



USAID
FROM THE AMERICAN PEOPLE



EARLY GRADE READING ASSESSMENT BASELINE REPORT

FEDERALLY ADMINISTERED TRIBAL AREAS

SEPTEMBER 2014

This publication was produced for review by the United States Agency for International Development by. It was prepared by Management Systems International (MSI) with School-to-School International (STS) under the Monitoring and Evaluation Program (MEP).

EARLY GRADE READING ASSESSMENT BASELINE REPORT FEDERALLY ADMINISTERED TRIBAL AREAS

Contracted under Order No. AID-391-C-13-00005

Monitoring and Evaluation Program (MEP)

DISCLAIMER

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

ACKNOWLEDGEMENTS

We would like to thank the Education team of USAID/Pakistan for their forward planning to be able to collect baseline data before the roll-out of the two important reading programs. Their support and responsiveness under a demanding timeline made this study possible. We would also like to thank the Department of Education of the Federally Administered Tribal Areas for their support of this activity. Finally, this effort would not have been possible without the dedication of our field teams of quality control officers and our local data collection partner, Basic Education for Awareness, Reforms and Empowerment (BEFARe).

CONTENTS

Executive Summary	1
Chapter 1: Introduction	8
Chapter 2: Design and Methodology	10
Chapter 3: Findings and Results	20
Chapter 4: Conclusions and Recommendations	31
Annexes	34
Annex 1: Complete Item Statistics by Grade	35
Annex 2: Box Plots for Phonics and Reading-Rate Fluency Tasks	36
Annex 3: Example of a Reading Fluency Score Threshold Calculation	39
Annex 4: Distribution of Reading Fluency and Comprehension Scores using Fixed Intervals.....	41

List of Tables and Figures

Table 1: Round 3 Timeline (January to May 2014).....	12
Table 2: Sample Schools by Agency, Gender, and Location	13
Table 3: Reliability Estimates	16
Table 4: EGRA Score Ranges and Calculations	18
Table 5: Example of EGRA Percent Correct and Summary Scores	18
Table 6: Example of EGRA Timed Task Scores.....	19
Table 7: Actual Student Sample by Grade and Gender.....	20
Table 8: Task Statistics	21
Table 9: Percent Correct Scores by Grade and Task	19
Table 10: Percent Correct Scores by Grade, Task, and Gender.....	20
Table 11: Baseline Maximum Scores on Fluency (timed) Tasks	21
Table 12: Phonics and Reading-Rate Fluency Task Means by Grade	22
Table 13: Phonics and Reading-Rate Fluency Task Means by Grade and Gender.....	22
Table 14: Summary Scores by Student Age	23
Table 15: Summary Scores by Reading the Quran at Home	23
Table 16: Summary Scores by the Presence of a Library at the School	24
Table 17: Summary Scores by the Presence of Newspapers at Home.....	24
Table 18: Summary Scores by the Presence of Magazines at Home	24
Table 19: Summary Scores by the Presence of Books at Home	25
Table 20: Summary Scores by Children Having Someone Read to Them at Home	25
Table 21: Summary Scores by Children Reading to Someone Else at Home	25
Table 22: Summary Scores by Children Reading Silently at Home	26
Table 23: Summary Scores by Teacher Academic Qualification.....	26
Table 24: Summary Scores by Teacher Professional Qualification	26
Table 25: Summary Scores by Teacher Age	27
Table 26: Summary Scores by Teacher Experience	27
Table 27: Summary Scores by Teacher In-Service Training	27
Table 28: Summary Scores by Head Teacher Academic Qualification.....	28
Table 29: Summary Scores by Head Teacher Professional Qualification.....	28

Table 30: Summary Scores by Head Teacher Experience.....	28
Table 31: Summary Scores by Head Teacher In-Service Training.....	29
Table 32: Summary Scores by Head Teacher Support to Teachers in Reading.....	29
Table 33: Summary Scores by Head Teacher Training in Teaching Reading	29
Table 34: Summary Scores by School Gender	30
Table 35: Summary Scores by PTA/SMC/PTSMC/PTC.....	30
Table 36: Summary Scores by Presence of a School Library	30
Table 37: Summary Scores by Infrastructure (Drinking Water, Electricity, Toilets)	30
Table A1: Complete Item Statistics by Grade.....	35
Table A2: Thresholds for CWPM with 80 Percent Comprehension	39
Table A3: Thresholds for CWPM with Fixed Intervals	40
Table A4: Grade 3 Reading Fluency and Comprehension.....	41
Table A5: Grade 5 Reading Fluency and Comprehension.....	42
Figure 1: Evaluation Design.....	10
Figure 2: Grade 3 Summary Scores.....	18
Figure 3: Grade 5 Summary Scores.....	18
Figure 4: Percent Correct Scores by Grade and Task.....	19
Figure 5: Grade 3 Percent Correct Scores by Task and Gender	20
Figure 6: Grade 5 Percent Correct Scores by Task and Gender	21
Figure A1: Understanding Boxplots	36
Figure A2: Phonics and Reading-Rate Fluency Box Plots for Grade 3	37
Figure A3: Phonics and Reading-Rate Fluency Box Plots for Grade 5	38
Figure A4: Grade 3 Reading Fluency and Comprehension	42
Figure A5: Grade 5 Reading Fluency and Comprehension	43

ACRONYMS

AJK	Azad Jammu and Kashmir
B.A.	Bachelor of Arts
BEFARe	Basic Education for Awareness, Reforms and Empowerment
B.Sc.	Bachelor of Science
C.T.	Certificate of Teaching (Grade 12 plus FA/FSC Certificate)
DOE	Department of Education
EGRA	Early Grade Reading Assessment
F.A.	Intermediate College (Grade 12) Certificate in Arts
FATA	Federally Administered Tribal Areas
F.Sc.	Fellow in Sciences
FR	Frontier Region
GB	Gilgit-Baltistan
ICT	Islamabad Capital Territory
KP	Khyber Pakhtunkhwa
M.A.	Master of Arts
Matric	Secondary School (Grade 10) Certificate (Matriculation)
M.Ed.	Master of Education
MEP	Monitoring and Evaluation Program
M.Sc.	Master of Science
MSI	Management Systems International
MT	Master Trainer
NEAS	National Education Assessment System
NEMIS	National Education Management Information System
NGO	Non-governmental Organization
PRP	Pakistan Reading Project
PTA	Parent Teacher Association
PTC	Parent Teacher Council
P.T.C.	Primary Teaching (Grade 12) Certificate
PTSMC	Parent Teacher School Management Committee
QCO	Quality Control Officer
SMC	School Management Committee
SPSS	Statistical Package for the Social Sciences
SRP	Sindh Reading Project
STS	School-to-School International
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

Overview

In 2013, Management Systems International (MSI) and School-to-School International (STS) conducted a baseline reading assessment for primary school children prior to the launching of two USAID-funded projects: the Pakistan Reading Project (PRP) and the Sindh Reading Program (SRP). PRP is targeting improved reading for 910,000 children in Azad Jammu and Kashmir (AJK), Balochistan, the Federally Administered Tribal Areas (FATA), Gilgit-Baltistan (GB), the Islamabad Capital Territory (ICT), Khyber Pakhtunkhwa (KP), and Sindh, while the SRP is targeting improved reading and mathematics for 750,000 children in Sindh. Targets will be achieved through support for 1) improved policies, laws, and guidelines for teachers and administrators, and 2) improved reading instruction for children in the primary grades.

To measure results from PRP and SRP, a rigorous external evaluation is being conducted. The evaluation will be comprised of data collection and analysis at baseline, midline, and endline. This report covers the baseline assessment in FATA. Due to the volume of data collection from across Pakistan, the baseline EGRA data were collected in three rounds. FATA, along with Balochistan and Punjab, was part of Round 3 of the baseline, which took place in October 2013. Data from Pakistan's other five provinces/areas/territories (hereafter simply referred to as provinces) were collected in Rounds 1 (AJK, GB, and ICT) and 2 (KP and Sindh) in May and September 2013, respectively. In the baseline part of the evaluation, the following activities were carried out for all of the provinces, including FATA: 1) design, 2) sampling, 3) instrumentation, 4) planning, 5) training, 6) implementation, 7) analysis, and 8) reporting.

The external evaluation design, which was developed prior to the baseline assessment, was tailored to the implementation of the PRP and SRP in each province. In most of the provinces, a quasi-experimental design will be used, with two treatment groups: “full treatment” and “light treatment.” The full treatment group will receive two kinds of support, i.e., 1) policy, laws, and guidelines, and 2) improved instruction. The light treatment group will only receive the first kind of support.

In FATA, the design was somewhat different based on a January 2013 meeting between Department of Education (DOE) officials and USAID representatives. Jointly, the officials and representatives decided to implement the full treatment in six agencies (Bajaur, Frontier Region [FR] Dera Ismail Khan, Khyber, Kurram, Mohmand, and Orakzai) and to exclude seven agencies due to security concerns (FR Bannu, FR Kohat, FR Lakki Marwat, FR Peshawar, FR Tan, North Waziristan, and South Waziristan). In the selected six agencies, all schools will receive the PRP full treatment.

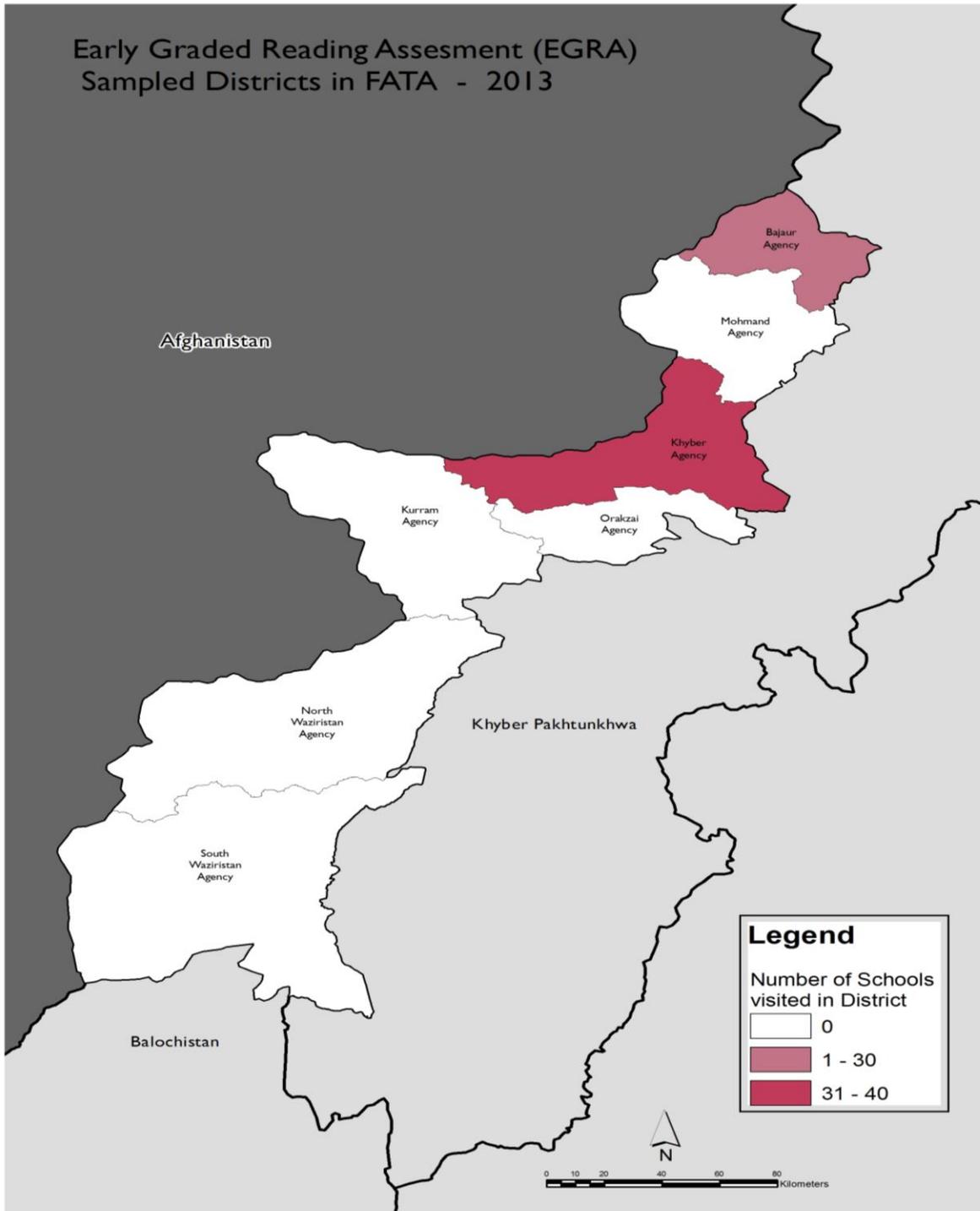
A simple random sample of two agencies (Bajaur and Khyber) was taken from the six full treatment agencies in FATA. A random sample of 70 schools – 35 male schools and 35 female schools – was then taken from the two agencies. In accordance with the USAID evaluation guidelines, students at two selected grade levels – grades 3 and 5 – will be assessed from the sampled agencies and schools at each of the three time points: baseline, midline, and endline. An internationally accepted assessment tool, the Early Grade Reading Assessment (EGRA), will be individually administered to a target sample of over 33,000 children in 1,120 schools throughout the country. Over the course of the evaluation, the evaluators will compare the baseline results with those at the midline and endline to examine success by the PRP and SRP in improving children's reading in Pakistan. The sampling was designed so that each province could be evaluated independently.

The long-term goal of this evaluation is to compare each province's baseline results to its midline and endline results, rather than other province's results. There are too many confounding variables – languages, curricula, administration dates, etc., that could render province-to-province comparisons meaningless. Furthermore, the

evaluation is designed to investigate reading performance of the full and light treatment groups across time: baseline, midline, and endline. The differences between treatments will be fully investigated later, given the baseline data as the starting point for comparisons. In-depth comparisons between the full and light treatment groups are not useful at this time; such comparisons at baseline could add some bias by facilitating competition between the two groups that could compromise the validity of the evaluation.

For the baseline in FATA, all activities were completed by the end of December, including a draft report. The results were presented and discussed at a consultative meeting in Islamabad in April 2014. Representatives from the DOE, USAID, PRP, the contractors (MSI and STS), and other donor and non-governmental organization (NGO) representatives attended the consultation. Revisions were then made to this report based on the discussions between the stakeholders.

Map of Sampled Agencies



Key Points

Several key points from the EGRA baseline assessment in FATA are highlighted below:

Implementation

1. All agencies in FATA that were selected during the initial consultative meetings between the DOE and USAID in February 2013 will receive “full treatment.” In FATA, there will not be a comparison of groups to determine the effects of the full treatment above and beyond those of the “light treatment”; rather, the results at the baseline will be compared against those in the midline and endline for the full treatment agencies only.
2. The baseline data were collected in a simple random sample of two agencies in FATA: Bajaur and Khyber. Within these agencies, a random sample of male and female schools was selected, followed by a random sample of grades 3 and 5 students within those schools. The results from this representative sample are presented below as a generalized view of the reading levels for children in the six target agencies in FATA.
3. A total of 70 schools from the full treatment group were selected for the baseline.
4. The EGRA tools, which have been administered in various forms in over 40 countries, were successfully adapted for use in Pakistan. These included individually administered reading tests for students, along with questionnaires for students, teachers, and head teachers. The Urdu versions of the tools, which were piloted in Mansehra (KP), Muzaffarabad (AJK), and Islamabad (ICT) prior to finalization, were used in FATA.
5. The results from this sample are presented in this report as a generalized view of the reading levels for students in the schools of the districts in FATA receiving full treatment. Please note that agency-level comparisons are not possible because the agencies were not evenly sampled; the number of sampled schools varied by district, and the sample sizes are limited for each district.
6. The EGRA testing window for FATA was October 2013. All schools were reached within this period of time.
7. The assessment tools were successfully administered in FATA’s two sample agencies (Bajaur and Khyber) as follows (with the percentage of the target reached in parentheses): in 70 schools (100.0 percent), to 1,985 students (94.5 percent), 106 teachers (75.7 percent), and 70 head teachers (100.0 percent). The percentage of participating teachers was relatively low; some teachers were teaching students in both grades while the remaining teachers were not available to participate in the study. There were 53 grade 3 and 53 grade 5 teachers who responded to the survey.
8. The validity and reliability of the tools was acceptable. Validity was assured through the adaptation process, which involved 17 educationists from throughout the country who participated in a workshop in Islamabad, and the piloting process. Reliability was assured through the high quality of the assessment tasks and the standardized administration of the tools. Reliability estimates (of internal consistency) were calculated using the coefficient alpha.
9. The data entry and data cleaning process followed international standards. All student data were entered twice into two separate databases. These databases were then compared. All data were reconciled across the two databases and with the assessment booklets. A clean data file was produced for analysis.
10. In the analysis phase, scores were calculated in three ways: 1) percentage correct scores for the reading tasks, 2) average percentage correct (grand means) for reading summary scores, and 3) adjusted raw scores for the timed reading tasks. These scores provide a comprehensive picture of

student performance. Analysis of student, teacher, head teacher, and school characteristics was carried out using the summary scores.

Analysis

1. EGRA was administered to 1,022 grade 3 students and 963 grade 5 students. The reliability estimates were acceptable for both grades ($\alpha = 0.87$ for grade 3 and 0.85 for grade 5), indicating that the items worked well in measuring reading constructs at each grade level.
2. In FATA, the subtest average task p-values at grade 3 ranged from 0.06 to 0.63 (See Table 5). Orientation to print fell about mid-range on the difficulty spectrum (0.63), while listening and passage comprehension were at the low end, 0.11 and 0.06, respectively. The other task p-values were in the low-moderate range (0.11 to 0.46). Grade 5 showed a similar pattern, although the scores were higher. Orientation to print fell at the moderate-high range of the difficulty scale while passage comprehension (0.20) and listening comprehension (0.32) were at the low end. The other subtests' difficulty p-values were in the moderate range (0.43 to 0.65).
3. Grade 3 students did relatively well in orientation to print, letter name recognition, familiar word reading, and passage reading, though their scores were under 50 percent in most of those areas. They had particularly low skills in passage comprehension, listening comprehension, letter sound knowledge, and non-word reading. Grade 5 students had the highest skills in orientation to print, familiar word reading, passage reading, and letter name recognition. In contrast, the scores for listening and passage comprehension were low.
4. Increases in scores from grade 3 to 5 are interesting to analyze because they indicate areas of low and high growth. Grade 5 students showed strong increases in familiar word reading, non-word reading passage reading, and listening comprehension. Phonemic awareness, letter sound knowledge, and passage comprehension were the areas of least improvement, and therefore there is much room for improvement. In areas where there are large differences, interventions at grade 3 could have particularly large effects in accelerating children's learning.
5. Gender comparisons revealed that boys performed better on nearly all of the reading tasks at both grade levels. At grade 3, boys had significantly higher scores ($p < 0.01$) in all areas except the two tasks with the lowest difficulty – orientation to print and letter name recognition. At grade 5, boys had significantly higher scores ($p < 0.05$) in all areas except for letter sound knowledge and passage comprehension. The differences in summary scores were 7.4 points and 6.3 points higher for the boys at grades 3 and 5, respectively.
6. Students were timed on five tasks as they read words or passages. These tasks were categorized into phonics (letter name recognition, letter sound knowledge, and non-word reading) and reading-rate fluency (familiar word and passage reading). Students in both grades had lower phonics scores than reading-rate fluency scores. For the timed tasks, familiar word and passage reading showed the most progression over the two grade levels. The lowest scores were in the area of non-word reading for both grades. The improvement from grade 3 to 5 was all statistically significant ($p < 0.001$) for all tasks. By gender, there were significant differences in favor of boys in grade 3 on all timed tasks. The largest differences were in familiar word reading and passage reading. Similarly for grade 5, boys had statistically higher scores in all tasks except letter sound knowledge. These timed task scores showed the same tendencies as the non-timed tasks; scores were generally higher for boys.
7. Student questionnaires revealed three positive findings. First, students reporting that they read the Quran and other books was related to higher reading scores for both grades. Second, higher reading outcomes were reported by students reading in the home, whether that meant being read to, reading to someone else, or reading silently. Third, attending school at an older-than-normal age seemed to have a positive effect on reading outcomes at grade 3, but that advantage waned and was not significant by grade 5.

8. Questionnaire findings for the teachers and head teachers were mostly inconclusive, due to small sample sizes and the lack of variation in the responses. The teacher and head teacher questionnaire results showed no conclusive patterns in terms of age, experience, education, certification, or training.
9. The school surveys revealed that, the presence of a library, a Parent Teacher Association (PTA)/School Management Committee (SMC)/Parent Teacher School Management Committee (PTSMC)/Parent Teacher Council (PTC), or better infrastructures were not related to higher reading scores. Only 6 percent of the schools reported having libraries, yet almost 60 percent had a PTA-like organization. Although infrastructure was not related to reading scores, gender and school size were associated. Male schools had higher scores than female schools and there was a moderate positive correlation ($r = 0.48$) between school size and EGRA scores (i.e., larger schools tended to have higher scores).

Evaluation Recommendations

Given the success of the baseline assessment in FATA (and in the other provinces), the methods used in 2013 should be repeated as much as possible for the midline and endline assessments in future years. This should be conducted as follows:

1. The EGRA instruments proved to be of high quality, and equivalent versions of those tools should be developed – through trans-adaptation, piloting, and revision – for the midline and endline assessments so that progress can be accurately measured over time.
2. The EGRA items and tasks had good discrimination (quality) values and covered the low-to-middle part of the difficulty range. At baseline, the reading scores were relatively low for both grades and show room for growth. In addition, histograms and box plots provided evidence that the tool is expected to measure higher levels of reading-rate fluency that are anticipated following project-led interventions. Therefore, the baseline data indicates that the EGRA is appropriate for measuring increases in reading ability at midline and endline.
3. The sampling was reasonable in terms of finding a balance between the resources available, the required sample size, and the geographic coverage. It should be maintained in the midline and endline, i.e., keep the same agencies and schools, along with the sampling methods at the school level. With the cross-sectional design, the children will be different, but the same sampling procedures to select students at the school level should be used.
4. The systems for field data collection should be replicated, with the same systems for recruitment and training for the master trainers (MTs), field supervisors, quality control officers (QCOs), and enumerators as used in the baseline.
5. The data entry system should continue to be used, with the same systems for recruitment and training of data entry supervisors and operators, along with implementation through networked computers, double data entry, and reconciliation of errors.
6. The analysis should follow the same procedures, with calculations of task scores, summary scores, and timed task scores. The baseline, midline, and endline scores should be comparable so that improvements in children’s reading can be accurately examined.
7. Reading proficiency levels should be created to provide educators and other stakeholders with meaningful results. Most parents and educators better understand reading achievement in useful terms or levels, such as emerging, proficient, or advanced, rather than interpreting a percent-correct test score that may differ by test or reading passage difficulty. Education officials are encouraged to select specific EGRA scores to serve as levels of reading proficiency for both grades. Percent correct

for each task, summary score, as well as fluency rates are recommended for this purpose. The baseline EGRA data can be used for establishing these reading proficiency levels.

8. Finally, it may be advisable to add items to the student, teacher, and head teacher questionnaires for collecting data on PRP- and SRP-supported interventions so that student scores can be correlated with these indicators.

CHAPTER I: INTRODUCTION

The Pakistan Reading Project (PRP) and the Sindh Reading Program (SRP) are two five-year initiatives funded by USAID. The projects/programs will cover over 40,000 government schools in Pakistan's eight provinces/areas/territories (hereafter referred to as provinces). PRP is targeting improved reading for 910,000 children in AJK, Balochistan, FATA, GB, ICT, KP, and Sindh, while the SRP is targeting improved reading and mathematics for 750,000 children in Sindh. Targets will be achieved through support for 1) improved policies, laws, and guidelines for teachers and educational administrators, and 2) improved reading instruction for children in primary grades. Some districts (or agencies) in Pakistan will receive both kinds of support, i.e., "full treatment," while others will receive only the policy support, i.e., "light treatment." All the schools within each individual district (or agency) will receive the same type of treatment.

To measure results from PRP and SRP, a rigorous external evaluation is being conducted. The evaluation baseline is taking place in 2013, prior to the launch of the reading interventions. In accordance with USAID program evaluation guidelines, samples of children in two selected grade levels – grade 3 and grade 5 – are being assessed throughout Pakistan so independent baselines can be established in each province. Children at the same grade levels will be assessed at the midline and endline time points to evaluate the success of the interventions, taking into account the two treatment groups.

This report covers FATA. Along with Balochistan and Punjab, FATA was part of the Round 3 baseline data collection in October 2013; data from Pakistan's other five provinces were collected in May 2013 (ICT, AJK, GB) and September 2013 (Sindh, KP). The following activities were planned for all of the provinces, including FATA:

1. Design – USAID required a cross-sectional design, i.e., assessing students at the same grade levels (grades 3 and 5) over the course of PRP and SRP. In most provinces (not FATA – see below), this was complemented by a quasi-experimental design with the two treatment groups.
2. Sampling – Schools were selected from the full treatment agencies. The sample enabled the collection of student reading assessment data that were representative of the treatment groups, grade levels, gender, and urban/rural zones.
3. Instrumentation – EGRA tools were developed, with tests at the grade 3 reading level in English, Sindhi, and Urdu, and questionnaires for teachers, head teachers, and students. Model EGRA instruments were trans-adapted, piloted, revised, and finalized for use in Pakistan.
4. Planning – A field administration plan was developed for the baseline administration that would ensure the reliability of the data collected. The plan specified the timeline, training, logistics, field activities, supervision, data entry, analysis, reporting, and quality control.
5. Training – Workshops were conducted to train all MTs, supervisors, enumerators, and QCOs. Enumerators and supervisors were observed to ensure they had clear comprehension and the skills adequate to implement the EGRA tools.
6. Implementation – The baseline survey was implemented according to the plan. It ensured that all of the field activities took place in a standardized manner, as verified by the QCOs. The fieldwork was followed by data entry and preparation of a clean data file.
7. Analysis – Data were analyzed using two different software packages (Excel and SPSS). Experienced statisticians/psychometricians conducted the analysis, produced data tables and graphs, and ensured quality control.

8. Reporting – Provincial-level reports were produced. A reporting template was developed according to guidelines from the USAID contract. These reports will be disseminated to the provincial education authorities.

This report is organized into four chapters: 1) introduction, 2) methodology, 3) findings and results, and 4) conclusions and recommendations. Annexes with item statistics, box plots for the timed tasks, and a possible process for establishing a reading proficiency threshold follow the chapters.

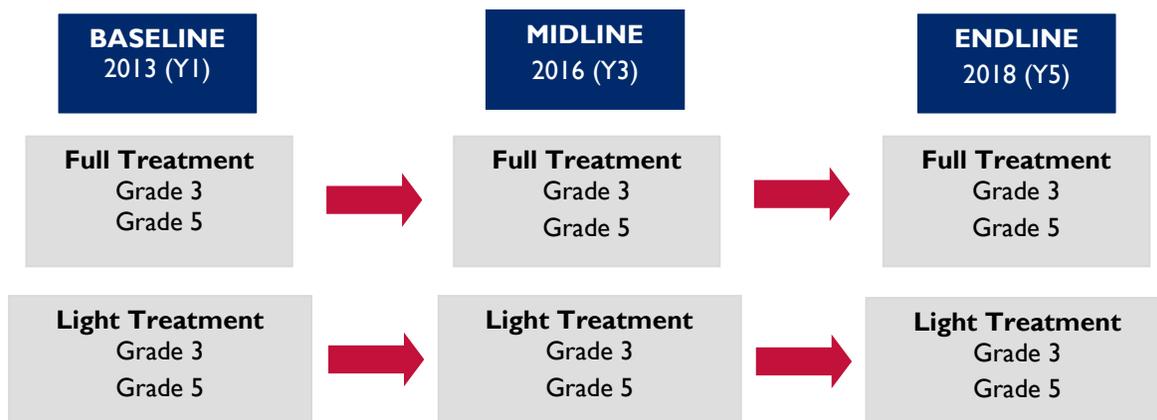
CHAPTER 2: DESIGN AND METHODOLOGY

This chapter presents the evaluation design and methodology, including the methods and systems used for collecting the EGRA baseline data. There are sections on the evaluation design, timeline, sampling, instrument development, data collection, data entry, and data analysis.

Evaluation Design

Following USAID policy, a cross-sectional evaluation design was developed prior to the baseline data collection. As shown in Figure 1, the design features two grade levels (3 and 5) and three time points (baseline, midline, and endline). Different groups of grade 3 and grade 5 students will be compared against each other across the three time points. In the figure, the years for the midline and endline are approximate and may be altered in accordance with implementation of the PRP and SRP interventions.

FIGURE 1: EVALUATION DESIGN



Districts (or agencies) for the “full” treatment group were pre-selected by the provincial DOE and USAID during consultations in January and February 2013. In addition, while most provinces have the “full” and “light” treatment groups, FATA – like AJK and ICT – will receive full treatment across six agencies. The remaining seven agencies in FATA will not receive PRP reading interventions due to security concerns. With this design, there will be no counterfactual (i.e., light treatment) for the reading interventions.

In FATA, students were tested in Urdu, their main language of instruction. Equal numbers of male and female schools, i.e., 35 male and 35 female schools, were sampled for the EGRA testing. The sampling design met the USAID requirements of adequate sample size and equal gender representation (see the sampling section below).

Timeline

The FATA baseline, like the other provinces, was conducted according to a timeline that started in January 2013 and continued through May 2014 with submissions of reports to USAID. The reports may then be distributed to the DOE and other stakeholders as appropriate (see the timeline in Table 1).

The assessment process began with the planning and design of activities, including preliminary sampling designs, selection of model EGRA tasks, recruitment of staff, and budgeting/contracting. This was followed in February by provincial consultations, including those for FATA. From February to April, the EGRA team, with participation from FATA and other provinces, prepared, piloted, and revised the EGRA tools and conducted the agency/school sampling. The data collection in FATA took place in October 2013, followed by the data entry, analysis, and reporting in October and November. A presentation of results for FATA to the DOE and USAID was done in April 2014. The final report for FATA was submitted in May 2014 (see Table 1 below).

TABLE 1: ROUND 3 TIMELINE (JANUARY TO MAY 2014)

Activity	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
Plan and design EGRA activities	X	X															
Participate in provincial consultations	X	X															
Prepare EGRA tools		X	X														
Prepare test administration manuals			X														
Train master trainers and enumerators									X								
Select and verify sample schools								X									
Administer EGRA										X							
Enter data										X	X						
Analyze baseline data											X						
Produce draft reports												X					
Produce presentations														X	X		
Make presentations																X	
Revise and finalize reports																X	X
Submit reports to USAID																	X

Sampling

The sampling for Round 3 started in January 2013 with the selection of the treatment agencies by the provincial DOEs and USAID. The EGRA team conducted the district and school sampling for Round 3, including FATA, in June and July. This included developing the sampling requirements, verifying the sample in the field, and finalizing the sample. The findings were provided in the sampling report for USAID.¹ As mentioned above, in FATA, six agencies were selected for full treatment. The sampling for FATA, as detailed in the sampling report, is briefly summarized in the following sub-sections of this report.

Sampling Requirements

The sampling frame for FATA included 13 agencies. During the consultation process, the DOE and USAID decided to intervene in six agencies, and to exclude seven agencies due to security concerns (FR Bannu, FR Kohat, FR Lakki Marwat, FR Peshawar, FR Tan, North Waziristan, and South Waziristan). In the selected six agencies (Bajaur, FR Dera Ismail Khan, Khyber, Kurram, Mohmand, and Orakzai), all schools will receive the PRP full treatment. A simple random sample of two agencies (Bajaur and Khyber) was taken.

In addition, since the minimum requirement was 15 students per grade level in grades 3 and 5, only schools meeting that requirement were eligible for sampling. Within the selected agencies and the eligible schools, equal numbers of male and female schools (35 each) were selected.

Sampling Process and Field Verification

Due to the need to balance representativeness, logistical demands, and resource availability (e.g., personnel, transportation, funding), one-third of the sampled population agencies in FATA were chosen using a simple random sample. This resulted in a clustered sample. For the 35 male and 35 female schools (70 total), the samples were divided among the selected agencies according to the proportions of schools within those agencies (stratified random sampling). The sampled agencies had zero schools classified as urban in the National Education Management Information System (NEMIS), a second stratification was not necessary at the “location” level. After sampling the 70 schools in FATA, an additional 10 male and 10 female schools were selected as replacements, if needed. Note that mixed schools may have been selected for some replacement schools due to not having enough options for replacement schools of strictly one gender. However, only students from the respective genders were included in those samples (i.e. if a mixed school was selected to replace a female school, only females were sampled).

The number of schools in the agencies and the apportioned number of samples from each agency by gender and location is shown in Table 2 below.

TABLE 2: SAMPLE SCHOOLS BY AGENCY, GENDER, AND LOCATION

Agency	Location	Schools	Pct.	Sample Schools		Replacement Schools	
				Boys	Girls	Boys	Girls
Bajaur	Rural	30	43%	15	15	4	4
Khyber	Rural	40	57%	20	20	6	6
Total		70	100%	35	35	10	10

¹ MSI (2013). *Pakistan EGRA Sampling Report*. 18 June 2013 (Revised).

Once the schools were sampled, the QCOs, supplemented by EGRA senior managers, verified the samples in the field. This step was necessary due to two factors: 1) some inaccuracies in the NEMIS data, and 2) changes in student numbers since the time period when the schools had submitted their data to NEMIS. If the original schools had fewer than 15 students in either grade 3 or 5, a replacement school was selected and verified. At times, schools were retained if their numbers were near the minimum.

Intended and Actual Samples

After conducting the field verification, 22 schools were replaced. For Bajaur Agency, eight schools – two boys and six girls schools – were all replaced due to lower than expected numbers of students in the original samples. In Khyber Agency, 14 schools – six boys and eight girls schools – were replaced. Four boys schools were inaccessible due to the law and order situation, one had no grade 5 class, and one had an inaccessible teacher and the Assistant Agency Education Officer did not have any information regarding that school. Four Khyber girls schools were replaced due to low enrollment, three due to the law and order situation, and in one school, the head mistress did not provide permission for the data collection due to fears that the school may be targeted by extremists (as had happened previously). The actual numbers of students, teachers, and head teachers in the survey are presented in the results section.

Instrument Development

A brief summary of the instrument development process is presented below. The full results from the trans-adaptation, which involved educationists from FATA, were presented in a report to USAID.² This report is available to provincial education officials.

Trans-adaptation

In February, the EGRA team used tasks from recent EGRA administrations in other countries to develop a model test. Led by two international and two national assessment specialists, the EGRA team then organized a trans-adaptation workshop in Islamabad. A total of 17 English, Sindhi, and Urdu language specialists from the DOEs and teacher training institutes throughout Pakistan – including one subject specialist from FATA – participated in the workshop.

The trans-adaptation process involved the following, with the local experts:

1. Discuss and choose reading tasks that would be of value to the baseline assessment in Pakistan;
2. Adapt each reading task using appropriate content in English, Urdu, and Sindhi; and
3. Ensure that the content would be suitable for grades 3 and 5 students.

The workshop resulted in a pilot EGRA test and pilot student, teacher, and head teacher questionnaires. The head teacher questionnaires included items about school characteristics.

Piloting

In March, the EGRA English and Urdu tools were piloted in selected schools in AJK, ICT, and KP provinces. The four tools included in the pilot were 1) a student response booklet (including the student questionnaire), 2) a student stimuli booklet, 3) a teacher questionnaire, and 4) a head teacher questionnaire. The EGRA team conducted the pilot sampling, trained the enumerators, arranged the logistics, and supervised the piloting. The team then entered the pilot data into a database, analyzed the data, and developed preliminary recommendations for final tools in preparation for the revision workshop. They also

² MSI (2013) *Pakistan EGRA Tools Trans-Adaptation Workshop Report*. June (Revised).

prepared a piloting report for USAID.³ As with the piloting report, the tools are available to provincial officials, though they must be kept secure since similar tasks will be used in the midline and endline.

Revision and Finalization

The EGRA team held a revision workshop in March 2013 for the Urdu and English tools with a limited number of experts from the trans-adaptation workshop. The Sindhi tools were revised in July with Sindhi language experts. Changes were made to the instruments based on the pilot data and field observations. These changes were summarized in the piloting report. The team then finalized the four instruments for each language and submitted them to USAID. USAID made suggestions, particularly around the inclusion of reading- and library-related items in the questionnaires that would provide information for the PRP and SRP. The instruments were approved and then used in the training workshops in advance of Round 1 data collection in May. The final instruments consisted of the following:

- Students: 16 informational items, 8 tasks (one with 2 sub-tasks), and 34 questionnaire items
- Teachers: 15 informational items and 52 questionnaire items
- Head teachers: 17 informational items and 37 questionnaire items

These instruments are available for use by education officials.

Data Collection

Subcontractor Selection

The EGRA team, with the participation of USAID, issued a request for proposals and followed a set of criteria to select local subcontractors for the field data collection and for data entry. In August, the Basic Education for Awareness, Reforms and Empowerment (BEFARe) was chosen for both activities (data collection and data entry). MSI, STS, and BEFARe collaborated on the data collection in FATA.

Data Collection

In September, EGRA senior managers trained MTs and QCOs during a two-week session in Islamabad. The MTs also spent one week in Islamabad, training the BEFARe FATA data collection team, which was comprised of a regional coordinator, two field supervisors, and 32 enumerators. The FATA team was trained alongside the teams from the other Round 3 provinces, i.e., Balochistan and Punjab. An EGRA senior manager and two QCOs were assigned to FATA to oversee and provide support for the BEFARe team. The QCOs, coordinator, supervisors, and enumerators organized the logistics for the data collection. Following the training and logistical preparations in Islamabad, one MT, the QCOs, and field supervisors conducted a three-day refresher course for the enumerators in Peshawar just prior to commencing data collection in the schools.

Over a 10-day period in October 2013, the enumerators spent a day in each of the 70 schools to collect the baseline data in FATA. The enumerators received frequent visits and mobile phone calls from the EGRA senior manager, QCOs, coordinator, and field supervisors to check on the status of data collection and to troubleshoot any issues. After collecting the data from the schools, the enumerators submitted their booklets to the supervisors and QCOs for verification and feedback. The supervisors then brought the booklets back to Islamabad for data entry.

³ MSI (2013). *Pakistan EGRA Instrument Development and Pilot Data Analysis*. August (Updated).

Data Entry

Data Entry

In May 2013, the EGRA team developed a customized data entry application so that 1) the exact data from the booklets and questionnaires could be entered into a database, and 2) the computers used for data entry could be networked with a server. In September, the team trained the BEFARe data coordinator, supervisors, and data entry operators on the application. In October and November, the EGRA and BEFARe teams completed data entry for over 23,000 student booklets, along with the questionnaires for the students, teachers, and head teachers (Rounds 2 and 3). This included approximately 2,000 booklets for FATA.

Data Cleaning

In November, the EGRA and BEFARe teams conducted the data verification and reconciliation. Following USAID requirements, 100 percent of the data were entered twice (double data entry) and any discrepancies between the first and second databases were reconciled. A clean data file was then provided to the data analysis team.

Data Analysis

Methodology

In June 2013, the EGRA statisticians and psychometrician developed a research plan that included the following steps: 1) reliability estimates, 2) task and item statistics, 3) mean and grand mean scores (percent correct scores), 4) data plots, 5) timed and untimed task scores, and 6) questionnaire results. They used both SPSS and Excel for the analysis. Some of the analyses were replicated to ensure that the calculations were accurate. Descriptive analyses and inferential statistical comparisons were conducted for the student scores by grade level and gender, and for the three sets of questionnaire data.

Please note that the analyses were only performed at the provincial level. This is because the sampling was conducted at the provincial level, i.e., the sample is only accurate at the provincial level. The samples at the district or school level are too small for analysis purposes, and any results at those levels would be misleading.

Validity and Reliability

Validity evidence for the tests was derived from previous experiences with EGRA in other developing countries, as well as through the trans-adaptation process in Pakistan. The test developers targeted grade 3 for the level of the tasks. An assumption was that the grade 5 students should perform better than the grade 3 students on each of the tasks.

For reliability, a generally accepted method is to estimate the internal consistency reliability (Coefficient Alpha) of the test. The minimum reliability threshold is approximately 0.75 to 0.80 for tests of this nature. Reliability was estimated for each province and language. Table 3 shows the reliability estimates for grades 3 and 5. These reliabilities are excellent and lend credibility to the internal consistency of the tests, indicating that the items are generally measuring similar reading constructs for both grade levels.

TABLE 3: RELIABILITY ESTIMATES

Language	Grade Level	Tasks	N-count	Alpha
Urdu	Grade 3	9	1,022	0.87
	Grade 5	9	963	0.85

Note that there were actually eight tasks, but one of the tasks (Task 7) was administered and scored in two parts, so the equivalent of nine tasks were used for the analysis.

Score Calculation

The EGRA data were analyzed three ways. First, p-values and item-total correlations were generated for assessing the difficulty and discrimination of the items and tasks. Second, the percent correct for each task provided an indication of the FATA students' mastery of the tasks, and third, FATA students' fluency was assessed.

Item P-values and Item-Total Correlations

P-values and item-total correlations are classical test theory statistics that are used to evaluate the performance of individual items and the tasks they comprise. Item difficulty is measured by p-values, which range from 0.00 to 1.00. Higher p-values indicate easier items, because a higher percentage of students posted correct responses. The other classical statistic is the item-total correlation, and it ranges from -1.00 to +1.00. This statistic measures how close the item or task relates to the overall percent correct on the summary score. Values above 0.2 are an indication of a good item or task.

Percent Correct

The results of the EGRA testing were calculated using task and summary scores. Table 4 lists the tasks, stimuli, raw score ranges, and the method for calculating the task and summary scores on the test. For each of the tasks, the stimuli (items) (i.e., questions, letters, sounds, words, and non-words) were worth one score point. The score points were added, and since the range of raw scores varies across the tasks, the percent of correct scores was used to report all results. No weighting was used with the tasks to calculate the summary scores. Each task summary score was calculated using the total number correct and dividing it by the number of items. The overall Reading Summary Score was calculated by adding all of the task summary scores and dividing by nine (total number of tasks) to arrive at the average.

Timed Tasks Scores

The scores on the timed tasks were calculated by taking the number of correct responses times 60 seconds then dividing that number by the number of seconds used to read the stimulus. For instance, if a student read 75 letters correctly in 30 seconds, their letters-correct-per-minute score would be 150 (75 words x 60 seconds/30 seconds). Given another example, if a student read 50 words correctly in 30 seconds, his or her timed task score would be 100 words per minute (50 words x 60 seconds/30 seconds). Table 4 lists the number of stimuli per task. Recall the percent correct scores ranged from zero to 100. The method for calculating phonics and fluency scores yielded much higher maximum values, upwards of 200 at baseline (see the task box plots in Annex 2, Figures A1 and A2).

TABLE 4: EGRA SCORE RANGES AND CALCULATIONS

Task (Subtest)	Stimuli	Score Range	Calculation
1. Orientation to print	5 questions (untimed)	0-5	Percent correct of answers
2. Letter name recognition	100 letters (timed)	0-100	Percent correct of letters
3. Phonemic awareness	10 questions (untimed)	0-10	Percent correct of words
4. Letter sound knowledge	100 sounds (timed)	0-100	Percent correct of sounds
5. Familiar word reading	50 words (timed)	0-50	Percent correct of words
6. Non-word reading	50 non-words (timed)	0-50	Percent correct of non-words
7a. Passage reading	60 words (timed)	0-60	Percent correct of words
7b. Passage comprehension	5 questions (untimed)	0-5	Percent correct of answers
8. Listening comprehension	3 questions (untimed)	0-3	Percent correct of answers
Reading Summary Score	-	-	Average of percent correct

An example of percent correct scores for each of the tasks and as a summary score is provided below. The raw score is divided by the maximum score (the highest score possible in the score range) to produce the percent correct score for each task. Then, the task scores are averaged to produce the summary score. Note that each of the task percent correct scores is weighted equally to provide the summary score.

TABLE 5: EXAMPLE OF EGRA PERCENT CORRECT AND SUMMARY SCORES

Task (Subtest)	Maximum Score	Raw Score	% Correct Score
1. Orientation to print	5	3	60.0%
2. Letter name recognition	100	68	68.0%
3. Phonemic awareness	10	5	50.0%
4. Letter sound knowledge	100	42	42.0%
5. Familiar word reading	50	34	68.0%
6. Non-word reading	50	25	50.0%
7a. Passage reading	60	50	83.3%
7b. Passage comprehension	5	2	40.0%
8. Listening comprehension	3	1	33.3%
Reading Summary Score	--	--	55.0%

An example of timed task scores (adjusted) is provided below for the five fluency tasks. The formula explained above is used (timed task score = raw score x 60 seconds/seconds used).

TABLE 6: EXAMPLE OF EGRA TIMED TASK SCORES

Task (Subtest)	Raw Score	Seconds Used	Timed Task Score
2. Letter name recognition	68	48	85.0
4. Letter sound knowledge	42	60	42.0
5. Familiar word reading	34	48	42.5
6. Non-word reading	25	40	37.5
7a. Passage reading	50	40	75.0

CHAPTER 3: FINDINGS AND RESULTS

This chapter presents the findings and results from the EGRA baseline in FATA. There are sections on the sample, task and item statistics, score calculation, task and summary scores, timed task scores, and questionnaire findings.

Student Sample

The intended sample was 70 full treatment schools. Within these schools, the target was to assess 15 students in each grade and gender per school; totaling 2,100 students; 1,050 for each gender, treatment, and grade. However, on the day of assessment, some students were not in attendance. Table 7 shows the number of students in the sample by grade and gender. For grades 3 and 5, the actual samples were 97.3 and 91.7 percent of the intended sample, respectively. The boys' percent (97.7) was higher than the girls' percent (90.8). A small number of students in grade 3 ($n = 1$) and grade 5 ($n = 5$) did not complete the gender item on the questionnaire. The total actual sample with EGRA scores in FATA was 1,985 students, 94.5 percent of the intended 2,100 sample records. A few schools were kept in the sample even though, during the field verification, their actual numbers were below the target. The main reasons, however, for the difference between the intended and actual samples was low student attendance and/or the law and order situations on the survey date. Still, the percentage of the target was high.

TABLE 7: ACTUAL STUDENT SAMPLE BY GRADE AND GENDER

Grade Level	Sample	Boys	Girls	Missing	Total
Grade 3	Students	518	503	1	1,022
	% of Target	98.7%	95.8%	--	97.3%
Grade 5	Students	508	450	5	963
	% of Target	96.8%	85.7%	--	91.7%
Total	Students	1,026	953	6	1,985
	% of Target	97.7%	90.8%	--	94.5%

Task and Item Statistics

Table 8 shows the statistics for the tasks on the test. Two classical statistics are provided: p-values and item-total correlations. P-values indicate the average score of the students on the tasks, or the difficulty of the tasks for the students. The item-total correlations in the table are actually task-total correlations, which indicate the degree to which the tasks can discriminate between low- and high-achieving students; this is an indicator of the quality of the items. P-values can range from 0.00 to 1.00, with higher values indicating easier items. Item-total correlations can range from -1.00 to +1.00, with values above +0.20 or +0.25 indicating that the item (or task) is of high quality.

In FATA, the subtests or task p-values for grade 3 ranged from 0.06 to 0.63. The p-value for orientation to print (0.63) was much higher than the other tasks, indicating a relatively easy item. In contrast, passage comprehension was relatively difficult (0.06). The remaining p-values were between 0.11 and 0.46. Thus, with the exception of orientation to print, the EGRA at grade 3 provided a spread on the lower half of the difficulty spectrum. The scores for grade 5 were higher and showed a similar pattern. The p-values for grade 5 ranged from 0.20 to 0.72, which is across the low-to-middle part of the spectrum. Such large variations in p-

values are helpful in terms of measuring pre- to post-test gains in student performance. These data suggest that the item difficulty ranges were suitable for the baseline testing and have “room to grow” for accommodating students’ potential increases in reading ability when measured at midline and endline. The item-total correlations for grades 3 and 5 were at or above 0.26 and 0.30 respectively, indicating that the tasks were of good quality.

TABLE 8: TASK STATISTICS

Task (Subtest)	Grade 3		Grade 5	
	P-Value	Item-Total	P-Value	Item-Total
1. Orientation to print (untimed)	0.63	0.26	0.72	0.30
2. Letter name recognition (timed)	0.46	0.65	0.61	0.60
3. Phonemic awareness (untimed)	0.35	0.43	0.43	0.39
4. Letter sound knowledge (timed)	0.29	0.64	0.43	0.42
5. Familiar word reading (timed)	0.42	0.86	0.69	0.79
6. Non-word reading (timed)	0.32	0.85	0.57	0.78
7a. Passage reading (timed)	0.42	0.86	0.66	0.79
7b. Passage comprehension (untimed)	0.06	0.51	0.20	0.56
8. Listening comprehension (timed)	0.11	0.48	0.32	0.49

Task and Summary Scores

The next part of the analysis involves plotting the scores. Histograms of the summary scores (Figures 2 and 3) show that the distributions are moving to the right from grade 3 to grade 5, which is strong evidence that the children are learning basic skills at the primary school level. As with the task and item statistics, it also shows that there is room for growth at each grade level, particularly at grade 3. The main goal of the intervention is to see movement of the distributions to the right within the same grade level (i.e., grades 3 and 5) from the baseline to midline to endline.

FIGURE 2: GRADE 3 SUMMARY SCORES

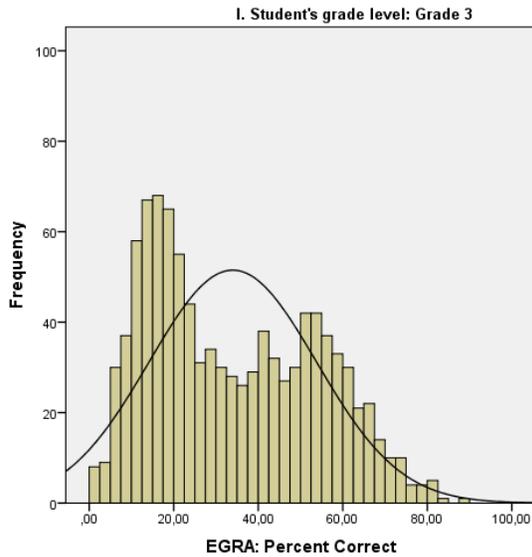


FIGURE 3: GRADE 5 SUMMARY SCORES

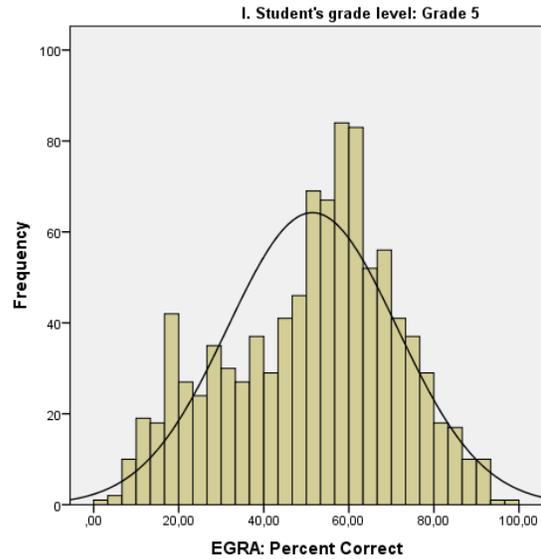


Table 9 and Figure 4 provide the average scores by task using percent correct scores. The score for each task was calculated using the total number correct and dividing by the number of items. For instance, a student who scored 3 out of 5 on Task 1 would receive a score of 60 percent. Averages were then calculated for all students on Task 1, which in FATA was 62.8 percent for grade 3 and 71.5 percent for grade 5. The same type of calculation was made for each student and each task. The table also includes the differences from grade 3 to grade 5, e.g., 71.5 percent minus 62.8 percent equals 8.7 percentage points.

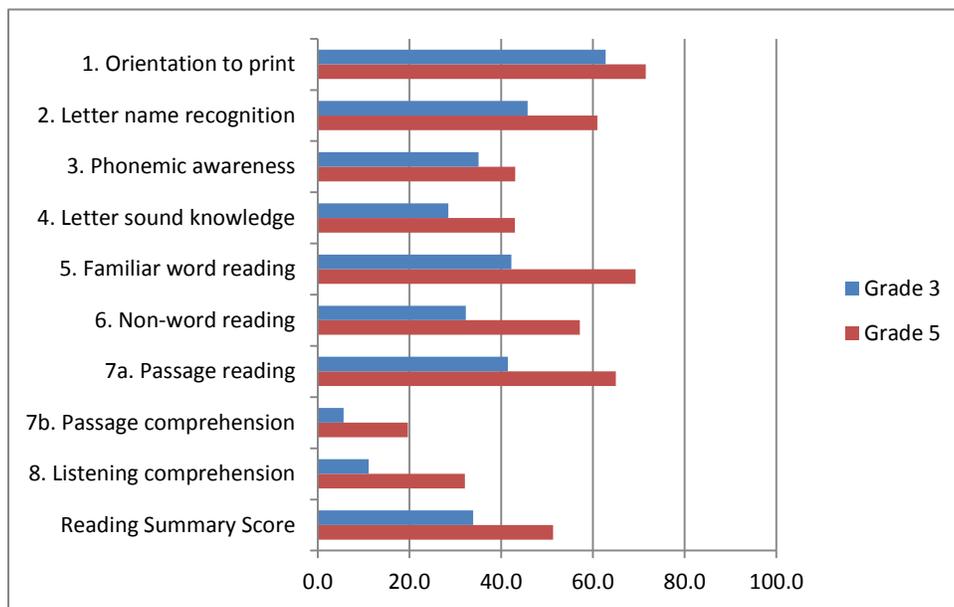
Grade 3 students did relatively well in orientation to print, letter name recognition, familiar word reading, and passage reading, though their scores were under 50 percent in most of those areas. They had particularly low skills in passage comprehension, listening comprehension, and letter sound knowledge. Grade 5 students had the highest skills in orientation to print, passage reading, familiar word reading, and letter name recognition. In contrast the scores for passage comprehension and listening comprehension were low.

The increases in scores from grade 3 to 5 are interesting to analyze because it provides an indication of areas of low and high growth. Grade 5 students showed strong increases in familiar word reading, non-word reading, passage reading, and listening comprehension. Phonemic awareness, letter sound knowledge, and passage comprehension are the three areas where the differences from grade 3 to grade 5 were the smallest and there is much room for improvement. In areas where there are large differences, interventions at grade 3 could have particularly large effects in accelerating children's learning.

TABLE 9: PERCENT CORRECT SCORES BY GRADE AND TASK

Task (Subtest)	Grade 3	Grade 5	Difference
1. Orientation to print	62.8%	71.5%	8.7% points
2. Letter name recognition	45.8%	61.0%	15.2% points
3. Phonemic awareness	35.1%	43.0%	8.0% points
4. Letter sound knowledge	28.5%	43.0%	14.4% points
5. Familiar word reading	42.2%	69.3%	27.1% points
6. Non-word reading	32.3%	57.1%	24.8% points
7a. Passage reading	42.5%	66.1%	23.9% points
7b. Passage comprehension	5.7%	19.7%	14.0% points
8. Listening comprehension	11.1%	32.1%	21.0% points
Reading Summary Score	34.0%	51.4%	17.4% points

FIGURE 4: PERCENT CORRECT SCORES BY GRADE AND TASK



When the scores were disaggregated by gender (Table 10 and Figures 5 and 6), most of the differences between boys and girls were statistically significant ($p < 0.01$ level) favoring boys. Statistically, boys in grade 3 scored significantly higher in all areas except the two tasks with the lowest difficulty - orientation to print and letter name recognition. The girls had a slightly higher score for orientation to print, and the two-point difference in letter name recognition was not statistically different, although it was higher for boys.

At grade 5, boys had significantly higher scores in all areas, except there was no statistical difference for letter sound knowledge and passage comprehension. The differences in summary scores were 7.4 points and 6.3 points higher for the boys at grades 3 and 5, respectively.

TABLE 10: PERCENT CORRECT SCORES BY GRADE, TASK, AND GENDER

Task (Subtest)	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
1. Orientation to print	62.2%	63.4%	73.9%*	69.0%
2. Letter name recognition	46.8%	44.7%	62.3%*	59.1%
3. Phonemic awareness	37.2%*	33.0%	47.0%*	38.8%
4. Letter sound knowledge	30.8%*	26.2%	42.8%	43.3%
5. Familiar word reading	50.2%*	34.3%	73.7%*	64.7%
6. Non-word reading	38.7%*	26.1%	60.9%*	53.2%
7a. Passage reading	50.0%*	34.1%	71.4%*	60.4%
7b. Passage comprehension	8.1%*	3.3%	21.0%	18.2%
8. Listening comprehension	14.9%*	7.2%	36.8%*	27.3%
Reading Summary Score	37.6%*	30.2%	54.4%*	48.1%

* Indicates that the group's performance the group was significantly higher $p < 0.01$

FIGURE 5: GRADE 3 PERCENT CORRECT SCORES BY TASK AND GENDER

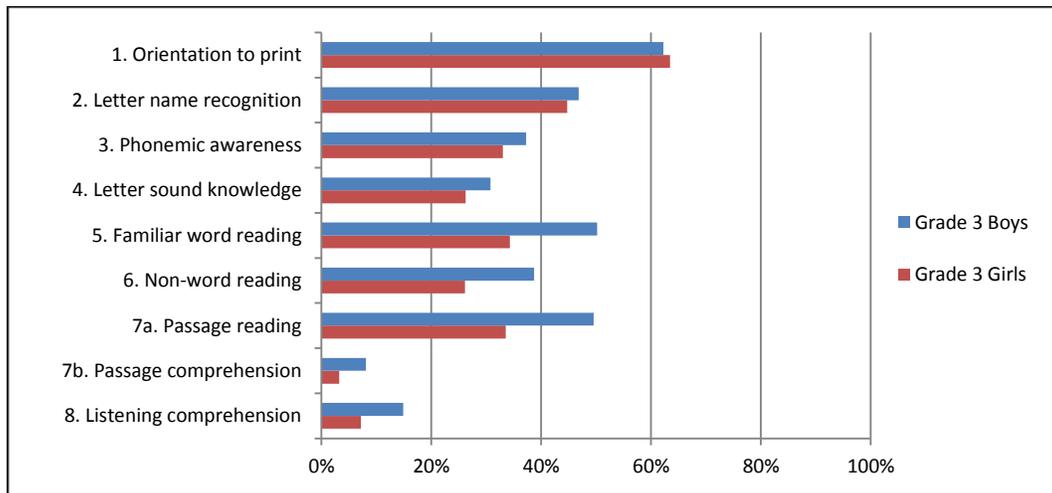
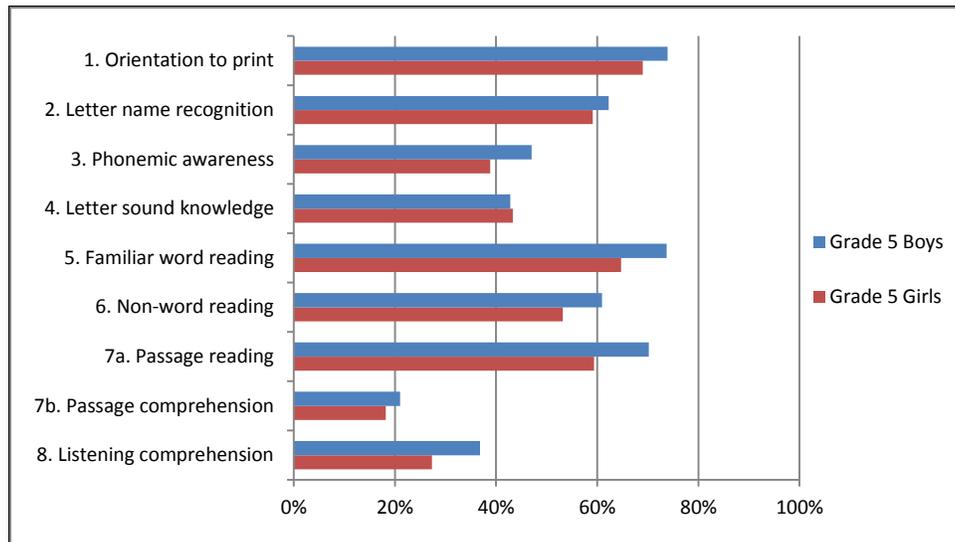


FIGURE 6: GRADE 5 PERCENT CORRECT SCORES BY TASK AND GENDER



Timed Tasks: Phonics and Reading-Rate Fluency Scores

Fluency is a measure of reading efficiency. On the Pakistan EGRA, there were two types of fluency measures: phonics and reading rate. The phonics-fluency subtest included letter name recognition, letter sound knowledge, and non-word reading, whereas, the reading-rate fluency subtest consisted of familiar word and passage reading.

Tables 11 to 13 below show scores in terms of raw scores (instead of the percent correct scores on the previous tables). Table 11 has the maximum raw scores attained by students on each task at each grade level. Tables 10 and 11 have mean scores for the students. In addition, adjustments were made to the raw scores for those students who finished the task before the end of one minute. For instance, if a student read 50 words correctly in 30 seconds, their words correct per minute score would be 100 (50 words x 60 seconds/30 seconds). Table 13 shows the scores by gender. Table 11 provides the baseline maximum scores at grade 3 and 5 for the five timed tasks.

TABLE 11: BASELINE MAXIMUM SCORES ON FLUENCY (TIMED) TASKS

Phonics Fluency Subtest	Grade 3	Grade 5
2. Letter name recognition	107	143
4. Letter sound knowledge	103	183
6. Non-word reading	96	152
Reading-Rate Fluency Subtest	Grade 3	Grade 5
5. Familiar word reading	128	104
7a. Passage reading	180	214

For the timed tasks, familiar word and passage reading showed the most progression over the two grade levels. The lowest scores were in the area of non-word reading for both grades. The improvement from grade 3 to 5 was statistically significant ($p < 0.001$) for all tasks. By gender, there were significant differences in favor of boys at grade 3 on all timed tasks, with the largest differences in familiar word reading and passage reading. Again, in grade 5, boys had statistically higher scores in all tasks except for a phonics item, letter sound knowledge. These fluency tasks showed similar tendencies as the percent-correct scores with boys and grade 5 students posting relatively higher scores.

TABLE 12: PHONICS AND READING-RATE FLUENCY TASK MEANS BY GRADE

Phonics Fluency Subtest	Grade 3	Grade 5	Difference (G5 – G3)
2. Letter name recognition	46.4	62.0	15.6 points
4. Letter sound knowledge	28.5	43.6	15.1 points
6. Non-word reading	18.1	34.3	16.3 points
Reading-Rate Fluency Subtest	Grade 3	Grade 5	Difference (G5 – G3)
5. Familiar word reading	27.9	52.0	24.0 points
7a. Passage reading	33.7	62.6	28.8 points

TABLE 13: PHONICS AND READING-RATE FLUENCY TASK MEANS BY GRADE AND GENDER

Phonics Fluency Subtest	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
2. Letter name recognition	48.2*	44.6	63.6*	60.1
4. Letter sound knowledge	30.9*	26.0	43.0	44.2
6. Non-word reading	21.6*	14.6	36.7*	31.7
Reading-Rate Fluency Subtest	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
5. Familiar word reading	33.8*	22.1	57.5*	45.5
7a. Passage reading	40.8*	26.8	68.9*	55.2

*Indicates that the performance of the group was significantly higher, $p < 0.01$

Questionnaire Findings

Selected results are presented below, particularly for those characteristics or items that showed significant results. Note that there were some students, teachers, and head teachers who did not respond to certain questionnaire items; they were labeled as missing. The total averages for the summary scores were calculated based on the grade averages. Statistical significance was determined based on t -tests for indicators with two categories and analysis of variance (and post hoc pairwise comparisons) for indicators with three or more categories.

Student Questionnaires

Table 14 shows the EGRA summary scores by student age. According to the National Education Policy (2009), the official age of the students at the beginning of the different grade levels of primary education is 6 to 10 years old. Since the baseline took place during the school year, the normal ages for this analysis were set at 8 to 9 years old for grade 3 and 10 to 11 years old for grade 5. The students were placed into three categories: younger than normal age for their grade, normal age, and older than normal age. There were significant differences in the scores. At grade 3, summary scores increased with relative age. The older than normal students had significantly higher scores than the normal age group ($p < 0.05$). However, there was no statistical difference between the younger group and the other two older groups. This could have been due to the low sample size of younger students (13). At grade 5, the younger group had lower scores, but they were not significantly different from the two older groups. Again, this could be due to the low sample size of the youngest group.

TABLE 14: SUMMARY SCORES BY STUDENT AGE

Age Group	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Younger than normal age	13	24.9%	14	50.0%
Normal age	168	28.7%	200	51.4%
Older than normal age	839	35.2%*	740	51.5%
Missing	2	-	18	-
Total	1,022	34.0%	963	51.4%

* Indicates that the group's performance the group was significantly higher $p < 0.05$ level

Table 15 shows the EGRA summary scores by whether the student reads the Quran at home. There are differences in scores for each grade level, however caution should be used in interpreting the significance of these findings due to large differences in sample sizes of those who read and do not read the Quran at home.

TABLE 15: SUMMARY SCORES BY READING THE QURAN AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	42	20.8%	17	37.1%
Yes	973	34.5%*	940	51.8%*
Missing	5	-	6	-
Total	1,022	34.0%	963	51.4%

* Indicates that the group's performance the group was significantly higher $p < 0.01$

Table 16 shows the differences in scores based on whether there is a library at the school. The missing category included students who responded they “did not know” if there was a library at school. At grade 3, this was 39 percent, and for grade 5 it was 32 percent. Surprisingly, at grade 3, students who reported the presence of no library had higher summary scores than those with a school library. In contrast, there was no significant difference at grade 5. Given the number of students with missing data, these results should be interpreted with caution since some students may not have understood this question. The school-level questionnaire revealed that only 6 percent of the schools had a library.

TABLE 16: SUMMARY SCORES BY THE PRESENCE OF A LIBRARY AT THE SCHOOL

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	533	37.5%*	583	54.4%
Yes	88	31.4%	69	50.7%
Missing	401	-	311	-
Total	1,022	34.0%	963	51.4%

* Indicates that the group's performance the group was significantly higher $p < 0.01$

In Tables 17 to 19, the data showed that the presence of newspapers, magazines, and books in the home made a difference in grade 5 reading scores. Scores were significantly higher for both grades for students who reported having books at home. There may be evidence that increasing the presence of reading materials in the home could contribute to raising children's reading levels.

TABLE 17: SUMMARY SCORES BY THE PRESENCE OF NEWSPAPERS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	631	34.2%	451	49.0%
Yes	391	33.6%	512	53.5%*
Missing	0	-	0	-
Total	1,022	34.0%	963	51.4%

* Indicates that the group's performance the group was significantly higher $p < 0.01$

TABLE 18: SUMMARY SCORES BY THE PRESENCE OF MAGAZINES AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	980	33.9%	905	50.8%
Yes	42	36.4%	58	60.8%*
Missing	0	-	0	-
Total	1,022	34.0%	963	51.4%

* Indicates that the performance of the group was significantly higher, $*p < 0.01$

TABLE 19: SUMMARY SCORES BY THE PRESENCE OF BOOKS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	841	31.0%	795	50.0%
Yes	181	47.9%*	468	58.0%*
Missing	0	-		-
Total	1,022	34.0%	963	51.4%

* Indicates that the performance of the group was significantly higher, *p < 0.01

The final set of student questions (in Tables 20 to 22) pertained to children’s reading habits at home. In general, these habits made a difference in their scores for both grades. Among all three reading habits, reading to someone else and reading silently at home had the greatest effects for both grades.

TABLE 20: SUMMARY SCORES BY CHILDREN HAVING SOMEONE READ TO THEM AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	492	33.0%	393	49.6%
Yes	496	35.8%*	561	52.8%*
Missing	34	-	9	-
Total	1,022	34.0%	963	51.4%

* Indicates that the performance of the group was significantly higher, *p < 0.05

TABLE 21: SUMMARY SCORES BY CHILDREN READING TO SOMEONE ELSE AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	557	32.0%	417	49.2%
Yes	433	37.5%*	541	53.3%*
Missing		-		
Total	1,022	34.0%	963	51.4%

* Indicates that the group’s performance the group was significantly higher p < 0.01

TABLE 22: SUMMARY SCORES BY CHILDREN READING SILENTLY AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	167	27.1%	83	37.5%
Yes	838	35.7%*	872	52.9%*
Missing	17	-	8	-
Total	1,022	34.0%	963	51.4%

* Indicates that the performance of the group was significantly higher, *p < 0.01

Teacher Questionnaires

With the smaller sample size, the analysis of the teacher questionnaires was limited to providing descriptive statistics on teacher characteristics and summary scores, i.e., with no group comparisons. Tables 23 to 25 provide information on teacher academic qualifications, professional qualifications, age, years of experience, and in-service training. Neither teacher academic or professional qualification showed a consistent pattern in student reading scores. Tables 25 to 26 also showed no clear pattern in age or years of experience of the teacher. Table 27 indicated no relation to the number of training sessions and student scores. Any observed differences should be treated with caution due to the small sample sizes among these teacher characteristics.

TABLE 23: SUMMARY SCORES BY TEACHER ACADEMIC QUALIFICATION

Academic Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.A./M.Sc.	14	33.0%	17	51.5%
B.A./B.Sc.	14	31.9%	17	54.5%
F.A./F.Sc.	13	33.2%	9	54.1%
Matric	11	29.0%	9	51.3%
Missing	1	-	1	-
Total	53	34.0%	53	51.4%

TABLE 24: SUMMARY SCORES BY TEACHER PROFESSIONAL QUALIFICATION

Professional Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.Ed./M.A.	3	31.6%	3	47.5%
B.Ed.	10	30.1%	16	55.1%
C.T.	7	37.8%	7	53.2%
P.T.C.	28	29.9%	20	51.7%
Missing	5	-	7	-
Total	53	34.0%	53	51.4%

TABLE 25: SUMMARY SCORES BY TEACHER AGE

Age Group in Years	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
40 and less	28	29.2%	28	52.8%
Between 41 and 50	17	35.0%	17	55.6%
51 and more	4	33.2%	3	42.0%
Missing	4	-	5	-
Total	53	34.0%	53	51.4%

TABLE 26: SUMMARY SCORES BY TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
10 or less	23	30.4%	18	52.7%
Between 11 and 20	18	34.6%	16	54.9%
Between 21 and 30	5	28.6%	21	50.7%
31 or more	4	33.2%	1	48.5%
Missing	3	-	6	-
Total	53	34.0%	53	51.4%

TABLE 27: SUMMARY SCORES BY TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	31	31.9%	32	52.6%
One time	13	34.0%	14	52.4%
Two times	3	28.8%	3	51.9%
Three times	6	32.3%	2	48.9%
Missing	0	-	2	-
Total	53	34.0%	53	51.4%

Head Teacher Questionnaires

Similar to the teacher questionnaires, the sample size for the head teacher questionnaires was small ($n = 70$), so interpretations of the data should be treated with caution. The characteristics presented are head teacher academic qualification, professional qualification, experience, in-service training, support to teachers in reading, and training in supporting reading (Tables 28 to 33). As with the teacher data, no clear conclusions were found with the head teacher characteristics and student scores.

TABLE 28: SUMMARY SCORES BY HEAD TEACHER ACADEMIC QUALIFICATION

Academic Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.A./M.Sc.	27	30.0%	27	49.7%
B.A./B.Sc.	13	34.1%	13	47.1%
F.A./F.Sc.	13	36.5%	13	55.1%
Matric	14	36.0%	14	53.2%
Missing	3	-	3	-
Total	70	33.2%	70	50.8%

TABLE 29: SUMMARY SCORES BY HEAD TEACHER PROFESSIONAL QUALIFICATION

Professional Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.Ed./M.A.	11	26.8%	11	50.0%
B.Ed.	16	34.9%	16	50.6%
C.T.	5	41.3%	5	51.5%
P.T.C.	36	34.4%	36	52.2%
Missing	2	-	2	-
Total	70	34.0%	70	51.4%

TABLE 30: SUMMARY SCORES BY HEAD TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
2 or less	2	31.3%	2	45.3%
3 to 5	10	36.9%	10	54.4%
6 to 10	16	34.9%	16	50.0%
11 or more	40	32.6%	40	50.4%
Missing	2	-	2	-
Total	70	34.0%	70	51.4%

TABLE 31: SUMMARY SCORES BY HEAD TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	37	34.8%	37	52.6%
1 time	19	30.0%	19	48.6%
2 times	8	39.2%	8	52.9%
More than 2 times	4	30.2%	4	48.4%
Missing	2	-	2	-
Total	70	34.0%	70	51.4%

TABLE 32: SUMMARY SCORES BY HEAD TEACHER SUPPORT TO TEACHERS IN READING

Support to Teachers	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	6	37.2%	6	46.7%
Yes	62	33.9%	62	52.1%
Missing	2	-	2	-
Total	70	33.9%	70	51.4%

TABLE 33: SUMMARY SCORES BY HEAD TEACHER TRAINING IN TEACHING READING

Support to Teachers	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	37	34.8%	37	52.6%
Yes	31	32.4%	31	49.7%
Missing	2	-	2	-
Total	70	33.9%	70	51.4%

School Characteristics

The final section provides information on school characteristics (from the head teacher questionnaires) by student summary scores. As with the teacher and head teacher characteristics, the sample size of the school characteristics means that any statistical comparisons should be interpreted with caution and may be of limited value in generalizing the results to FATA as a whole (Tables 34 to 37).

The school surveys revealed that the presence of a library, PTA/SMC/PTSMC/PTC, or infrastructure was not related to higher reading scores. Only 6 percent of the schools reported having libraries, yet almost 60 percent had a PTA-like organization. In contrast, school gender and school size were related to the EGRA scores. Male schools had higher scores than female schools, and there was a moderate positive correlation ($r_p = 0.48$) between school size and EGRA scores (i.e., larger schools tended to have higher scores).

TABLE 34: SUMMARY SCORES BY SCHOOL GENDER

School Gender	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Male school	38	36.6%	38	54.0%
Female school	28	31.5%	28	49.1%
Mixed school	1	41.4%	1	50.9%
Missing	3	--	3	--
Total	70	34.0%	70	51.4%

TABLE 35: SUMMARY SCORES BY PTA/SMC/PTSMC/PTC

Parent Teacher Committee	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	24	34.1%	24	52.7%
Yes	40	34.0%	40	51.9%
Missing	6	-	6	-
Total	70	34.0%	70	51.4%

TABLE 36: SUMMARY SCORES BY PRESENCE OF A SCHOOL LIBRARY

School Library	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	60	34.1%	60	51.9%
Yes	8	26.7%	8	43.4%
Missing	2	-	2	-
Total	70	34.0%	70	51.4%

TABLE 37: SUMMARY SCORES BY INFRASTRUCTURE (DRINKING WATER, ELECTRICITY, TOILETS)

Number of Infrastructures (Water, Electricity, Toilets)	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	20	41.6%	20	57.4%
1	22	34.2%	22	50.9%
2	18	28.3%	18	44.4%
3	10	27.4%	10	52.4%
Missing	2	-	2	-
Total	70	34.0%	70	51.4%

CHAPTER 4: CONCLUSIONS AND RECOMMENDATIONS

This final chapter provides conclusions from the FATA EGRA baseline. It is organized according to the two main sections in the report: 1) design and methodology, and 2) findings and results. There are also recommendations based on the instrument development, data collection, data entry, and analysis.

Design and Methodology

1. The design followed USAID evaluation guidelines for a cross-sectional approach. In FATA, not all agencies have been selected for implementation of the PRP. Seven of the 13 agencies have been eliminated due to security concerns, leaving six agencies for implementation. For the same reasons, these seven agencies were not considered as possible light treatment groups during the EGRA baseline. Due to the selection of all of schools in the PRP intervention agencies in FATA for full treatment, there is no counterfactual against which to measure the effects of the full treatment above and beyond the light treatment. With the cross-sectional design, the evaluation will be limited to examining the progress of children in grades 3 and 5 over the course of the life of the PRP project.
2. The sampling issues were addressed as well as could have been expected. The main issues were inaccessibility due to security or safety and the lack of schools with the requisite number of children per grade level. However, the actual sample of schools was 100 percent, and the actual sample of students reached over 94.5 percent of the intended sample.
3. The EGRA test was of good quality. The reliability estimates were in the range of previous EGRA administrations in other countries. The task statistics were acceptable, with an appropriate range of p-values and item-total correlations that were at a good level of quality. The characteristics of the test were such that it should be a strong measure of progress over time due to project-led interventions. As with any test, there may be ways to improve on the task and item statistics for the midline and endline.
4. The field implementation was successful, though there were difficulties to overcome, including logistical challenges with the security and safety situations in FATA. In spite of these challenges, there was a high level of standardization reported by the QCOs, which they attributed to the effective training process by the EGRA team. The team paid careful attention to detail in the logistics and test administration, which was reflected in the low error rates in the booklets and in the data entry.

Findings and Results

Several key findings emerged from the baseline assessment in FATA. These are as follows:

1. EGRA was administered to a robust sample at each grade level (3 and 5). Test reliabilities were very good, showing that the EGRA tasks and items worked well in measuring reading constructs at both grade levels. The task and item statistics showed that EGRA discriminates well between low- and high-achieving students in both grades. They also showed that there is adequate room for growth by students in each grade level.
2. In the analysis phase, scores were calculated in three ways: 1) percentage correct scores for the reading tasks, 2) average percentage correct (grand means) for reading summary scores, and 3) adjusted raw scores for the timed reading tasks. These scores provide a comprehensive picture of student performance. Analysis of student, teacher, head teacher, and school characteristics was carried out using the summary scores.

3. Grade 3 students did relatively well in orientation to print, letter name recognition, familiar word reading, and passage reading, though their scores were under 50 percent in most of those areas. They had relatively low skills in passage comprehension, listening comprehension, letter sound knowledge, and non-word reading. Grade 5 students had the highest skills in orientation to print, familiar word reading, passage reading, and letter name recognition. In contrast, the scores for listening and reading comprehension were relatively low.
4. Scores showed strong increases from grade 3 to grade 5 (17 percentage points). Grade 5 students showed strong increases in familiar word reading, non-word reading, passage reading, and listening comprehension. Phonemic awareness, letter sound knowledge, and passage comprehension were the areas of least improvement, and therefore there is much room for growth.
5. Gender comparisons revealed that boys performed better on nearly all of the reading tasks at both grade levels. At grade 3, boys had significantly higher scores in all areas except the two tasks with the lowest difficulty – orientation to print and letter name recognition. At grade 5, boys had significantly higher scores in all areas except for letter sound knowledge and passage comprehension. The differences in summary scores were 7.4 points and 6.3 points higher for the boys at grades 3 and 5, respectively.
6. Students were timed on five tasks as they read words or passages. These tasks were categorized into phonics fluency (letter name recognition, letter sound knowledge, and non-word reading) and reading-rate fluency (familiar word and passage reading). Students at both grades had lower phonics fluency scores than reading-rate fluency. Moreover, gains from grade 3 to grade 5 were lower for phonics than reading-rate fluency tasks. For the timed tasks, familiar word and passage reading showed the most progression over the two grade levels. The lowest scores were in non-word reading for both grades. The improvement from grade 3 to 5 was statistically significant for all tasks. By gender, there were significant differences in favor of boys at grade 3 on all timed tasks. The largest differences were in familiar word reading and passage reading. Similarly for grade 5, boys had statistically higher scores in all tasks except letter sound knowledge. These timed task scores showed the same tendencies as the non-timed tasks; scores were generally higher for boys.
7. Student questionnaires revealed three positive findings. First, reading the Quran and other books was related to higher reading scores for both grades. Second, higher reading outcomes were reported by students reading in the home, whether that meant being read to, reading to someone else, or reading silently. Third, attending school at an older-than-normal age seemed to have a positive effect on reading outcomes at grade 3, but that advantage waned and was not significant by grade 5.
8. Questionnaire findings for the teachers and head teachers were mostly inconclusive, due to small sample sizes and the lack of variation in the responses. The teacher and head teacher questionnaire results showed no conclusive patterns in terms of age, experience, education, certification, or training.

Evaluation Recommendations

Given the success of the baseline assessment in FATA (and in the other provinces), the methods used in 2013 should be repeated as much as possible for the midline and endline assessments in future years. This should be conducted as follows:

1. The EGRA instruments proved to be of high quality, and equivalent versions of those tools should be developed – through trans-adaptation, piloting, and revision – for the midline and endline assessments so that progress can be accurately measured over time.

2. The EGRA items and tasks had good reliability values and covered the low-to-middle difficulty range. At baseline, the reading scores were relatively low for both grades, and show room for growth. In addition, histograms and box plots provided evidence that the tool is expected to measure higher levels of reading that are anticipated due to project-led interventions. Therefore, the baseline data indicates that the EGRA is appropriate for measuring increases in reading ability at midline and endline.
3. The sampling was reasonable in terms of finding a balance between the resources available, the required sample size, and the geographic coverage. It should be maintained in the midline and endline, i.e., keep the same districts and schools, along with the methods at the school level.
4. The data entry process took time to develop, but it eventually proved to be advantageous in terms of having the data entry operators connect to a central server. This facilitated the two rounds of data entry and the reconciliation process. This system should also be repeated in subsequent data entry activities.
5. The analysis should follow the same procedures, with calculations of reliability, difficulty, task percent-correct scores, summary scores, and fluency (timed) task scores. The baseline, midline and endline scores should be comparable, so that improvements in students' reading can be accurately examined.
6. Reading proficiency levels should be created to provide educators and other stakeholders with meaningful results. Most parents and educators better understand reading achievement in useful terms or levels, such as emerging, proficient, or advanced, rather than interpreting a percent-correct test score that may differ by test or reading passage difficulty. Education officials are encouraged to select specific EGRA scores to serve as levels of reading proficiency for both grades. Percent correct for each task, summary score, as well as fluency rates are recommended for this purpose. The baseline EGRA data can be used for establishing these reading proficiency levels.
7. Finally, it may be advisable to add items to the student, teacher, and head teacher questionnaires to collect data on PRP- and SRP-supported interventions so that student scores can be correlated with these indicators.

In general, the FATA baseline was successful in providing accurate data on which to base decisions for implementation of the PRP interventions, and also for tracking student reading progress over time. It provides a solid foundation for the midline and endline assessments.

ANNEXES

Annexes 1 to 4 provide additional information on the EGRA baseline. Specifically, the annexes have the following:

Annex 1 gives complete item statistics – p-values (the difficulty of the items) and item-total correlations (the quality of the items) by grade – for the items associated with the various tasks. These are more detailed than the task statistics presented in Chapter 3 of the report. Measurement specialists often request these kinds of item statistics for the purposes of quality control, analysis, and test equating.

Annex 2 provides box plots for the fluency tasks. The box plots are more task-specific than the overall score distributions (histograms) presented in the report. They show the median (middle score), the range (highest and lowest scores), and the distribution of scores (by quartiles) for each task. The task-specific distributions are useful to EGRA specialists who place emphasis on the fluency tasks.

Annex 3 gives two examples of categorizing passage reading fluency scores using performance levels. The categorizations – along with raw scores and scale scores -- are often used to interpret test scores. The first example combines reading speed with comprehension, while the second example only uses reading speed. Each example uses a set of cut-scores for placing the students into performance categories.

Annex 4 provides detailed information on the second example, with results for each category of fluency and each level of comprehension. These data can be used as evidence on the reliability of using a combined measure of fluency and comprehension for setting performance cut-scores. The validity of combining these scores is more of an issue for reading experts.

Annex I: Complete Item Statistics by Grade

Table A1 presents item statistics for the untimed tasks, each of which have multiple items. For instance, task 1 (orientation to print) has item statistics for its five items (Q1 to Q5). Note that the timed tasks are lists of letters, sounds, and words, i.e., not items, so it is not necessary to calculate item statistics for them.

Previously, we presented task statistics (Chapter 3, Table 8) with explanations of how they are calculated. These item statistics are calculated in the same way. They show the difficulty and quality of the items. Recall that when constructing a test, we strive for tasks and items that have difficulty values (p-values) that are spread across a range of about 0.05 to 0.90 and quality values (item-total correlations) of at least 0.20. The difficulty values ranged from 0.03 to 0.91 for grade 3 and 0.12 to 0.95 for grade 5, indicating an acceptable range of item difficulties. At both grade levels, a total of 21 items out of the 23 total items per grade had item-total correlations of at least 0.20, indicating high quality items.

TABLE A1: COMPLETE ITEM STATISTICS BY GRADE

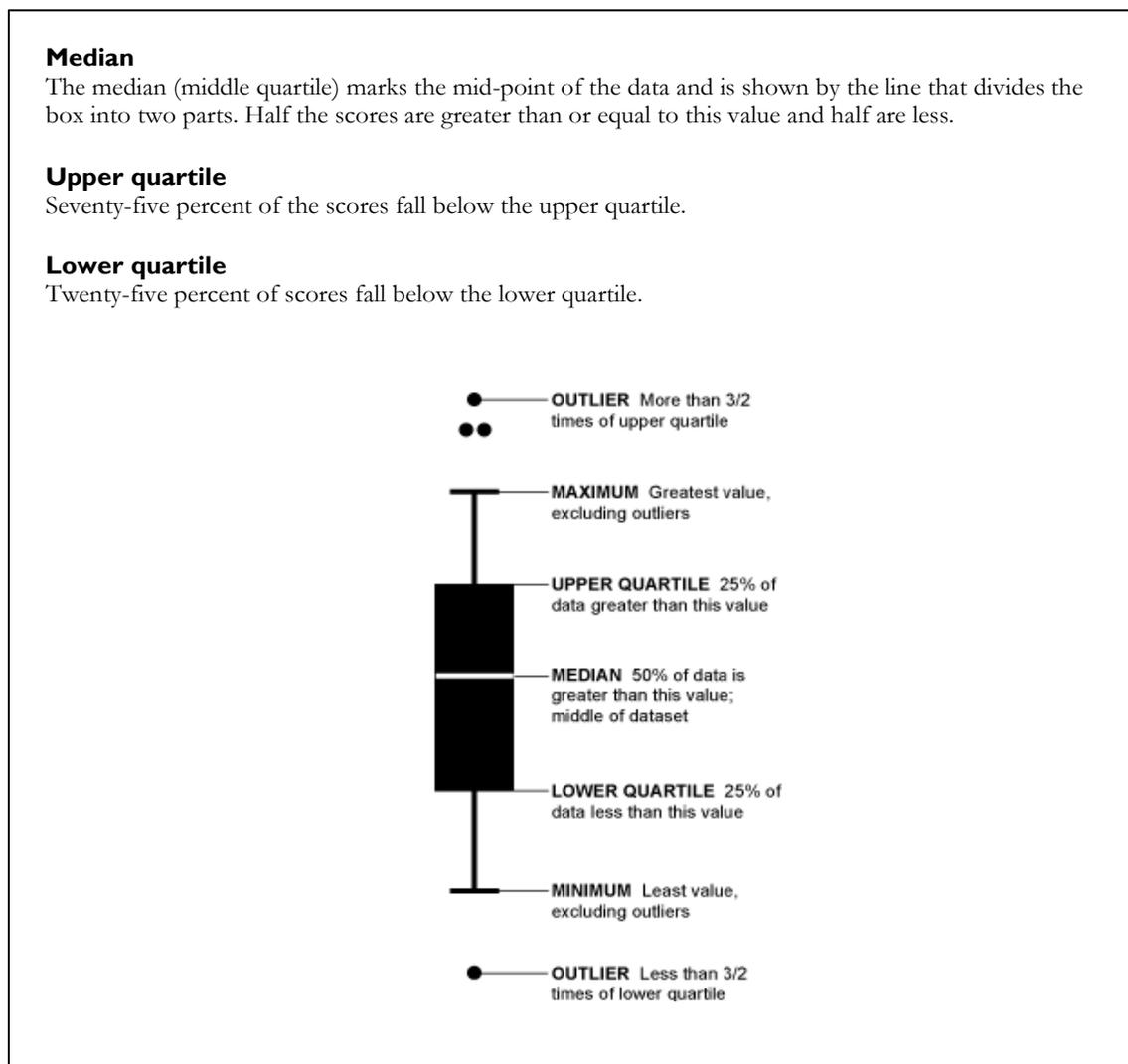
Task (Subtest)	Item	Grade 3		Grade 5	
		P-Value	Item-Total	P-Value	Item-Total
1. Orientation to print (untimed)	Q1	0.87	0.34	0.91	0.26
	Q2	0.91	0.45	0.95	0.32
	Q3	0.72	0.28	0.70	0.15
	Q4	0.06	0.02	0.22	0.09
	Q5	0.58	0.22	0.81	0.25
2. Letter name recognition (timed)	--				
3. Phonemic awareness (untimed)	Q1	0.55	0.45	0.65	0.41
	Q2	0.29	0.46	0.44	0.52
	Q3	0.34	0.38	0.40	0.39
	Q4	0.25	0.27	0.29	0.35
	Q5	0.36	0.42	0.43	0.45
	Q6	0.49	0.44	0.59	0.41
	Q7	0.20	0.31	0.28	0.40
	Q8	0.26	0.42	0.34	0.42
	Q9	0.28	0.36	0.30	0.41
	Q10	0.50	0.43	0.60	0.43
4. Letter sound knowledge (timed)	--				
5. Familiar word reading (timed)	--				
6. Non-word reading (timed)	--				
7a. Passage reading (timed)	--				
7b. Passage comprehension (untimed)	Q1	0.05	0.31	0.19	0.57
	Q2	0.03	0.35	0.16	0.50
	Q3	0.04	0.17	0.12	0.37
	Q4	0.09	0.48	0.26	0.62
	Q5	0.08	0.46	0.24	0.61
8. Listening comprehension (untimed)	Q1	0.09	0.31	0.30	0.40
	Q2	0.03	0.21	0.12	0.32
	Q3	0.20	0.33	0.52	0.39

Annex 2: Box Plots for Phonics and Reading-Rate Fluency Tasks

EGRA places a high emphasis on fluency (timed) tasks. In addition to the descriptive statistics in Table 9 (percent correct scores) and Table 14 (fluency task means), we show box plots for the different fluency tasks. Widely used since their development in the 1960s, box plots are a convenient way for graphically presenting numerical data.

Box plots have two characteristics: 1) central tendency (i.e., the median, or the middle score in the data) and 2) variation (i.e., the range, with scores grouped by quartile). The boxes (which are actually rectangles) represent the two middle quartiles of the scores and the “whiskers” represent the upper and lower quartiles. The small circles on the ends of the whiskers represent outliers. The figure below provides a more detailed explanation for interpreting box plots.

FIGURE A1: UNDERSTANDING BOXPLOTS



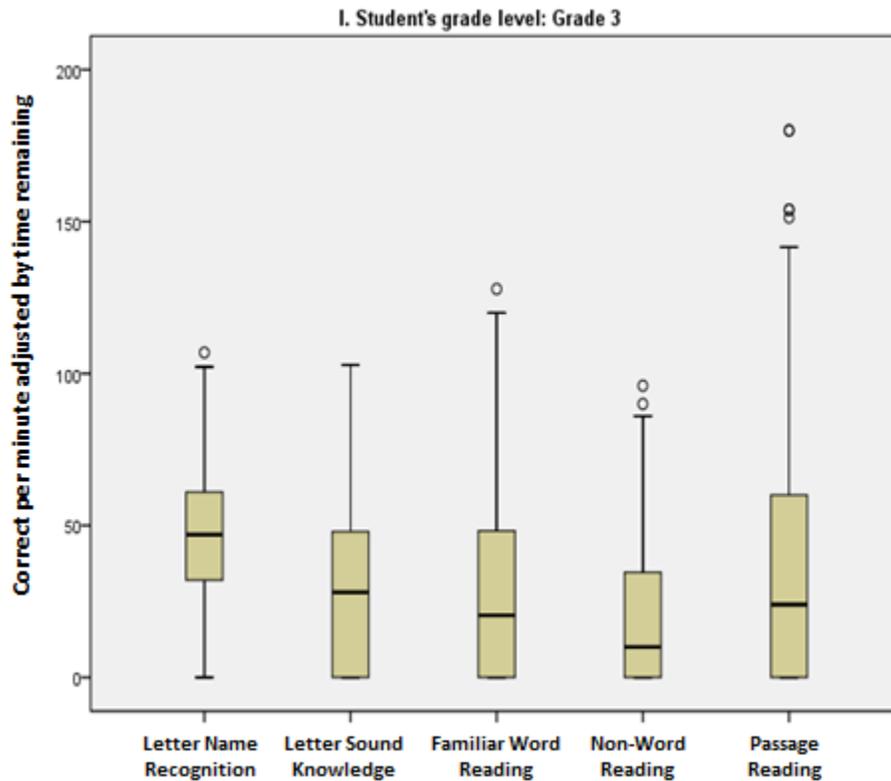
Box plots are presented below (Figures A2 and A3) for the results by grade level on the five fluency (timed) tasks: letter name recognition (task 2), letter sound knowledge (task 4), familiar word reading (task 5), non-word reading (task 6), and passage reading (task 7a).

Grade 3

For grade 3, the central tendency (i.e., the median speed, or the line in the middle) for each of the tasks ranged from about 10 (non-word reading) to about 40 (letter name recognition) items per minute. It shows that the students had better knowledge of letter names than grapheme-morpheme correspondence.

The variation (i.e., the range of scores, without outliers) for each of the tasks varied from about 80 (non-word reading) to about 140 (passage reading). It shows that the scores were more spread out when reading connected words than sounding out pseudo-words.

FIGURE A2: PHONICS AND READING-RATE FLUENCY BOX PLOTS FOR GRADE 3



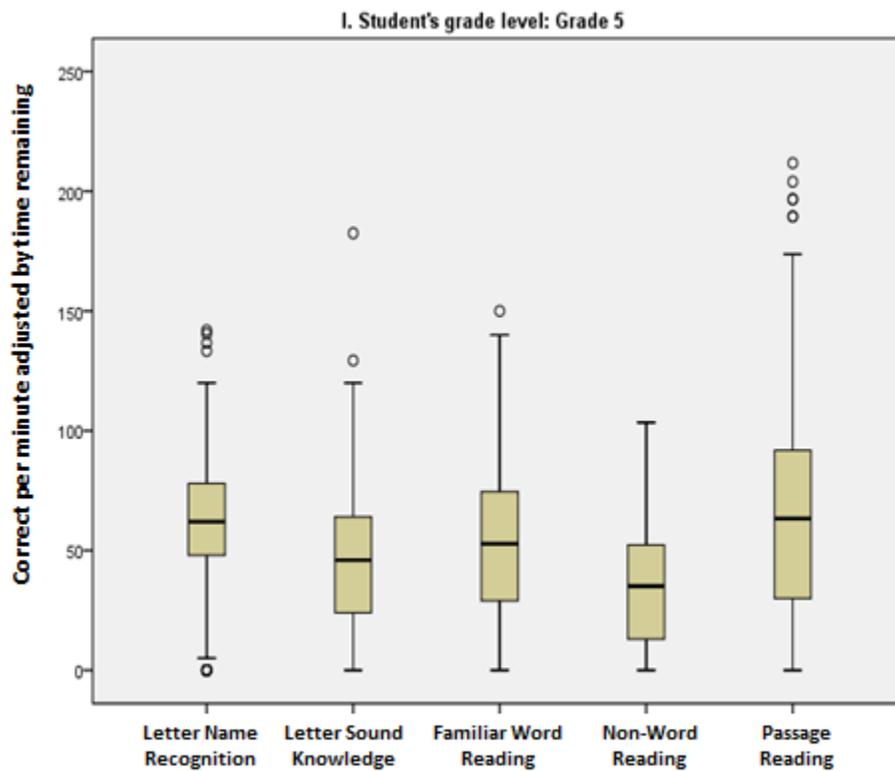
Grade 5

For grade 5, the central tendency (the median speed) for each of the tasks ranged from about 40 (non-word reading) to about 60 (passage reading) items per minute. It shows that the students had more fluency reading connected words than conducting grapheme-morpheme correspondence.

The variation (range of scores) for each of the tasks varied from about 100 (non-word reading) to about 180 (passage reading). It shows that the scores were more spread out when reading connected words than sounding out pseudo-words.

Note also that the medians and the ranges increased from grade 3 to grade 5 for all fluency tasks. Many students are becoming more fluent readers at grade 5, but there are also those students who are either non-readers or very low readers. These children lack of knowledge of letter names, sight words, connected text, and (especially) phonics.

FIGURE A3: PHONICS AND READING-RATE FLUENCY BOX PLOTS FOR GRADE 5



Annex 3: Examples of Fluency Score Threshold Calculations

There are different ways of interpreting test scores. Three of the main ways are 1) raw scores (e.g., number correct), 2) scale scores (e.g., percent correct), and 3) percentile scores (e.g., rank in relation to other students). In the report, we presented scores in terms of number correct (for the fluency tasks) and percent correct (for all tasks). We could also calculate the percentile scores for each student, though this is not normally done with EGRA. Note that these kinds of calculations do not change or affect the actual results, but they do involve issues of interpretability.

A fourth main way of interpreting scores is through performance categories, e.g., low, middle, and high. This requires setting cut-scores, or thresholds, to separate the student scores into categories, e.g., two cut-scores lead to three performance categories. The following analysis shows two examples of calculating thresholds for passage reading scores (CWPM), which allows us to place the student scores into different performance categories. Note that performance categories are often accompanied by performance level descriptors (PLDs), which give a text-based explanation of the meaning of the scores in each category. We have not developed PLDs for these examples since 1) the threshold setting is at a preliminary stage and 2) reading specialists with knowledge of local curricula and context generally develop the PLDs.

Fluency using an 80 percent comprehension threshold

In the first example, we used a method that has been suggested by some EGRA specialists. It involves calculating the mean reading speed associated with 80 percent comprehension for those that can read at least one word correctly and then applying it as a fluent cut-score. In other words, the mean reading speed for these students signifies whether the students are fluent readers through using both passage reading speed *and* comprehension in the calculation; the fluent cut-score separates the fluent readers from the non-fluent readers. To establish a second threshold, we again followed the suggested method and used the lowest level of reading (1 CWPM) as the non-fluent cut-score. The two cut-scores resulted in three performance levels: non-readers (low), non-fluent readers (middle), and fluent readers (high).

At grade 3, the mean reading speed on the passage reading task (Task 7a) for students who scored 80 percent on the passage comprehension task (Task 7b) was 102.3 (rounded to 102). With this method, 102 CWPM becomes a threshold for grade 3 students who are proficient at passage reading *and* comprehension. At grade 5, the mean speed on the passage reading task (Task 7a) for students who scored 80 percent on the passage comprehension task (Task 7b) was 112.0 (or 112). Then 112 CWPM becomes a threshold for grade 5 students who are proficient at passage reading and comprehension.

The definitions of the three categories in terms of CWPM and the percentages of grades 3 and 5 students in the categories for grades 3 and 5 are shown in Table A2 below.

TABLE A2: THRESHOLDS FOR CWPM WITH 80 PERCENT COMPREHENSION

Category (Performance Level)	Grade 3		Grade 5	
	CWPM	% of Students	CWPM	% of Students
Non-Reader	0	37.8%	0	17.1%
Non-Fluent Reader	1 to 101	56.8%	1 to 111	68.7%
Fluent Reader	102 and above	5.4%	112 and above	14.2%
Total	--	100.0%	--	100.0%

Note that the majority of the students are in the middle category at each grade level. This is due the large range of scores for this category, i.e., from the students who score just above non-readers to those who score just below fluent readers are in the non-fluent reader (middle) category.

Fluency using fixed interval thresholds

In the second example, we used fixed intervals of CWPM for the performance levels. This reduced the problem of having a large range of students in the middle category by creating early reader and intermediate reader categories. It also follows common practice when setting performance categories of having between three and five levels for student scores. We used an interval of 40 CWPM to produce five performance levels, along with a category for the non-readers. The five levels were: non-readers (0 CWPM); early readers (1-40 CWPM); intermediate readers (41-80 CWPM); fluent readers (81-120 CWPM); and advanced readers (121 and above CWPM).

TABLE A3: THRESHOLDS FOR CWPM WITH FIXED INTERVALS

Category (Performance Level)	CWPM	% of Students	
		Grade 3	Grade 5
Non-Reader	0	37.8%	17.1%
Early Reader	1 to 40	23.6%	14.2%
Intermediate Reader	41 to 80	26.4%	34.7%
Fluent Reader	81 to 120	10.0%	24.1%
Advanced Reader	121 and above	2.3%	9.8%
Total	--	100.0%	100.0%

At both grades 3 and 5, the fixed interval method allowed for more distribution of the scores across the categories. We can also see a shift in percentages of students in each category from grade 3 to grade 5; the performance categories allow for a score interpretation showing that students are improving across the grade levels, with more scores in the lower categories at grade 3 and more scores in the higher categories at grade 5.

Remarks

While it is possible to use such percentages to set cut-scores for interpretation purposes at the baseline, midline and endline, this analysis should be taken as preliminary. For instance, more well-known and accepted method of setting thresholds – which is commonly called “standard setting” by measurement specialists – involve holding a workshop with local reading experts to set the cut-scores according to the experts’ conceptions of what students should know and be able to do in order to be classified into a performance category. There are several well-known methods, e.g., Angoff and Bookmark, which have been judged as valid and reliable for this purpose.⁴ Further discussions on setting thresholds involving local reading experts are recommended.

⁴ References include: Zieky, M. & Perie, M. (2006). *A primer on setting cut-scores on tests of educational achievement*. Princeton, New Jersey: Educational Testing Service; Cizek, G. (1996). *Standard-setting guidelines*. Educational Measurement: Issues and Practices, Spring 1996, p. 13-21; Cizek, G., Bunch, M., & Koons, H. (2004). *Setting performance standards: Contemporary methods*. Educational Measurement: Issues and Practices, Winter 2004.

Annex 4: Distribution of Reading Fluency and Comprehension Scores using Fixed Intervals

In this last annex, we provide more information on the relationship between reading fluency (speed) and comprehension using information from the fixed interval method. While the data show a positive relationship between speed and comprehension, there are sizeable numbers of “fluent” readers with little comprehension. Our conclusion is that setting a cut-score using a less than reliable indicator, such as the mean speed of students with 80 percent comprehension (i.e., using *both* speed and comprehension), can be problematic. The result is categorizing some students as fluent readers who in fact, according to the definition, are not, i.e., they have high reading speed but low comprehension. It may be better to set thresholds based solely on a single indicator – reading speed – rather than mixing it with comprehension.

The figures and tables below (Tables A4-A5 and Figures A4-A5) expand on the data in Table A3. They show the results for reading fluency (in terms of speed) by comprehension level for grades 3 and 5. We used the categories based on intervals of 40 CWPM, along with a category for the CWPM non-readers (0 CWPM). Comprehension levels were calculated in terms of percent correct scores (e.g., 20 percent is the same as correctly answering one question out of five total questions). For instance, at grade 3, 100 percent of the non-readers have 0 percent comprehension and 9 percent of the advanced readers have 80 percent comprehension.

TABLE A4: GRADE 3 READING FLUENCY AND COMPREHENSION

Category (Performance Level)	CWPM	% of Students by Comprehension Level						
		0%	20%	40%	60%	80%	100%	Total
Non-Reader	0	100%	0%	0%	0%	0%	0%	100%
Early Reader	1 to 40	90%	8%	1%	0%	0%	0%	100%
Intermediate Reader	41 to 80	64%	20%	11%	4%	1%	0%	100%
Fluent Reader	81 to 120	55%	22%	14%	6%	3%	0%	100%
Advanced Reader	121 and above	43%	22%	17%	9%	9%	0%	100%

FIGURE A4: GRADE 3 READING FLUENCY AND COMPREHENSION

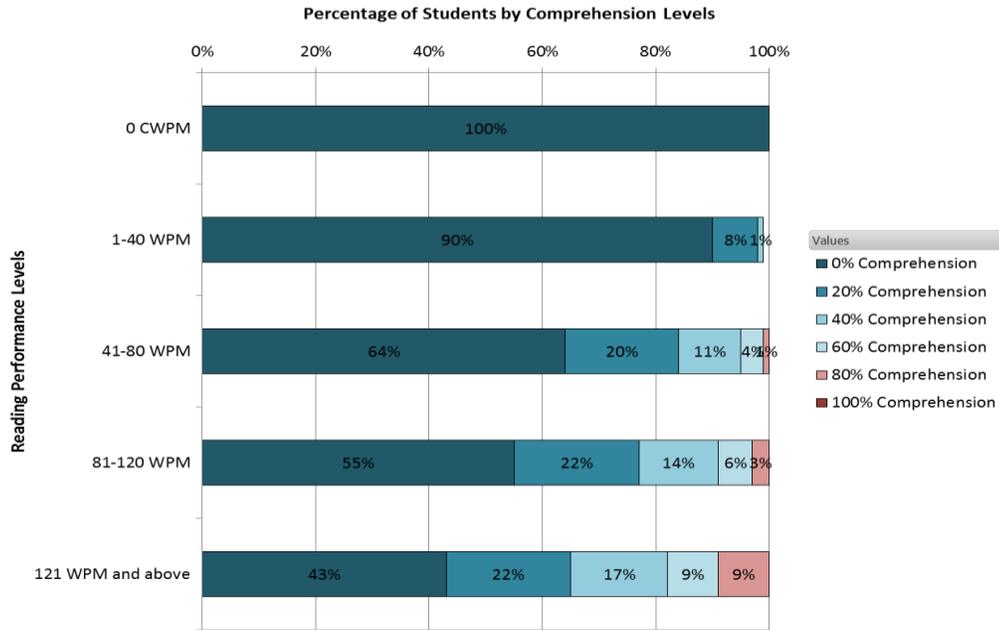
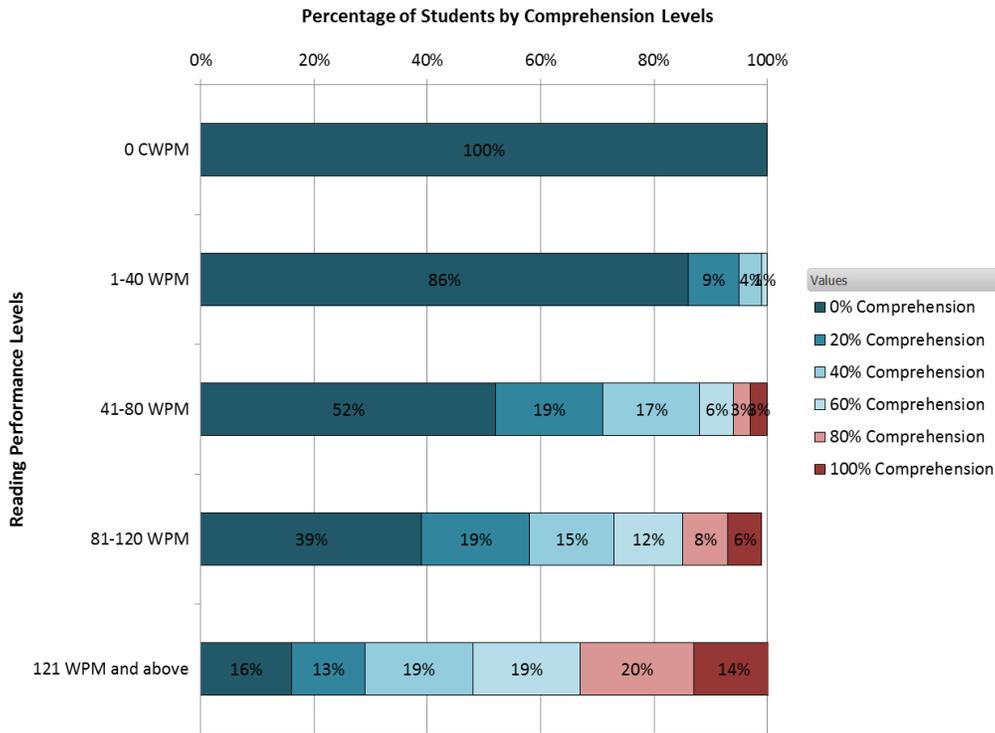


TABLE A5: GRADE 5 READING FLUENCY AND COMPREHENSION

Category (Performance Level)	CWPM	% of Students by Comprehension Level						Total
		0%	20%	40%	60%	80%	100%	
Non-Reader	0	100%	0%	0%	0%	0%	0%	100%
Early Reader	1 to 40	86%	9%	4%	1%	0%	0%	100%
Intermediate Reader	41 to 80	52%	19%	17%	6%	3%	3%	100%
Fluent Reader	81 to 120	39%	19%	15%	12%	8%	6%	100%
Advanced Reader	121 and above	16%	13%	19%	19%	20%	14%	100%

FIGURE A5: GRADE 5 READING FLUENCY AND COMPREHENSION



The main results for the categories of reading speed (from non-readers to advanced readers) in relation to comprehension levels (from 0 percent to 100 percent) for grades 3 and 5 are summarized as follows:

- Non-Readers (0 CWPM) – All of the non-readers had 0 percent comprehension.
- Early Readers (1-40 CWPM) – Most of the early readers (90 percent at grade 3 and 86 percent at grade 5) had 0 percent comprehension. None of them achieved 80 percent comprehension and only 1 percent at grade 5 achieved 60 percent comprehension.
- Intermediate Readers (41-80 CWPM) – More than half of the intermediate readers (64 percent at grade 3 and 52 percent at grade 5) had 0 percent comprehension. A tiny minority of them (1 percent at grade 3 and 6 percent at grade 5) achieved at least 80 percent comprehension.
- Fluent Readers (81-120 CWPM) – A substantial percentage of fluent readers (55 percent at grade 3 and 39 percent at grade 5) had 0 percent comprehension. Only 3 percent at grade 3 and 14 percent at grade 5 achieved at least 80 percent comprehension.
- Advanced Readers (121 CWPM and above) – Fewer than two fifths of the advanced readers (9 percent at grade 3 and 34 percent at grade 5) achieved at least 80 percent comprehension.

The key point from the data is that most of the fluent and advanced readers – at both grade levels – did not reach 80 percent comprehension. Setting a threshold under the assumption that fluent readers (in terms of speed) have a high level of comprehension can be misleading. Conversely, using a single indicator, i.e., reading speed, to set thresholds can be a more reliable way of interpreting the results.