseline study, there were three questionnaires (student, teacher, and principal), each providing many contextual variables that can be used in building multiple regression models. The selection of the questionnaire variables to be included in the regression modelling was based on two principles: 1) Content of the question (selecting those that can be reasonably expected to be associated

with student reading performance; and 2) Univariate analysis of associations between the questions and student performance. It was also necessary to recode some response options to preserve the ordinality of response categories.

'Stepwise' multiple regression modeling was used for the ACR Baseline study. This modeling strategy includes putting the predictors into the equation one at a time and testing the significance of change in the proportion of explained variance of the criterion (R2 change). At every step of the 'stepwise' strategy, predictors that are already included in the model are re-evaluated for significance and if any of them becomes insignificant it would be removed from the model. The 'stepwise' procedure stops when there are no more predictors whose inclusion in the model would make a significant R2 change. In such a way, the final model includes only predictors that demonstrate statistical significance in explaining the variation of the outcome variable.

7.1. Multiple Regressions of Student Contextual Variables (Dari)

The results of MR modeling for student contextual variables are presented in Table 19. There are 26 student questionnaire responses that showed significant associations with student reading performance.

The contextual variables showing the strongest positive associations were found in the categories of **home support** ('Last time had good marks family congratulated and encouraged me' and 'Last time had good marks family gave me a gift'), **student motivation for school** ('Like school: In order to become a good human'), and **absenteeism** ('Absent last week: No, I wasn't absent').

The strongest negative associations were found in the category of **teacher's reactions to student low performance** ('When cannot answer question the teacher beats me'), student negative emotions ('I am afraid fighting' and 'I am afraid of students from other classes') and **receiving help at home** ('Does anyone help you with your homework?'). For the latter variable, although receiving help with homework sounds positive, in this case it is associated with lower student performance, likely because high performing students do not have a need for help with homework.

Students' sound motivation for school (to learn something new and to become a good human), **extra-curricular activities** (such as attending lessons in the mosque), and their **emotional status** (happy/laughed yesterday) demonstrate the relevance of intrinsic motivation and emotional factors for school performance.

It is very informative that multiple teacher practices showed significant associations with student performance: *checking homework*; *awarding students for success* (admiring and giving tokens); *constructive support in case of failure* (asking the question in another way; encouraging to try once more); whereas adverse reactions in case of failure (*beating students*) are negatively associated with student performance.

It is also strikingly evident how home environment has a strong association with student early reading performance; for example: *awarding success* (congratulating, encouraging, hugging, kissing, taking for a sightseeing, and giving a gift); *constructive support in case of failure* (they encouraged me to work hard); and various home resources indicating socio-economic status (*having other books*, *eating before coming to school*, and *having electricity*).

Table 19: Dari Students: Results of multiple regression for background factors

Questionnaire Variables (Predictors)	Standardized Coefficients	Sig.
Like school: Because I learn something new	.040	.034
Like school: In order to become a good human	.123	.000
Do you still attend lessons in the mosque after you are admitted to school?	.051	.006
Does the teacher check your homework notebook (Dari subject)?	.055	.003
When lessons well learned teacher: He admires me	.059	.003
When lessons well learned teacher: He awards me (gives me a gift)	.090	.000
When cannot answer question the teacher: The teacher asks the question in another way to explain in better	.089	.000
When cannot answer question the teacher: The teacher encourages me to try it once more	.043	.037
When cannot answer question the teacher: The teacher beats me	-.071	.000
Absent last week: No, I wasn't absent	.108	.000
Absent last week: I had been to a wedding	.100	.000
Absent last week: Yes! Because the weather was bad	.055	.004
Can your mother read and write?	-.040	.033
Last time had good marks family: They congratulated and encouraged me	.147	.000
Last time had good marks family: They hugged and kissed me	.098	.000
Last time had good marks family: They took me for sightseeing	.048	.009
Last time had good marks family: They gave me a gift	.120	.000
The last time you had not got good marks, did your family/guardian know it?	-.043	.040
Last time had not good marks family: They encouraged me to work hard	.084	.000
Does anyone help you with your homework?	-.063	.001
Besides the school books, do you have other books (story books, magazine, newspaper, etc.) at home for reading?	.078	.000
Have you eaten something before coming to school today?	.044	.016
Do you have electricity at home?	.052	.005
Afraid of anything at home: Yes. I am afraid of students of other classes.	-.042	.026
Afraid of anything at home: Fighting	-.054	.005
Were you happy/have you laughed yesterday?	.038	.039

7.2. Multiple Regressions of Teacher Contextual Variables (Dari)

The results of MR analysis for teacher contextual variables are presented in Table 20. Strong positive associations with student reading performance are found when teachers report high expectations about their students (indicating expectations that *students should write and read simple texts in earlier grades*), and when they report using the following practices: *encouraging students* (writing encouraging remarks

in their notebooks); *getting in touch with family in case of misbehaving students*; *consulting other teachers when facing problems in teaching*.

Other teaching practices that appeared positively associated with student reading performance are: *preparing for lessons by arranging study items*; *using aid materials made by themselves*; *writing in the student's textbook and reading silently as methods used for teaching reading*; and *using story books prepared by school administration*.

Other background factors significantly associated with student performance are **school type** (CBEs performing lower than public schools), **teacher gender** (being male associated with lower performance of students, likely indicating that male teachers predominantly work in remote low-performing rural areas), **being a permanent teacher** (meaning that schools with frequent substitutes perform lower), **teacher age** (indicating that older teachers work with higher performing students), and **receiving training** (in schools where teachers received training on children rights, students demonstrate higher reading performance).

Negative associations with reading performance are observed when teachers report that they *don't use student results* and when they *criticize teacher guidebooks* (e.g., spelling and writing problems).

Table 20: Dari Teachers: Results of multiple regression for background factors

Questionnaire Variables (Predictors)	Standardized Coefficients	Sig.
School type	-.046	.015
Are you a permanent teacher for this class at this school?	.047	.015
Gender	-.081	.000
How old are you?	.063	.001
Training on children rights: Yes	.052	.008
What to do misbehaving students: Getting in touch with the family	.070	.001
What to do misbehaving students: Other treatment	.060	.002
Encourage students: I write encouraging remarks in [his/her] notebook	.072	.000
Consults when facing problems teaching: I arrange a meeting with other teachers	.076	.000
How prepares for new lesson: I arrange study aid items	.044	.039
Items used as teaching aid: I use materials that I make made myself	.051	.012
Methods used teaching reading: Writing the student's book text.	.054	.004
Methods used teaching reading: Reading silently	.068	.001
Teacher guide book problems: Guide [book] has spelling and writing problems	-.048	.017
Who prepared story books: School's administration	.043	.028
Use of students results: I do not use them at all	-.049	.011
What grade students do you expect to be able to write a simple text?	-.091	.000
What grade students do you expect to be able to write a simple passage?	.052	.034

7.3. Multiple Regressions of Principal Contextual Variables (Dari)

The results of MR analysis for principal contextual variables are presented in Table 21. There were 11 contextual variables that remained significant in the multiple regression model. The same with teachers, principal's gender shows a strong association with student performance – being a male is associated with low performance of students, which likely means that male directors are more typically appointed in remote low performing rural areas.

The number of teachers in grades 2-5 is positively associated with student reading performance, meaning that larger schools (likely urban schools) are higher performing schools. Again, similarly as with teachers' expectations, when principals expect that students should be able to read and write in earlier grades, it is associated with higher student performance in their schools.

Other positive associations include principals' *monitoring the examination results of students*; *holding council meetings more frequently*; *council's actions on construction of classes*; *council's following up on absences of teachers*; and *during inspections receiving advice on using different teaching methods*.

Negative associations are found when in the case of teacher absence, a *principal brings a backup teacher* (likely indicating that absenteeism is a problem), and when *actions during inspections point to the use of different methods of student assessment* (likely indicating that student assessment is not present in low performing schools). Interestingly, when the principal reported that *students received training on keeping safe on the way to school*, it was negatively associated with reading performance in their schools, likely meaning that student low performance may be more typical in the areas with safety issues.

Table 21: Dari Principals – Results of multiple regression for background factors

Questionnaire Variables (Predictors)	Standardized Coefficients	Sig.
Being a male	-.114	.000
Average number of teachers in grades 2-5	.096	.000
If teacher is absent: I bring them a backup teacher	-.076	.001
Ensure students learn: I monitor the examination results of students	.075	.001
What grade students do you expect to be able to read a simple text fluently?	-.107	.000
When did you hold the last council meeting?	.084	.000
Council actions on maintenance: construction of the classes	.052	.023
Council actions on administrative: follow up on absences of teachers	.059	.007
Actions during inspections: advice on using different teaching methods	.066	.004
Actions during inspections: advice on different methods of student assessment	-.105	.000
Have the students received any training on how to keep safe on their way to school?	-.050	.023

7.4. Multiple Regressions of Student Contextual Variables (Pashto)

The results of MR modeling for student contextual variables are presented in Table 22. There are 17 student questionnaire responses that showed significant associations with student reading performance.

The strongest positive associations were found in the following categories: **student characteristics** (attending school to become a professional, attending school to have a better future), **home support** (family hugging and kissing students when they get good marks), and **student absenteeism** (not being absent from school).

Teaching practices that showed significant association with reading performance are: *teachers awarding students for success* (admiring the students) and *helping students when they make a mistake* (asking questions in a different way to help students understand, and correcting students' mistakes when they are reading).

Home environment (*having books at home*; *having a television set at home*; *having a notebook to do homework*; *not having to work to help support the family*), *family providing constructive support by hugging and kissing the student when they get good marks* or *encouraging them to work hard when they do not get good marks*, and *student motivation to attend school* (learning something new) are also positively associated with reading performance.

The categories with the strongest negative association with reading performance are: teacher activities (*beating students who cannot answer questions correctly*), home support (*receiving help with homework from a cousin*; *having a mother that can read and write*), and home environment (*helping to financially support the family by doing agricultural work with the father*).

Table 22: Pashto Students: Results of multiple regression for background factors

Questionnaire Variables (Predictors)	Standardized Coefficients	Sig.
Likes coming to school: Because I learn something new	.058	.003
Likes coming to school: In order to have a better future	.069	.000
Likes coming to school: In order to become a doctor, engineer and teacher	.095	.000
Do you have a notebook for doing homework?	.047	.008
When lessons well learned teacher: He admires me	.064	.001
When cannot answer question the teacher: asks the question in another way to explain in better	.064	.000
When cannot answer question the teacher: beats me	-.084	.000
Was absent last week: No, I was not absent	.077	.000
Can your mother read and write?	-.049	.007
Last time had good marks family: They hugged and kissed me	.079	.000
Last time did not have good marks family: They encouraged me to work hard	.066	.000
Who helps with homework: Cousin	-.052	.004
Besides the school books, do you have other books (story books, magazine, newspaper, etc.) at home for reading?	.067	.000
Helps family financially: No	.051	.004
Helps family financially: Yes, I work with my father on the land	-.038	.033
Do you have a television set at home?	.043	.017
Does your teacher help you if you make mistakes during reading?	.039	.028

7.5. Multiple Regressions of Teacher Contextual Variables (Pashto)

The results of MR analysis for teacher contextual variables are presented in Table 23. For teachers, 16 responses showed significant associations with student reading performance.

Strong positive associations with student reading performance are found when teachers report they *choose the teaching profession because they like it*, when they *receive advice from their supervisors on preparing the teaching plans*, and when they report *being a female teacher*, the latter likely indicating that male teachers predominantly work in remote low-performing rural areas.

Additional teacher background factors that show positively significant associations are: *having a degree in education*; *being a permanent teacher at the school or CBE* (meaning that schools with frequent substitutes perform lower) and *having six or more years of experience*. In terms of school environment, *feeling safe at school* is also significantly associated with reading performance.

Teaching practices that appeared positively associated with student reading performance are: *preparing new lessons by consulting the department head* and *using assessment results to change the teaching methodology to serve the needs of the students*. Also, teachers who reported that the *reading material used at school was prepared by the Ministry of Education or a Non-Governmental Agency* are also associated with higher performance of their students.

Negative associations with reading performance are observed when teachers report that they *don't use student results*, that *the department head or teaching deputy does not check teaching plans*, when they report *the teacher guidebook contains inappropriate methods*, when they *write on the board as a method for teaching reading* and *discussing teaching problems in a casual way with other teachers*.

Table 23: Pashto Teachers: Results of multiple regression for background factors

Questionnaire Variables (Predictors)	Standardized Coefficients	Sig.
Sex	.132	.000
What is your field of study?	.050	.004
Are you a permanent teacher for this class at this school?	.044	.014
How long have you been working as a teacher?	.042	.019
Choose to work as teacher: Because I like it	.062	.000
Consults when facing problems teaching: I discuss it with other teachers in a usual manner	-.053	.002
How many times did s/he check your plan during the past one month?	-.040	.026
Advice received from supervisors: Advices regarding the preparation of teaching plans	.059	.001
How prepares for new lesson: I consult the department head	.036	.034
Methods used teaching reading: Writing on the board	-.050	.004
Teacher guide book problems: Teaching methods in guide [book] are not appropriate	-.044	.011
Who prepared story books: Ministry of education	.036	.040
Who prepared story books: Non-Governmental agency	.053	.004
Use of students results: I do not use them at all	-.039	.023
Use of students results: Changing the teaching methodology and complying it with students' needs	.047	.007
Do you feel safe at school?	.045	.009

7.6. Multiple Regressions of Principal Contextual Variables (Pashto)

The results of MR analysis for principal contextual variables are presented in Table 24. There are 14 contextual variables that showed significant associations with reading performance.

The strongest positive associations are found when principals reported they think *students are safe on their way to school* and that they *monitor student learning by reviewing the results of examinations*. The *number of sections in grade 2-5* is also positively associated with student reading performance.

Other variables that show positive associations are the *number of female teachers at school*; *students receiving training on how to keep safe on the way to school*; *distribution of textbooks by the Ministry of Education*; *having discipline regulations at school*; and *school council actively participating in the preparation of the school progress plan*.

Negative associations related to actions by the school council are: *construction of the school surrounding wall* and *monitoring the progress plan of the school*. Additional variables with negative associations are *checking the students' homework to ensure they have learned the lessons* and *average number of teachers in grade 2-5* (likely that larger schools are lower performing). Finally, variables related to actions by school inspectors that show negative associations are: *providing advice on different teaching methods* and the *frequency of school inspections*, where inspecting the school more than three times a year does not improve student performance.

Table 24: Pashto Principals – Results of multiple regression for background factors

Questionnaire Variables (Predictors)	Standardized Coefficients	Sig.
Does your school have education [ministry's] discipline/order regulations?	.045	.026
What is the number of female teachers at your school?	.067	.003
Average number of sections in grades 2-5	.330	.000
Average number of teachers in grades 2-5	-.211	.002
Ensure students learn: I monitor the examination results of students	.075	.000
Ensure students learn: I check the student's homework	-.061	.003
In the beginning of the current academic year, did the Ministry of Education distribute books to all your students?	.049	.017
Council actions on maintenance: construction of the school surrounding wall	-.058	.005
Council actions on administrative: preparation of the school progress plan	.072	.001
Council actions on administrative: monitoring of the school progress plan	-.098	.000
In the past one year, how many times has your school been inspected by members of the (central, provincial and district) education monitoring team?	-.069	.001
Actions during inspections: advice on using different teaching methods	-.049	.013
Are students safe on their way to school?	.092	.000
Have the students received any training on how to keep safe on their way to school?	.051	.010

8. GENERAL CONCLUSIONS AND IMPLICATIONS

The EGRA 2017 baseline results in the *Afghan Children Read* project's targeted provinces of Herat, Nangarhar and Laghman are intended to serve as a means to mediate important policy and pedagogical discussions around the issue of where, when and how to improve the level of reading among early grade students in Afghanistan. In this section we provide a brief summary of the general conclusions from the findings and some key implications and recommendations.

Overall, a close examination of the baseline results indicates that there are some encouraging signs amidst a strong sense of need for rapid improvement. The results relate to Grade 2 and Grade 3 students in the targeted provinces, in Dari and Pashto, two key national languages in Afghanistan. One would hope to see results in Grade 2 that would suggest that students are well on their way to acquiring solid reading skills – and by reading we refer to the acquisition of skills to comprehend key information contained in a grade-level text and on a topic that students are expected to be able to relate to. These expectations should be clearly consolidated by the end of Grade 3, a grade by which it is important for students to be able to read in order to ensure their ongoing involvement and success in the education system. By Grade 4 we expect students to be able to "read to learn", whereas the emphasis in earlier grades has been more on "learning to read". Where students fail to acquire effective reading skills by the end of Grade 3, we typically see students abandoning the school system in the following grades.

The second point to emphasize is the fact that the scores reported in this study relate to a baseline evaluation, i.e., a study of levels of achievement in aspects of early grade reading at the outset of a program of interventions designed to improve reading skills. We thus expect scores to be low in a baseline since they presumably reflect a prior absence of appropriate instructional methodology for reading. They represent the motivation for change and instructional improvement.

Two positive signs should be highlighted that would seem to suggest that, with the application of an effective and sustained reading pedagogy, performance goals can be achieved. The first of these is that in the area of listening comprehension, a fundamental early skill that supports the acquisition of comprehension-based reading skills, scores are solid in the targeted grades: in Grade 2, LC is 61% and in Grade 3 is 68% for Dari, while in Pashto the scores are slightly higher at 73% in Grade 2 and 77% in Grade 3. These scores across languages are only comparable in a general sense, not a statistical sense, since the listening tests are different for each language and clearly based on quite different language and sociocultural contexts. We should conclude from these results that students in both language acquire solid listening skills, and are able to identify key information in grade-appropriate texts that they listen to. We can also conclude that comprehension of text as a general construct does not appear to be a problem in learning for students of Dari and Pashto. They receive solid experiences in listening for information and develop useful skills in this area. We can also conclude that students' lack of success in a parallel skill for reading is more to do with not having developed appropriate decoding skills, key pre-comprehension skills. These skills can be developed, given the right pedagogical approach and given the availability of appropriate materials and resources.

The other positive sign is that we notice a clear trend for improvement of skills across all early grade reading measures. Improvements in scores from Grade 2 to Grade 3 on the key measures of Oral Reading Fluency and Reading Comprehension are the following: ORF in Dari - 32 words-per-minute to 47 words-per-minute (an increase of almost 50%), and in Pashto 18 words-per-minute to 26 words-per-minute (an increase of 44%); on RC in Dari 24% to 35% (close to 50% increase) while RC in Pashto increased from 10% to 16% (a 60% increase). All other measures similarly improved from Grade 2 to Grade 3. This suggests a clear ability to improve and one would assume that, again given the right

approach and support to reading instruction, these increases should be sooner and lead to appropriate levels of gains by the end of Grade 3.

Quite clearly from these baseline results however, scores in Grade 2 foretell the obvious fact that by the end of Grade 3 students in the targeted provinces in Afghanistan in Dari and Pashto are still some way off demonstrating solid reading skills. There is plenty of evidence for this, including the ORF words-per-minute scores and RC percentage scores presented above which clearly show how large the gap is from acceptable levels of comprehension. The percentage of students who scored zero on key early grade reading measures, especially Oral Reading Fluency and Reading Comprehension, is high, especially in RC: 55% in Grade 2 in Dari, falling to 44% in Grade 3; for Pashto, 79% falling to 70%; for ORF, in Dari 25% in Grade 2 falling to 15% in Grade 3 and in Pashto 53% in Grade 2 falling to 43% in Grade 3.

Another piece of evidence reinforcing the low baseline scores in both Dari and Pashto is the percentage of students who achieved appropriate levels of ORF words-per-minute (70 words-per-minute and above) which reached 14% of all students in Grade 2 for Dari rising to 28% in Grade3. For Pashto the corresponding percentages were 8% in Grade 2 rising to 16% in Grade3.

Highlights in terms of subgroup scores are summarized in the following bullet points:

- There are no significant differences in girls' performance versus that of boys in either of the 2 grades measured and in either of the two languages. In some respects, this is good news because resources do not have to be channeled into supporting one group over the other. We can conclude that the educational system in Afghanistan, at least in the targeted provinces, is equitable at least in terms of Sex. In terms of achievement, the observations made above of low scores in both grades although improving from one grade to the next do not approximate acceptable levels by the end of Grade 3.
- It appears that it is not a significant disadvantage to be a home speaker of a language that is not the language of instruction (i.e., the language of assessment). This was evident in the case of both Dari and Pashto as languages of instruction. We must bear in mind that the sample sizes of students in Grade 3 representing languages other than the targeted language of instruction were very low.
- When we compare the performance of public-school students over those students receiving Community Based Education, we found a difference between performance in the Province of Herat over that of the Provinces of Nangarhar and Laghman. For the Dari language, public schools outperformed CBE students in both grades, with differences increasing from Grade 2 to 3. For Pashto, Nangarhar/Laghman public school students did not outperform CBE students in either grades.
- Expectedly urban students consistently outperformed rural students in both grades and in both targeted regions.

In Table 25 below, we provide a comparison of the general conclusions drawn on baseline scores in Herat versus Nangarhar and Laghman. These comparisons should not be used to highlight numeric differences since scores are obtained from different test items in different languages. The scores can indicate trends in the targeted provinces and languages, and, in that sense, we mostly see similar trends in Dari and Pashto.

Table 25: Comparison of Summary and Conclusions for Dari and Pashto

Dari	Pashto
1. Summary and conclusions regarding overall baseline scores	
• Student achievement levels significantly improve from Grade 2 to Grade 3 in all subtasks measured in the baseline; however, overall, these improvements still leave students far short of an acceptable level in Reading Comprehension (of 80% and above). In Grade 2, the mean ORF score is 32 words-per-minute (with Reading Comprehension at 24%), in Grade 3 ORF is 47 words-per-minute (with Reading Comprehension at 35%). • The percentage of students who score zero points on Reading Comprehension (55% in Grade 2 and nearly 40% in Grade 3) is extremely high; the percentage of zero scores on Oral Reading Fluency (the predictor variable for Reading Comprehension) is also high, with a quarter of students in Grade 2 scoring zero, which falls to 15% in Grade 3. • Listening comprehension mean scores are high (at 61% in Grade 2 and 68% in Grade 3). Apart from indicating that oral-based skills develop easily in Dari, these scores also indicate that comprehension of text as a general construct does not appear to be a problem in learning; what inhibits comprehension of reading text is the still under-developed decoding skills as seen in the low ORF, FWR, IWR and SSI scores. • For a baseline, these results are hardly surprising and leave room for significant improvement with the application of appropriate pedagogical strategies contemplated under Afghan Children Read. This will be ascertained in the midline evaluation.	• Despite significant improvement on all subtasks from Grade 2 to Grade 3, mean achievement scores on all subtasks are extremely low with only 8% of students in Grade 2 reaching an acceptable level of Reading Comprehension (of 80% or above) which rises to 16% in Grade 3. • In Grade 2, the mean ORF score is 18 words-per-minute and RC at 10% while in Grade 3 mean ORF score is 26 words-per-minute and RC is 16%; lack of comprehension skill is entirely predictable from the low level of fluency. • In terms of percentage of students scoring zero on subtasks, 53% in Grade 2 and 43% in Grade 3 scored zero on ORF, with 79% in Grade 2 and 70% in Grade 3 scoring zero on RC. • Students in both grades for Pashto scored high on Listening Comprehension (73% and 77% on Grades 2 and 3 respectively) – indicating that comprehension as a construct is not a problem for acquisition, rather students lack sufficient decoding skills to achieve the same levels of comprehension in reading.

Dari	Pashto
2. Summary and conclusions regarding scores by sex	
• In Grade 2, girls outperform boys on most measures (6 of 7) but by differences that are statistically insignificant. By Grade 3 it appears that these differences are even smaller and in the case of 3 of 7 subskills boys slightly outperform girls, but differences in Grade 3 continue to be statistically insignificant. • Zero scores on Reading Comprehension in Grade 2 are high, and significantly higher for boys – 60% of boys scored zero versus 51% of girls. These drop to equal percentages by sex in Grade 3 (38%). By contrast, zero scores on Listening Comprehension are very low, ranging from 5% to 3% by sex across grades, reinforcing the notion that the obstacle to Reading Comprehension is lack of appropriate fluency rates. • By the end of Grade 3 both boys and girls appear to be somewhat distant from developing appropriate levels of Oral Reading Fluency to permit acceptable levels of Reading Comprehension (viz. 48 and 47 words-per-minute on ORF with 35% and 36% RC).	• There is no significant difference between girls' performance and boys' performance on any reading indicator for either grade. • All reading scores improve by some 10% from Grade 2 to 3 although remain significantly low by the end of Grade 3 (Grade 2 ORF scores are 17 words-per-minute (boys) and 19 words-per-minute (girls) improving to 25 and 26 words-per-minute in Grade 3. These ORF scores correctly predict the low comprehension scores that were achieved. • Comprehension scores in Grade 2 are at 8% (boys) and 11% (girls) respectively increasing to 14% and 18% in Grade 3 respectively.
Dari	**Pashto**
3. Summary and conclusions regarding scores by home language	
• These baseline results seem to suggest that from Grade 2 to Grade 3 students who do not speak Dari at home, i.e., they are native speakers of languages other than Dari, significantly improve some of their decoding skills in Dari to the point where they appear to have caught up with or surpassed the level of their Dari peers. Note that the numbers of students who were non-Dari speakers in the home are extremely low in this sample. The issue requires further investigation. Regardless of language background, all students are remote from developing acceptable decoding skills to permit appropriate levels of Reading Comprehension, as we have seen in other sub-group scores.	• While sample sizes for non-native speakers of Pashto in Nangarhar and Laghman are too low to provide valid comparisons, the tentative evidence from these results suggests that non-native speakers of Pashto are not at a significant disadvantage over native Pashto speakers. Note that on Listening Comprehension, non-native speakers of Pashto score reasonably well in both grades (69% and 74% in Grades 2 and 3) although were outperformed by native speakers of Pashto. • In terms of the key variables of interest, decoding skills improve from Grade 2 to Grade 3 for both groups, but remain very low by the end of Grade 3 (ORF: 18 and 17 words-per-minute to 26 and 26 words-per-minute,

• Of those students in Grade 2 who scored zero on Reading Comprehension, 20% were speakers of Pashto (i.e., not the language of the test, Dari, represented by 78% of the total of zero scorers). In Grade 3, a similar pattern exists: 16% of students in Grade 3 who scored zero were Pashto speakers while 81% were Dari speakers. It would appear from this information that being a speaker of Pashto in a Dari-speaking instructional environment may continue to be a problem in Grade 3. Again, given the weak representation of native speakers of languages other than Dari in this baseline sample, the issue would merit further research. • Dari speakers outperform their non-Dari peers by 11% in Listening Comprehension in Grade 2, which is halved by Grade 3 (a difference of 6%).	with Pashto native speakers cited first; in RC, 10% and 8% in Grade 2 to 17% and 15% in Grade 3, Pashto native speakers are cited first).
Dari	**Pashto**
4. Summary and conclusions regarding scores by school type	
• Public schools outperform CBEs with significant differences on most subskills measured. Differences get larger from Grade 2 to Grade 3. Bear in mind that achievement in both Public schools and CBEs remains well below the targeted achievement level in ORF and RC. Also note that the grade 3 sample is small and does not provide a reliable comparison • When a comparison is made of urban public schools, rural public schools, and CBEs (which are only found in rural locations), there is a significant decline with CBEs achieving lowest scores on all subskills measured, rural public schools outperforming CBEs, and public schools outperforming all others. • Note: The Grade 3 sample is too small to provide a reliable comparison.	• In Grade 2, scores on all reading subtask measures indicate no difference in performance between students from public schools and those from CBEs. • In Grade 3, superficially, scores in ORF and RC (as well as familiar word and invented word decoding) appear to favor CBE students but valid comparisons are not possible between public schools and CBEs given the low Ns for students in CBEs in this grade.

Dari	Pashto
5. Summary and conclusions regarding scores by school location	
• Urban schools outperform rural schools on all subskills measured in Grade 2 and 3. Improvements on all subskills are evident from Grade 2 to 3 but insufficient to qualify students as good readers, with Oral Reading Fluency falling well short in Grade 3 of a minimum target of 72 words-per-minute (see Figure 1 above) associated with 80% or more comprehension.	• In Grade 2, urban students outperformed rural students on all reading measures with statistically significant levels of FWR and RC. As we have seen with other sub-groups, ORF and RC scores were extremely low in the case of both urban and rural students. • At Grade 3, urban students continue to outperform rural students and there is clear improvement from Grade 2, by some 5% on Reading Comprehension and by some 7-9 words-per-minute on ORF, but levels remain extremely low. By the end of Grade 3, ORF levels need to increase by at least 100% to approach acceptable Reading Comprehension.

REFERENCES

USAID/Afghanistan. (n.d.). Afghanistan Education. Retrieved from United States Agency for International Development: https://www.usaid.gov/afghanistan/education

Central Intelligence Agency. (n.d.). The World Factbook/Afghanistan. Retrieved from Central Intelligence Agency: https://www.cia.gov/library/publications/the-world-factbook/geos/af.html

Cohen, Jacob. (1988). Statistical power analysis for the Behavioral sciences. SERBIULA (Sistema Librum 2.0). 2nd.

Comeau, Liane; Cormier, Pierre; Grandmaison, Éric; Lacroix, Diane. (1999). A Longitudinal Study of Phonological Processing Skills in Children Learning to Read in a Second Language. Journal of Educational Psychology.

Dubeck, M. M., & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. International Journal of Educational Development, 40, 315–322.

Geva, E., & Genesee, F. (2006). First-Language Oral Proficiency and Second-Language Literacy. In D. August & T. Shanahan (Eds.), Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth (pp. 185-195). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Gough, P. B., Hoover, W. A., & Peterson, C. L. (1996). Some observations on a simple view of reading. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 1-13). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Gove, A. and P. Cvelich. 2011. Early Reading: Igniting Education for All. A report by the Early Grade Learning Community of Practice. Revised Edition. Research Triangle Park, NC: Research Triangle Institute.

Hoover, W., & Gough, B. (1990). The simple view of reading. Reading and Writing: An Interdisciplinary Journal 2, 127–160. The Netherlands: Kluwer Academic Publishers.

Joshi, R. Malatesha; Tao, Sha; Aaron, P. G.; Quiroz, Blanca. (2012). Cognitive Component of Componential Model of Reading Applied to Different Orthographies. Journal of learning disabilities. 45. 480-6.

Kim, J., & Mueller, C. (1978). Introduction to factor analysis: What it is and how to do it. Quantitative Applications in Social Sciences. Sage University Press.

Lervåg, Arne & Grøver Aukrust, Vibeke. (2009). Vocabulary knowledge is a critical determinant of the difference in reading comprehension between first and second language learners. Journal of child psychology and psychiatry, and allied disciplines.

Lesaux, Nonie K., Koda, Keiko; Siegel, Linda; Shanahan, Timothy. (2006). Development of Literacy, Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth (pp. 75-122). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

National Reading Panel. (2000). Teaching Children To Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction. Bethesda, MD: National Reading Panel.

Proctor, C. P., August, D., Carlo, M., & Snow, C. (2005). Native Spanish-speaking children reading in English: Toward a model of comprehension. *Journal of Educational Psychology*, *97*(2), 246-256.

RTI International (2016). Early Grade Reading Assessment (EGRA) Toolkit, Second Edition.

Washington, DC: United States Agency for International Development.

Vaughn, S., & Linan-Thompson, S. (2004). Research-based methods of reading instruction grades K-3. Alexandria, VA: Association for Supervision and Curriculum Development.

Verhoeven, Ludo & van Leeuwe, Jan & Vermeer, Anne. (2011). Vocabulary Growth and Reading Development across the Elementary School Years. Scientific Studies of Reading - SCI STUD READ. 15. 8-25.

APPENDICES

Appendix 1. The EGRA Subtasks

The Phonemic Awareness (PA) subtask determined the ability of students to identify and produce the beginning sound of each of 10 words. The test administrator read each word twice and asked the students to make the first sound of the word. If a learner did not answer within 3 seconds, a "no answer" response was recorded. The maximum score for this section was 10 points, 1 point for each correct answer. This was an untimed activity.

The Syllable Sound Identification (SSI) subtask was designed to measure the ability of students to produce the sounds of the written form of a series of syllables. The test administrator asked the students to make the sound of each of the syllables shown to them. If a learner did not answer within 3 seconds, a "no answer" response was recorded. The SSI tasks were scored on a syllables-per-minute calculation that called for the administrator to determine how many syllables were attempted, how many were read correctly, and in what time over the course of 60 seconds. The SSI task included 100 syllables.

The Invented Word Reading (IWR) subtask assessed the ability of students to decode one- to two-syllable non-words that could plausibly exist in the target language. The IWR task provided a measure of decoding related to that of the Familiar Word Reading task but had the advantage of not allowing respondents to sight-read words, that is, the task provides a measure of a learner's ability to apply the phonological rules of the target language in new unseen contexts. The IWR task 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, and in what time over the course of 60 seconds. The IWR task included 50 words.

The Familiar Word Reading (FWR) subtask assessed a learner's ability to recognize and read high-frequency age and grade appropriate words. This subtask presents a list of unrelated words that are not presented as a story or complete text; the words were randomly arranged on a card that formed part of the learner stimulus. The FWR 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, and in what time over the course of 60 seconds. The FWR task included 50 words.

The Oral Reading Fluency (ORF) subtask can be best understood as a measure of a learner's ability to read with speed and accuracy a passage made up of grade-appropriate words (familiar words) and presented in the learner stimulus booklet. The ORF task is "oral" in that students read the passage aloud. The ORF 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, and in what time over the course of 60 seconds. The ORF subtask included a paragraph with 66 words.

The Reading Comprehension (RC) subtask, which relied on questions about the text read in the Oral Reading Fluency subtask, measured a learner's understanding of the text he or she has just read through a series of factual and inferential questions. The RC subtask included five untimed questions related to the ORF subtask, each with a value of 1 score point for a correct answer.

The Listening Comprehension (LC) subtask measured students' understanding of a short text, read by the test administrator, through a set of factual and inferential questions based on what was read to them. The LC subtask included five untimed questions, each with a value of 1 score point for a correct answer.

Appendix 2. Sampling Method

The following steps guided the random selection of students in each grade:

Stage 1 – Cluster Sampling (CS) for selection of schools

Cluster sampling was applied to the total number of students in each relevant grade (in Herat for Dari and in Nangarhar and Laghman for Pashto) for the public schools. Schools were considered clusters for efficiency purposes. The necessary CS sample size was calculated at a 95% confidence level and 5% confidence interval. The design effect (DEFF) for schools was 3, due to the large and scattered population to get a significant sample size. To avoid loss of sample due to security or other reasons and to keep intact the significance of sample and confidence level, the sample was topped up by 25%.

Stage 2 - Proportionate allocation of sampled clusters to districts

The cluster samples of schools in the province was proportionately allocated to the districts based on the district's proportional representation in the total population. This allowed appropriate representation of each districts in sample and reduce skewing towards a district or set of districts. Schools were selected using a probability proportional technique in each district as per proportionate sample size.

Stage 3 – Probability proportionate sampling for selection of random selection of schools

For the random selection of schools, EMIS data was used. The Grade 2 and Grade 3 enrollment was summed up for each school to get the total population in each school. Probability proportionate to size sampling was used for random selection of the schools using the cumulative school population. The following steps were followed to draw probability proportionate samples:

1. Cleaned the EMIS data and filter all schools where G2 & G3 population doesn't exist in the target districts. Schools with less than 10 G2 or G3 students were also be filtered as 10 students per grade were required for assessment.
2. Added cumulative population column to the filtered schools list.
3. Calculated the sampling interval for each district sampling frame. i.e.

   Sampling Interval (SI) = Population total for the district / sample size of clusters (S)

4. Chose a random starting point (RSP) for the district (using Excel function *rand between*). The RSP was taken between 1 and SI for each district.
5. The first cluster (school; sample point) selected was the one corresponding (greater than or equal) to the row in cumulative population column containing the obtained RSP number
6. Picked the rest of sample clusters from the corresponding rows of cumulative population column to the series calculated as given below:

   For 2nd Sample = cumulative population row corresponding to [*RSP + 1xSI*]

   For 3rd Sample = cumulative population row corresponding to [*RSP + 2xSI*]

   And so on and so forth until Last Sample = cumulative population row corresponding to [*RSP + (S-1) x SI*].

7. Repeated this process for each district.

Stage 4 – Random selection of students in sampled schools (clusters)

Ten students per grade per school was the sample size assessed as for the EGRA. The following steps guided the random selection of students in each grade:

1. Students were randomly selected one class for each grade (Grade 2 & Grade 3) where the number of classes for a grade was more than one. This was done through a draw method. The names of all classes were written on equal size small sheets, folded alike and mixed. Then one sheet was drawn randomly and was the class where students would be assessed.
2. For random selection of students in the selected class, average interval method was used on class attendance register. The total number of students in the attendance registered were divided by 10 to get average interval (*m*, rounded off to the nearest lower whole number). Thus, from the attendance register, every m^{th} student was picked for assessment starting from the 1st student in the attendance list.

Table 26: Final Sample – Dari (Herat)

No.	District	Schools		Proportionate Sample	Students to be Assessed		
		Total	% age		G2 Completers	G3 Completers	Total
1	Hirat	80	13%	19	188	188	376
2	Injil	134	21%	31	315	315	630
3	Guzara	86	14%	20	202	202	404
4	Pashtoon Zarghoon	51	8%	12	120	120	240
5	Karokh	55	9%	13	129	129	258
6	Kuhsan	38	6%	9	89	89	179
7	Ghoreyan	50	8%	12	117	117	235
8	Zenda Jan	29	5%	7	68	68	136
9	Oba	41	7%	10	96	96	193
10	Kushk (Rubatak i Sangi)	66	10%	16	155	155	310
	Total	630		148	1,480	1,480	2,960

Table 27: Final Sample Pashto (Nangarhar/Laghman)

No.	District	Schools		Proportionate Sample	Students to be Assessed		
		Total	% age		G2 Completers	G3 Completers	Total
1	Jalalabad	79	19%	27	274	274	548
2	Surkhrud	55	13%	19	191	191	381
3	Goshta	23	5%	8	80	80	159
4	Kama	34	8%	12	118	118	236
5	Dara i Noor	32	7%	11	111	111	222
6	Koz Kunar (Khiwa)	42	10%	15	146	146	291
7	Behsud	57	13%	20	198	198	395

No.	District	Schools		Proportionate Sample	Students to be Assessed		
		Total	% age		G2 Completers	G3 Completers	Total
8	Mehtarlam Baba	59	14%	20	204	204	409
9	Qarghayee	46	11%	16	159	159	319
	Totals:	427		148	1480	1480	2960

Appendix 3. EGRA Subtask Quality (Item Level Analyses)

AIR carried out basic analyses of baseline data at the item and sub-task level to scrutinize item difficulties and discrimination, particularly to ensure the homogeneous distribution of stimulus items (letters and words) across the rows in student reading sheets. In addition, reliability and factor analyses were performed where appropriate. After the pilot, review committees examined several items from each of the language versions to identify potentially ambiguous items. Subtask items were replaced or revised as necessary, based on the item reviewers' recommendations.

The major criteria for evaluation of EGRA items are:

1. Item Difficulty: For timed sub-tasks it is important that all items are relatively easy and that their difficulty does not differ abruptly. Items are thus distributed across the rows with even difficulty. For untimed tasks it is important to include items that cover a range of difficulties in order to discriminate the performance of students at different levels of ability.
2. Item Discrimination: It is desirable to demonstrate that items can discriminate between higher and lower performing students. This evaluation will be particularly relevant for items in untimed tasks as high item-total correlations contribute to reliability.
3. Differential Item Functioning: This analysis is relevant for verification of test fairness to make sure that the test content does not place any particular group at a disadvantage.

Item difficulty

The computation of item difficulty was conducted using classical item difficulty for untimed tasks and conditional item difficulty for timed tasks where a decreasing number of students reach items at increasing positions in the item list. For example, in the subtask Syllable Sound Identification (SSI), the stimuli are sorted from the 1st to the 100th in 10 rows with 10 syllables each. Because of the timed nature of the subtask, a different number of examinees reach items (stimuli) at different locations in the list. The items that are located furthest from the beginning of the list are reached by a decreasing number of students that have an increasing level of ability (the poorer students don't get to the end of the list because they time out of the activity). Thus, the straightforward measures of item difficulty are not comparable between items at different locations on the list (items at the beginning of the list are taken by examinees of different ability level than the items towards the end of the list, because higher ability students reach items towards the end of the list). We have developed a method for estimation of conditional item difficulty as if every item is taken by all students regardless of their level of ability. In this way, item difficulties are comparable regardless of their position in the list, which is a very useful feature for designing the instrument with rows of homogeneous difficulty to avoid low performing examinees getting stuck with difficult stimuli at the beginning of the list. The tables that follow show the difficulty indices for each item in the EGRA subtasks; they also show the average row difficulty for each subtask where the stimuli are presented in rows. We first present all Herat data followed by tables for Nangarhar/Laghman.

Dari – Herat

Table 28: Item Difficulty Indices for Familiar Word Reading (Dari)

It	N	Diff	It	N	Diff	It	N	Diff	It	N	Diff	It	N	Diff
1	3555	0.47	11	2210	0.803	21	1792	0.322	31	1221	0.446	41	787	0.797
2	3548	0.57	12	2197	0.912	22	1730	0.252	32	1163	0.443	42	770	0.629
3	3545	0.45	13	2183	0.527	23	1662	0.469	33	1122	0.578	43	742	0.376
4	3542	0.44	14	2153	0.451	24	1573	0.355	34	1082	0.648	44	709	0.565
5	3541	0.49	15	2118	0.466	25	1521	0.531	35	1049	0.37	45	675	0.379
6	2251	0.73	16	2070	0.802	26	1437	0.331	36	1003	0.587	46	621	0.463
7	2244	0.70	17	2033	0.361	27	1370	0.461	37	953	0.335	47	576	0.561
8	2239	0.86	18	1984	0.356	28	1331	0.541	38	900	0.325	48	547	0.576
9	2233	0.25	19	1924	0.77	29	1282	0.583	39	846	0.558	49	521	0.586
10	2219	0.87	20	1889	0.463	30	1253	0.6	40	813	0.584	50	494	0.368
Row Diff		0.58	Row Diff		0.591	Row Diff		0.4445	Row Diff		0.487	Row Diff		0.530

Table 29: Item Difficulty Indices for Invented Word Reading (Dari)

It	N	Diff	It	N	Diff	It	N	Diff	It	N	Diff	It	N	Diff
1	3555	0.405	11	1690	0.641	21	1167	0.59	31	551	0.325	41	176	0.277
2	3548	0.45	12	1665	0.252	22	1106	0.406	32	486	0.177	42	154	0.34
3	3547	0.248	13	1627	0.248	23	1025	0.203	33	427	0.139	43	133	0.07
4	3542	0.281	14	1589	0.652	24	966	0.533	34	385	0.293	44	113	0.148
5	3535	0.309	15	1543	0.406	25	926	0.351	35	350	0.273	45	98	0.153
6	1743	0.636	16	1488	0.471	26	840	0.268	36	319	0.281	46	78	0.169
7	1739	0.289	17	1430	0.382	27	769	0.412	37	297	0.12	47	72	0.286
8	1733	0.562	18	1365	0.588	28	722	0.301	38	261	0.181	48	65	0.266
9	1723	0.234	19	1322	0.365	29	671	0.255	39	226	0.308	49	60	0.081
10	1707	0.407	20	1256	0.294	30	604	0.257	40	208	0.188	50	47	0.038
Row Diff		0.382	Row Diff		0.430	Row Diff		0.3576	Row Diff		0.229	Row Diff		0.183

Table 30: Item Difficulty Indices for Oral Reading Fluency (Dari)

It	N	Diff	It	N	Diff	It	N	Diff	It	N	Diff	It	N	Diff
1	3555	0.557	11	2495	0.735	21	2262	0.584	31	1905	0.503	41	1492	0.394
2	3532	0.308	12	2486	0.529	22	2226	0.326	32	1853	0.723	42	1451	0.438
3	3530	0.645	13	2468	0.742	23	2183	0.357	33	1821	0.657	43	1409	0.682
4	3526	0.673	14	2456	0.804	24	2145	0.643	34	1794	0.628	44	1385	0.444
5	3525	0.573	15	2450	0.818	25	2101	0.75	35	1762	0.678	45	1350	0.558
6	3522	0.649	16	2439	0.362	26	2076	0.819	36	1695	0.334	46	1325	0.676
7	3518	0.653	17	2398	0.457	27	2058	0.811	37	1656	0.638	47	1291	0.57
8	3517	0.535	18	2358	0.792	28	2034	0.407	38	1634	0.694	48	1267	0.433
9	3509	0.621	19	2338	0.439	29	1977	0.415	39	1590	0.783	49	1223	0.783
10	3506	0.641	20	2307	0.647	30	1934	0.783	40	1564	0.336	50	1215	0.764
Row Diff		0.586	Row Diff		0.633	Row Diff		0.590	Row Diff		0.597	Row Diff		0.574

It	N	Diff
51	1194	0.591
52	1178	0.267
53	1105	0.381
54	1078	0.777
55	1066	0.78
56	1048	0.726
57	1033	0.824
58	1016	0.813
59	1004	0.697
60	974	0.581
Row Diff		0.644

It	N	Diff
61	940	0.77
62	904	0.812
63	896	0.695
64	881	0.604
65	864	0.795
66	844	0.498
Row Diff		0.696

Table 31: Item Difficulty Indices for Syllable Sound Identification (Dari)

It	N	Diff
1	3555	0.588
2	3552	0.502
3	3552	0.436
4	3551	0.542
5	3550	0.405
6	3550	0.478
7	3549	0.445
8	3546	0.457
9	3545	0.449
10	3539	0.432
Row Diff		0.473

It	N	Diff
11	2397	0.575
12	2384	0.632
13	2369	0.513
14	2355	0.478
15	2335	0.495
16	2316	0.473
17	2291	0.446
18	2257	0.614
19	2225	0.451
20	2196	0.496
Row Diff		0.517

It	N	Diff
21	2147	0.533
22	2098	0.341
23	2063	0.454
24	2039	0.468
25	2006	0.448
26	1961	0.46
27	1921	0.56
28	1870	0.443
29	1827	0.457
30	1776	0.48
Row Diff		0.464

It	N	Diff
31	1718	0.496
32	1660	0.572
33	1625	0.463
34	1589	0.44
35	1546	0.452
36	1502	0.434
37	1459	0.42
38	1407	0.322
39	1344	0.378
40	1294	0.372
Row Diff		0.435

It	N	Diff
41	1235	0.456
42	1173	0.431
43	1132	0.399
44	1098	0.433
45	1063	0.353
46	1012	0.332
47	942	0.368
48	907	0.38
49	867	0.299
50	821	0.353
Row Diff		0.380

It	N	Diff
51	760	0.32
52	717	0.334
53	698	0.295
54	671	0.288
55	640	0.287
56	615	0.295
57	595	0.283
58	559	0.248
59	522	0.188
60	490	0.24
Row Diff		0.278

It	N	Diff
61	451	0.32
62	410	0.337
63	395	0.348
64	372	0.279
65	345	0.261
66	327	0.239
67	306	0.24
68	282	0.222
69	270	0.214
70	257	0.235
Row Diff		0.270

It	N	Diff
71	240	0.263
72	222	0.287
73	206	0.235
74	190	0.218
75	178	0.25
76	156	0.177
77	138	0.162
78	128	0.169
79	117	0.129
80	108	0.151
Row Diff		0.204

It	N	Diff
81	97	0.199
82	90	0.174
83	85	0.142
84	79	0.132
85	75	0.102
86	69	0.104
87	64	0.11
88	54	0.076
89	53	0.09
90	49	0.083
Row Diff		0.121

It	N	Diff
91	43	0.147
92	42	0.227
93	40	0.223
94	38	0.213
95	35	0.125
96	33	0.201
97	31	0.12
98	31	0.12
99	30	0.175
100	29	0.026
Row Diff		0.158

Table 32: Item Difficulty Indices for Listening Comprehension (Dari)

Item	N	Difficulty
1	3488	0.859
2	3488	0.452
3	3488	0.690
4	3488	0.436
5	3488	0.685

Table 33: Item Difficulty Indices for Phonemic Awareness (Dari)

Item	N	Difficulty
1	3488	0.478
2	3488	0.547
3	3488	0.626
4	3488	0.541
5	3488	0.590

Item	N	Difficulty
6	3488	0.625
7	3488	0.580
8	3488	0.552
9	3488	0.585
10	3488	0.555

Table 34: Item Difficulty Indices for Reading Comprehension (Dari)

#	N	Difficulty
1	3460	0.332
2	3249	0.258
3	2706	0.482
4	1887	0.337
5	1811	0.201

Pashto – Nangarhar and Laghman

Table 35: Item Difficulty Indices for Familiar Word Reading (Pashto)

It	N	Diff
1	3738	0.41
2	3728	0.30
3	3725	0.32
4	3723	0.32
5	3722	0.25
6	1728	0.58
7	1706	0.23
8	1703	0.10
9	1699	0.49
10	1690	0.26
Row Diff		0.33

It	N	Diff
11	1674	0.27
12	1668	0.45
13	1650	0.59
14	1631	0.10
15	1595	0.25
16	1559	0.25
17	1503	0.18
18	1450	0.45
19	1404	0.53
20	1369	0.22
Row Diff		0.33

It	N	Diff
21	1300	0.56
22	1234	0.19
23	1177	0.59
24	1119	0.18
25	1047	0.40
26	989	0.07
27	932	0.34
28	890	0.51
29	868	0.39
30	846	0.52
Row Diff		0.38

It	N	Diff
31	815	0.32
32	775	0.22
33	743	0.21
34	712	0.25
35	677	0.07
36	635	0.28
37	612	0.43
38	594	0.18
39	573	0.15
40	548	0.44
Row Diff		0.26

It	N	Diff
41	528	0.04
42	497	0.22
43	469	0.24
44	448	0.35
45	420	0.14
46	394	0.44
47	378	0.27
48	349	0.32
49	328	0.41
50	316	0.08
Row Diff		0.25

Table 36: Item Difficulty Indices for Invented Word Reading (Pashto)

It	N	Diff
1	3738	0.37
2	3727	0.19
3	3723	0.4
4	3723	0.18
5	3723	0.14
6	1623	0.61
7	1595	0.46
8	1594	0.13
9	1591	0.21
10	1587	0.09
Row Diff		0.28

It	N	Diff
11	1574	0.38
12	1551	0.19
13	1518	0.1
14	1480	0.3
15	1427	0.13
16	1355	0.16
17	1291	0.22
18	1233	0.47
19	1179	0.02
20	1111	0.11
Row Diff		0.21

It	N	Diff
21	1010	0.27
22	921	0.13
23	853	0.1
24	775	0.06
25	712	0.25
26	660	0.23
27	612	0.15
28	562	0.02
29	489	0.05
30	442	0.07
Row Diff		0.13

It	N	Diff
31	398	0.06
32	356	0.03
33	308	0.08
34	285	0.08
35	261	0.04
36	228	0.01
37	191	0.05
38	173	0.03
39	157	0.02
40	132	0.01
Row Diff		0.04

It	N	Diff
41	120	0.02
42	106	0.03
43	96	0.05
44	85	0.01
45	75	0.01
46	65	0.01
47	59	0.01
48	46	0.02
49	42	0.01
50	42	0
Row Diff		0.02

Table 37: Item Difficulty Indices for Oral Reading Fluency (Pashto)

It	N	Diff
1	3738	0.44
2	3697	0.37
3	3692	0.32
4	3687	0.37
5	3679	0.38
6	3637	0.37
7	3628	0.42
8	3622	0.4
9	3621	0.3
10	3616	0.23
Row Diff		0.36

It	N	Diff
11	1755	0.55
12	1727	0.25
13	1719	0.63
14	1713	0.15
15	1696	0.27
16	1677	0.53
17	1669	0.44
18	1655	0.27
19	1632	0.49
20	1611	0.57
Row Diff		0.42

It	N	Diff
21	1588	0.63
22	1565	0.47
23	1538	0.28
24	1481	0.24
25	1436	0.3
26	1408	0.32
27	1348	0.55
28	1317	0.28
29	1281	0.51
30	1257	0.49
Row Diff		0.41

It	N	Diff
31	1236	0.65
32	1217	0.33
33	1182	0.08
34	1143	0.64
35	1125	0.58
36	1113	0.61
37	1100	0.38
38	1065	0.51
39	1047	0.62
40	1028	0.09
Row Diff		0.45

It	N	Diff
41	982	0.36
42	953	0.63
43	941	0.57
44	929	0.66
45	921	0.66
46	914	0.66
47	912	0.47
48	897	0.06
49	875	0.26
50	857	0.3
Row Diff		0.46

It	N	Diff
51	821	0.61
52	808	0.67
53	798	0.69
54	796	0.51
55	785	0.64
56	782	0.36
57	764	0.63
58	750	0.61
59	745	0.38
60	736	0.41
Row Diff		0.55

It	N	Diff
61	717	0.44
62	701	0.42
63	671	0.75
64	661	0.78
65	655	0.75
66	651	0.78
67	639	0.54
68	623	0.19
Row Diff		0.58

Table 38: Item Difficulty Indices for Syllable Sound Identification (Pashto)

It	N	Diff
1	3738	0.37
2	3734	0.39
3	3734	0.25
4	3734	0.43
5	3732	0.28
6	3724	0.38
7	3723	0.25
8	3722	0.25
9	3722	0.3
10	3721	0.27
Row Diff		0.32

It	N	Diff
11	2194	0.2
12	2164	0.27
13	2150	0.29
14	2134	0.29
15	2112	0.34
16	2097	0.36
17	2077	0.26
18	2048	0.25
19	2023	0.27
20	1998	0.45
Row Diff		0.30

It	N	Diff
21	1961	0.27
22	1912	0.36
23	1869	0.39
24	1817	0.2
25	1745	0.21
26	1675	0.2
27	1623	0.24
28	1575	0.34
29	1516	0.16
30	1460	0.36
Row Diff		0.27

It	N	Diff
31	1380	0.28
32	1310	0.22
33	1249	0.25
34	1194	0.18
35	1125	0.27
36	1065	0.16
37	1017	0.17
38	982	0.13
39	939	0.15
40	899	0.14
Row Diff		0.20

It	N	Diff
41	853	0.28
42	810	0.14
43	759	0.17
44	720	0.14
45	674	0.11
46	651	0.12
47	614	0.1
48	585	0.13
49	549	0.08
50	515	0.1
Row Diff		0.14

It	N	Diff
51	495	0.15
52	468	0.11
53	446	0.12
54	429	0.08
55	410	0.09
56	388	0.07
57	363	0.12
58	345	0.07
59	334	0.05
60	312	0.06
Row Diff		0.09

It	N	Diff
61	292	0.06
62	274	0.07
63	249	0.04
64	227	0.04
65	213	0.03
66	204	0.02
67	187	0.03
68	174	0.01
69	164	0.03
70	159	0.02
Row Diff		0.04

It	N	Diff
71	149	0.04
72	143	0.03
73	133	0.03
74	130	0.03
75	126	0.02
76	117	0.01
77	113	0.02
78	102	0.02
79	100	0.02
80	99	0.02
Row Diff		0.02

It	N	Diff
81	92	0.03
82	88	0.02
83	84	0.01
84	81	0.01
85	77	0.01
86	72	0.01
87	70	0.01
88	68	0.01
89	66	0
90	65	0.01
Row Diff		0.01

It	N	Diff
91	60	0.04
92	57	0.02
93	54	0.03
94	52	0.01
95	51	0.01
96	51	0.01
97	49	0.02
98	48	0.01
99	48	0.02
100	46	0.01
Row Diff		0.02

Table 39: Item Difficulty Indices for Listening Comprehension (Pashto)

Item	N	Difficulty
1	3647	0.72
2	3647	0.88
3	3647	0.89
4	3647	0.84
5	3647	0.32

Table 40: Item Difficulty Indices for Phonemic Awareness (Pashto)

Item	N	Difficulty
1	3647	0.52
2	3647	0.49
3	3647	0.57
4	3647	0.47
5	3647	0.47

Item	N	Difficulty
6	3647	0.38
7	3647	0.29
8	3647	0.40
9	3647	0.51
10	3647	0.36

Table 41: Item Difficulty Indices for Reading Comprehension (Pashto)

Item	N	Difficulty
1	3500	0.17
2	2632	0.12
3	2477	0.16
4	2399	0.14
5	2399	0.12

Item discrimination

The computation of item discrimination was conducted for untimed subtasks using the Pearson correlation between an item and the total score. This evaluation is relevant for items in untimed tasks as high item-total correlations are indicative of higher internal consistency and contribute to reliability. The computation of item discrimination was thus carried out for the following subtasks: listening comprehension, phonemic awareness, and reading comprehension. The discrimination indices for these subtasks are shown in Table 50.

Dari – Herat

Table 42: Item Discrimination Indices for Untimed Tasks (Dari)

Subtask	Item	N	Discrimination
LC	1	3488	0.267
	2	3488	0.207
	3	3488	0.327
	4	3488	0.271
	5	3488	0.342
PA	1	3488	0.534
	2	3488	0.610
	3	3488	0.642
	4	3488	0.631
	5	3488	0.665
	6	3488	0.683
	7	3488	0.650
	8	3488	0.638
	9	3488	0.677
	10	3488	0.664

Subtask	Item	N	Discrimination
RC	1	3460	0.441
	2	3249	0.347
	3	2706	0.497
	4	1887	0.532
	5	1811	0.421

LC = listening comprehension, PA = phonemic awareness, RC = reading comprehension

Pashto – Nangarhar and Laghman

Table 43: Item Discrimination Indices for Untimed Tasks (Pashto)

Subtask	Item	N	Discrimination
LC	1	3647	0.26
	2	3647	0.27
	3	3647	0.33
	4	3647	0.29
	5	3647	0.34
PA	1	3647	0.60
	2	3647	0.57
	3	3647	0.54
	4	3647	0.58
	5	3647	0.60
	6	3647	0.57
	7	3647	0.51
	8	3647	0.61
	9	3647	0.64
	10	3647	0.58
RC	1	3500	0.39
	2	2632	0.40
	3	2477	0.47
	4	2399	0.45
	5	2399	0.44

LC = listening comprehension, PA = phonemic awareness, RC = reading comprehension

Differential item functioning (DIF)

This analysis is relevant for verification of test fairness to make sure that the test content does not place any particular group at a disadvantage. DIF computation for gender was done by using the Mantel-Haenszel alpha method. Table 52 shows 24 EGRA items that showed DIF for the Province of Herat. As can be observed from the table, all DIF items are level B (moderate DIF), however, most of them favor boys. This may be of concern for the ORF task (passage reading) where 6 items (words) favor boys and just 2 favor girls. It is recommended to scrutinize the content of this passage for possible gender bias.

EGRA items that showed DIF for the Provinces of Nangarhar and Laghman are to be found in Table 53.

Dari – Herat

Table 44: Subtask Items Showing DIF (Dari)

Subtask	Item	DIF level	Favored group	N Boys	N Girls
FW	1	B	Boys	998	1269
	10	B	Girls	977	1242
	11	B	Girls	971	1239
	18	B	Boys	859	1125
	19	B	Girls	838	1086
	21	B	Boys	784	1008
	25	B	Boys	660	861
	32	B	Girls	521	642
	38	B	Boys	401	499
	43	B	Boys	340	402
	44	B	Boys	327	382
IW	7	B	Boys	808	931
	21	B	Boys	550	617
	27	B	Boys	363	406
OR	3	B	Boys	1115	1411
	10	B	Girl	1104	1398
	16	B	Boys	1071	1367
	21	B	Boys	982	1280
	23	B	Boys	952	1231
	28	B	Boys	890	1144
	32	B	Girl	805	1048
	40	B	Boys	667	897
SI	20	B	Boys	1007	1189
PA	9	B	Girls	1605	1883

FW = familiar word reading, IW = invented word reading, OR = oral reading fluency, SI = syllable sound identification, PA = phonemic awareness

Pashto – Nangarhar and Laghman

Table 45: Subtask Items Showing DIF (Pashto)

Subtask	Item	DIF level	Favored group	N Boys	N Girls
FW	35	B	Boys	395	282
	38	B	Boys	340	254
IW	27	B	Girls	380	232
OR	36	B	Boys	625	488
	50	B	Girls	488	369

FW = familiar word reading, IW = invented word reading, OR = oral reading fluency

Appendix 4. Grade 2 Mean Score and Zero Scores

Table 46: Grade 2 Per Minute Scores for Timed Tasks (Dari)

EGRA Subtask	N	Mean	Zero scores (%)
Syllable sound identification	1826	22	26
Invented words reading	1826	10	49
Familiar word reading	1825	19	35
Oral reading fluency	1823	32	26

Table 47: Grade 2 Percent Correct Scores for Untimed Tasks (Dari)

EGRA Subtask	N	Mean	Zero scores (%)
Listening comprehension	1830	61	5
Phonemic awareness	1830	62	17
Reading comprehension	1830	23	55

Table 48: Grade 2 Per Minute Scores for Timed Tasks (Pashto)

EGRA Subtask	N	Mean	Zero scores (%)
Syllable sound identification	2241	10	41
Invented words reading	2224	5	60
Familiar word reading	2243	8	55
Oral reading fluency	2273	18	53

Table 49: Grade 2 Percent Correct Scores for Untimed Tasks (Pashto)

EGRA Subtask	N	Mean	Zero scores (%)
Listening comprehension	2319	73	4
Phonemic awareness	2319	45	22
Reading comprehension	2319	10	79

Appendix 5. Grade 3 Mean Scores and Zero Scores

Table 50: Grade 3 Per Minute Correct Scores for Timed Tasks (Dari)

EGRA Subtask	N	Mean	Zero scores (%)
Syllable sound identification	1724	28	21
Invented words reading	1720	13	37
Familiar word reading	1721	27	20
Oral reading fluency	1724	47	15

Table 51: Grade 3 Percent Correct Score for Untimed Tasks (Dari)

EGRA Subtask	N	Mean	Zero scores (%)
Listening comprehension	1725	68	3
Phonemic awareness	1725	64	15
Reading comprehension	1725	35	38

Table 52: Grade 3 Per Minute Correct Scores for Timed Tasks (Pashto)

EGRA Subtask	N	Mean	Zero scores (%)
Syllable sound identification	1345	12	36
Invented words reading	1303	6	50
Familiar word reading	1323	12	48
Oral reading fluency	1337	26	43

Table 53: Grade 3 Per Minute Correct Scores for Untimed Tasks (Pashto)

EGRA Subtask	N	Mean	Zero scores (%)
Listening comprehension	1418	77	2
Phonemic awareness	1418	48	19
Reading comprehension	1418	16	70

Appendix 6. Results of T-tests Across Grades 2 and 3

Table 54: T-tests Grades 2 and 3 (Dari)

Subtask	Grade 2			Grade 3			T-Test	Cohen's D	
	N	SD	Mean	N	SD	Mean	(2 tail)	D	Size
SSI	1826	23.0	22.39	1724	24.8	27.91	0.00	-0.2	SM
IWR	1826	12.6	9.69	1720	14.1	13.35	0.00	-0.3	SM
FWR	1825	20.9	18.59	1721	22.6	26.74	0.00	-0.4	SM
ORF	1823	33.4	31.96	1724	38.3	47.32	0.00	-0.4	SM
LC	1830	27.2	60.89	1725	26.0	67.82	0.00	-0.3	SM
PA	1830	36.6	61.62	1725	35.5	63.56	0.11	-0.1	
RC	1830	30.4	23.27	1725	34.3	35.14	0.00	-0.4	SM

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Table 55: T-tests Grades 2 and 3 (Pashto)

Subtask	Grade 2			Grade 3			T-Test	Cohen's D	
	N	SD	Mean	N	SD	Mean	(2 tail)	D	Size
SSI	2241	14.1	10.27	1345	14.9	12.10	0.00	-0.1	
IWR	2224	7.3	4.62	1303	7.5	5.77	0.00	-0.2	
FWR	2243	13.1	8.41	1323	16.0	11.99	0.00	-0.3	SM
ORF	2273	28.0	17.70	1337	33.2	25.75	0.00	-0.3	SM
LC	2319	24.3	73.34	1418	21.9	76.64	0.00	-0.1	
PA	2319	34.3	44.67	1418	34.0	48.17	0.00	-0.1	
RC	2319	22.8	9.74	1418	28.8	15.99	0.00	-0.2	SM

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Appendix 7. Results of T-tests by Sex

Table 56: T-tests Grade 2 by Sex (Dari)

Grade 2 Subtasks	Boys			Girls			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	878	21.90	22.71	948	22.85	23.29	0.38	0.0	
IWR	879	9.43	12.58	947	9.94	12.60	0.39	0.0	
FWR	879	17.21	21.00	945	19.86	20.71	0.01	-0.1	
ORF	880	29.14	33.13	942	34.59	33.49	0.00	-0.2	
LC	880	61.28	27.16	950	60.52	27.29	0.55	0.0	
PA	880	59.42	38.11	950	63.67	35.07	0.01	-0.1	
RC	880	20.46	29.8	950	25.88	30.7	0.00	-0.2	

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Table 57: T-tests Grade 3 by Sex (Dari)

Grade 3 Subtasks	Boys			Girls			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	670	27.73	23.81	1053	28.03	25.41	0.81	0.0	
IWR	669	14.51	14.43	1051	12.60	13.91	0.01	0.1	
FWR	669	27.23	22.53	1052	26.43	22.57	0.47	0.0	
ORF	671	47.56	38.87	1053	47.17	37.96	0.84	0.0	
LC	671	68.11	27.26	1054	67.63	25.22	0.71	0.0	
PA	671	59.61	36.45	1054	66.08	34.65	0.00	-0.2	
RC	671	34.54	34.6	1054	35.52	34.1	0.56	0.0	

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Table 58: T-tests Grade 2 by Sex (Pashto)

Grade 2 Subtasks	Boys			Girls			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	1163	10.39	14.4	1078	10.15	13.8	0.69	0.0	
IWR	1133	4.64	7.3	1090	4.59	7.3	0.86	0.0	
FWR	1144	7.95	12.7	1099	8.89	13.6	0.09	-0.1	
ORF	1162	16.93	26.9	1111	18.50	29.1	0.18	-0.1	
LC	1189	73.61	24.0	1130	73.05	24.7	0.58	0.0	
PA	1189	43.75	35.0	1130	45.64	33.5	0.19	-0.1	
RC	1189	8.33	21.2	1130	11.22	24.2	0.00	-0.1	

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Table 59: T-tests Grade 3 by Sex (Pashto)

Grade 3 Subtasks	Boys			Girls			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	714	12.89	15.2	631	11.22	14.5	0.04	0.1	
IWR	669	6.30	7.7	634	5.20	7.3	0.01	0.1	
FWR	687	12.20	15.7	636	11.77	16.2	0.63	0.0	
ORF	698	25.40	32.1	639	26.12	34.5	0.69	0.0	
LC	747	76.15	22.0	671	77.18	21.9	0.38	0.0	
PA	747	48.31	33.8	671	48.01	34.2	0.87	0.0	
RC	747	14.20	25.8	671	17.99	31.7	0.01	-0.1	

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Appendix 8. Results of T-tests by Language of Instruction

Table 60: T-tests Grade 2 by Language of Instruction (Dari)

Grade 2 Subtasks	Dari			Other language			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	1660	22.60	23.21	166	20.34	20.88	0.23	0.1	
IWR	1660	9.55	12.56	166	11.08	12.80	0.14	-0.1	
FWR	1658	18.59	20.96	166	18.53	20.23	0.97	0.0	
ORF	1656	32.08	33.56	166	30.82	32.09	0.64	0.0	
LC	1664	61.85	26.71	166	51.21	30.36	0.00	0.4	SM
PA	1664	61.74	36.68	166	60.41	36.01	0.66	0.0	
RC	1664	23.57	30.5	166	20.27	29.1	0.18	0.1	

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Table 61: T-tests Grade 3 by Language of Instruction (Dari)

Grade 3 Subtasks	Dari			Other			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	1574	27.38	24.84	149	33.52	23.67	0.00	-0.2	SM
IWR	1573	13.19	14.08	148	14.98	14.73	0.14	-0.1	
FWR	1574	26.38	22.27	148	30.51	25.13	0.03	-0.2	
ORF	1575	47.10	38.35	149	49.64	37.91	0.44	-0.1	
LC	1576	68.33	25.58	149	62.40	29.87	0.01	0.2	SM
PA	1576	63.28	35.64	149	66.50	33.91	0.29	-0.1	
RC	1576	35.01	34.3	149	36.55	34.6	0.60	0.0	

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Table 62: T-tests Grade 2 by Language of Instruction (Pashto)

Grade 2 Subtasks	Dari			Other			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	1648	10.38	14.2	593	9.99	13.7	0.57	0.0	
IWR	1640	4.71	7.4	583	4.37	6.9	0.34	0.0	
FWR	1656	8.50	13.2	587	8.17	12.9	0.61	0.0	
ORF	1674	17.84	28.1	599	17.29	27.7	0.68	0.0	
LC	1713	75.01	23.6	606	68.60	25.6	0.00	0.3	SM
PA	1713	46.11	34.3	606	40.57	34.1	0.00	0.2	
RC	1713	10.43	23.2	606	7.77	21.3	0.01	0.1	

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Table 63: T-tests Grade 3 by Language of Instruction (Pashto)

Grade 3 Subtasks	Dari			Other			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	979	11.93	14.5	366	12.57	15.9	0.48	0.0	
IWR	957	5.75	7.5	346	5.81	7.7	0.90	0.0	
FWR	962	11.58	15.7	361	13.11	16.6	0.12	-0.1	
ORF	981	25.63	33.6	356	26.07	32.4	0.83	0.0	
LC	1035	77.87	21.2	383	73.30	23.4	0.00	0.2	SM
PA	1035	49.61	33.5	383	44.26	35.0	0.01	0.2	
RC	1035	16.52	29.3	383	14.58	27.1	0.26	0.1	

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Appendix 9. Results of T-tests by School Type

Table 64: T-tests Grade 2 by School Type (Dari)

Grade 2 Subtasks	Community Based			School			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	72	17.12	23.64	1754	22.61	22.96	0.05	-0.2	SM
IWR	72	6.95	11.92	1754	9.81	12.61	0.06	-0.2	SM
FWR	72	13.49	19.51	1753	18.80	20.92	0.03	-0.3	SM
ORF	72	24.11	31.93	1750	32.29	33.45	0.04	-0.2	SM
LC	72	58.74	30.18	1758	60.97	27.10	0.49	-0.1	
PA	72	51.99	39.35	1758	62.02	36.46	0.02	-0.3	SM
RC	72	17.96	27.6	1758	23.49	30.5	0.13	-0.2	

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Table 65: T-tests Grade 3 by School Type (Dari)

Grade 3 Subtasks	CBE			School			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	68	18.31	27.60	1656	28.31	24.60	0.00	-0.4	SM
IWR	67	9.03	13.97	1653	13.52	14.13	0.01	-0.3	SM
FWR	67	18.69	23.53	1654	27.07	22.46	0.00	-0.4	SM
ORF	67	33.82	41.65	1657	47.87	38.07	0.00	-0.4	SM
LC	68	63.29	30.28	1657	68.00	25.83	0.14	-0.2	
PA	68	45.61	40.09	1657	64.30	35.11	0.00	-0.5	MO
RC	68	24.20	33.5	1657	35.59	34.3	0.01	-0.3	SM

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Table 66: T-tests Grade 2 by School Type (Pashto)

Grade 2 Subtasks	CBE			School			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	236	9.74	15.1	2005	10.34	14.0	0.54	-0.0421	
IWR	232	4.23	7.3	1992	4.66	7.3	0.39	-0.0598	
FWR	234	8.22	14.0	2009	8.43	13.0	0.82	-0.0160	
ORF	236	16.20	27.8	2037	17.87	28.0	0.39	-0.0595	
LC	244	72.17	24.8	2075	73.47	24.3	0.43	-0.0538	
PA	244	41.79	35.8	2075	45.01	34.1	0.17	-0.0936	
RC	244	9.59	23.0	2075	9.76	22.8	0.91	-0.0074	

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Table 67: T-tests Grade 3 by School Type (Pashto)

Grade 3 Subtasks	CBE			School			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	20	16.60	18.0	1325	12.04	14.9	0.17	0.3069	SM
IWR	20	8.93	9.7	1283	5.72	7.5	0.06	0.4281	SM
FWR	20	15.47	18.0	1303	11.94	15.9	0.33	0.2213	SM
ORF	20	33.78	37.1	1317	25.63	33.2	0.28	0.2453	SM
LC	22	75.93	20.8	1396	76.65	21.9	0.88	-0.0326	
PA	22	45.59	39.2	1396	48.21	33.9	0.72	-0.0771	
RC	22	22.26	32.1	1396	15.89	28.7	0.30	0.2215	SM

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Appendix 10. Results of T-tests by School Location

Table 68: T-tests Grade 2 by School Location (Dari)

Grade 2 Subtasks	Community Based			School			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	25.97	22.3	545	20.87	23.2	1281	0.00	0.2	SM
IWR	12.55	13.3	545	8.48	12.1	1281	0.00	0.3	SM
FWR	24.20	22.7	542	16.22	19.6	1283	0.00	0.4	SM
ORF	40.71	36.3	538	28.30	31.4	1285	0.00	0.4	SM
LC	65.13	25.0	545	59.08	28.0	1285	0.00	0.2	SM
PA	67.99	34.1	545	58.92	37.3	1285	0.00	0.2	SM
RC	30.18	32.0	545	20.34	29.2	1285	0.00	0.3	SM

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Table 69: T-tests Grade 3 by School Location (Dari)

Grade 3 Subtasks	Community Based			School			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	31.50	23.6	516	26.38	25.1	1208	0.00	0.2	SM
IWR	16.46	14.6	517	12.01	13.7	1204	0.00	0.3	SM
FWR	32.60	22.7	517	24.23	22.0	1205	0.00	0.4	SM
ORF	56.43	37.7	517	43.42	37.9	1208	0.00	0.3	SM
LC	71.40	24.3	517	66.29	26.6	1208	0.00	0.2	
PA	68.35	33.3	517	61.51	36.2	1208	0.00	0.2	
RC	41.39	32.0	517	32.47	34.9	1208	0.00	0.3	SM

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Table 70: T-tests Grade 2 by School Location (Pashto)

Grade 2 Subtasks	Community Based			School			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	867	10.75	14.0	1375	9.97	14.1	0.21	0.0550	
IWR	863	5.38	7.9	1360	4.14	6.9	0.00	0.1705	
FWR	878	10.06	14.5	1365	7.35	12.0	0.00	0.2077	SM
ORF	895	20.28	29.8	1379	16.02	26.7	0.00	0.1523	
LC	911	74.09	22.6	1408	72.85	25.4	0.23	0.0508	
PA	911	46.88	33.9	1408	43.23	34.5	0.01	0.1064	
RC	911	12.65	26.3	1408	7.86	20.0	0.00	0.2115	SM

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8

Table 71: T-tests Grade 3 by School Location (Pashto)

Grade 3 Subtasks	Community Based			School			T-Test Prob (2tail)	Cohen's D	
	N	Mean	SD	N	Mean	SD		D	Size
SSI	574	13.06	15.8	771	11.39	14.2	0.04	0.1126	
IWR	549	6.37	7.8	754	5.32	7.3	0.01	0.1404	
FWR	564	14.03	17.5	759	10.48	14.5	0.00	0.2243	SM
ORF	577	29.49	35.4	760	22.91	31.2	0.00	0.1989	
LC	611	77.64	20.7	807	75.88	22.8	0.13	0.0805	
PA	611	50.17	34.1	807	46.65	33.8	0.05	0.1036	
RC	611	19.33	31.3	807	13.47	26.5	0.00	0.2048	SM

Significant at .05 level
D is Cohen's d and appears only when the differences are statistically significant.

Cohen's D	**Effect size**
Small (SM)	= 0.2
Moderate (MO)	= 0.5
Large (LA)	= 0.8