



USAID
FROM THE AMERICAN PEOPLE



EARLY GRADE READING ASSESSMENT BASELINE REPORT

GILGIT-BALTISTAN

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 GILGIT-BALTISTAN

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 Government of Gilgit-Baltistan, Directorate of Education 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, the Institute for Social and Applied Policy Studies (I-SAPS).

CONTENTS

Executive Summary	1
Chapter 1: Introduction	7
Chapter 2: Design and Methodology	9
Chapter 3: Findings and Results	19
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 1 Timeline (January 2013 to May 2014).....	11
Table 2: Sample Schools by District, Gender, and Location for Full Treatment.....	13
Table 3: Reliability Estimates	16
Table 4: EGRA Score Ranges and Calculations	17
Table 5: Example of EGRA Percent Correct and Summary Scores	18
Table 6: Example of EGRA Timed Task Scores.....	18
Table 7: Actual Student Sample by Grade and Gender.....	19
Table 8: Task Statistics (Full and Light Treatment Groups).....	20
Table 9: Percent Correct Scores by Grade and Task (Full and Light Treatment Groups).....	16
Table 10: Scores by Grade, Task, and Group	16
Table 11: Scores by Grade, Task, and Gender (Full and Light Treatment Groups)	18
Table 12: Percent Correct Scores by Group, Grade, and Gender.....	19
Table 13: Baseline Maximum Scores on Fluency (Timed) Tasks (Full and Light Treatment Groups)	20
Table 14: Phonics and Reading-Rate Fluency Task Means by Grade (Full and Light Treatment Groups).....	21
Table 15: Timed Task Scores by Grade and Group.....	21
Table 16: Timed Task Scores by Grade and Gender (Full and Light Treatment Groups).....	22
Table 17: Phonics and Reading-Rate Fluency Task Means by Group, Grade, and Gender.....	22
Table 18: Summary Scores by Student Age.....	23
Table 19: Summary Scores by Reading the Quran at Home	23
Table 20: Summary Scores by the Presence of a Library at the School.....	24
Table 21: Summary Scores by the Presence of Newspapers at Home.....	24
Table 22: Summary Scores by the Presence of Magazines at Home	24
Table 23: Summary Scores by the Presence of Books at Home	25
Table 24: Summary Scores by Children Having Someone Read to Them at Home	25
Table 25: Summary Scores by Children Reading to Someone Else at Home	25
Table 26: Summary Scores by Children Reading Silently at Home	25
Table 27: Summary Scores by Teacher Academic Qualification.....	26
Table 28: Summary Scores by Teacher Professional Qualification	26
Table 29: Summary Scores by Teacher Age	26

Table 30: Summary Scores by Teacher Experience	27
Table 31: Summary Scores by Teacher In-Service Training	27
Table 32: Summary Scores by Head Teacher Academic Qualification.....	28
Table 33: Summary Scores by Head Teacher Professional Qualification.....	28
Table 34: Summary Scores by Head Teacher Experience.....	28
Table 35: Summary Scores by Head Teacher In-Service Training.....	29
Table 36: Summary Scores by Head Teacher Support to Teachers in Reading.....	29
Table 37: Summary Scores by Head Teacher Training in Teaching Reading	29
Table 38: Summary Scores by School Gender	30
Table 39: Summary Scores by PTA/SMC/PTSMC/PTC.....	30
Table 40: Summary Scores by Presence of a School Library	30
Table 41: 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.....	9
Figure 2: Grade 3 Summary Scores.....	15
Figure 3: Grade 5 Summary Scores.....	15
Figure 4: Full Treatment Scores by Grade and Task	17
Figure 5: Light Treatment Scores by Grade and Task	17
Figure 6: Grade 3 Scores by Task and Gender (Full and Light Treatment Groups).....	18
Figure 7: Grade 5 Scores by Task and Gender (Full and Light Treatment Groups).....	19
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
B.Sc.	Bachelor of Science
C.T.	Certificate of Teaching (Grade 12 plus FA/FSC Certificate)
DEO	Data Entry Operators
EGRA	Early Grade Reading Assessment
F.A.	Intermediate College (Grade 12) Certificate in Arts
FATA	Federally Administered Tribal Areas
F.Sc.	Intermediate College (Grade 12) Certificate in Sciences
GB	Gilgit-Baltistan
ICT	Islamabad Capital Territory
I-SAPS	Institute for Social and Applied Policy Studies
KP	Khyber Pakhtunkhwa
M.A.	Master of Arts
Matric	Secondary School (Grade 10) Certificate (Matriculation)
M.Ed.	Master of Education
MOE	Ministry of Education
M.Sc.	Master of Science
MSI	Management Systems International
MT	Master Trainers
NEAS	National Education Assessment System
NEMIS	National Education Management Information System
PRP	Pakistan Reading Project
P.T.C.	Primary Teaching (Grade 12) Certificate
QCO	Quality Control Officer
SPSS	Statistical Package for the Social Sciences
SQL	Structured Query Language
SRP	Sindh Reading Project
STS	School-to-School International
TTI	Teacher Training Institute
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. This report covers the baseline assessment in Gilgit-Baltistan. In May 2013, GB, along with AJK and ICT, was part of Round 1 of the baseline data collection; data from Pakistan's other five provinces/areas/territories (hereafter referred to as provinces) were collected in Rounds 2 and 3 in September and October 2013, respectively. The following activities were carried out for all of the provinces, including GB: 1) design, 2) sampling, 3) instrumentation, 4) planning, 5) training, 6) implementation, 7) analysis, and 8) reporting.

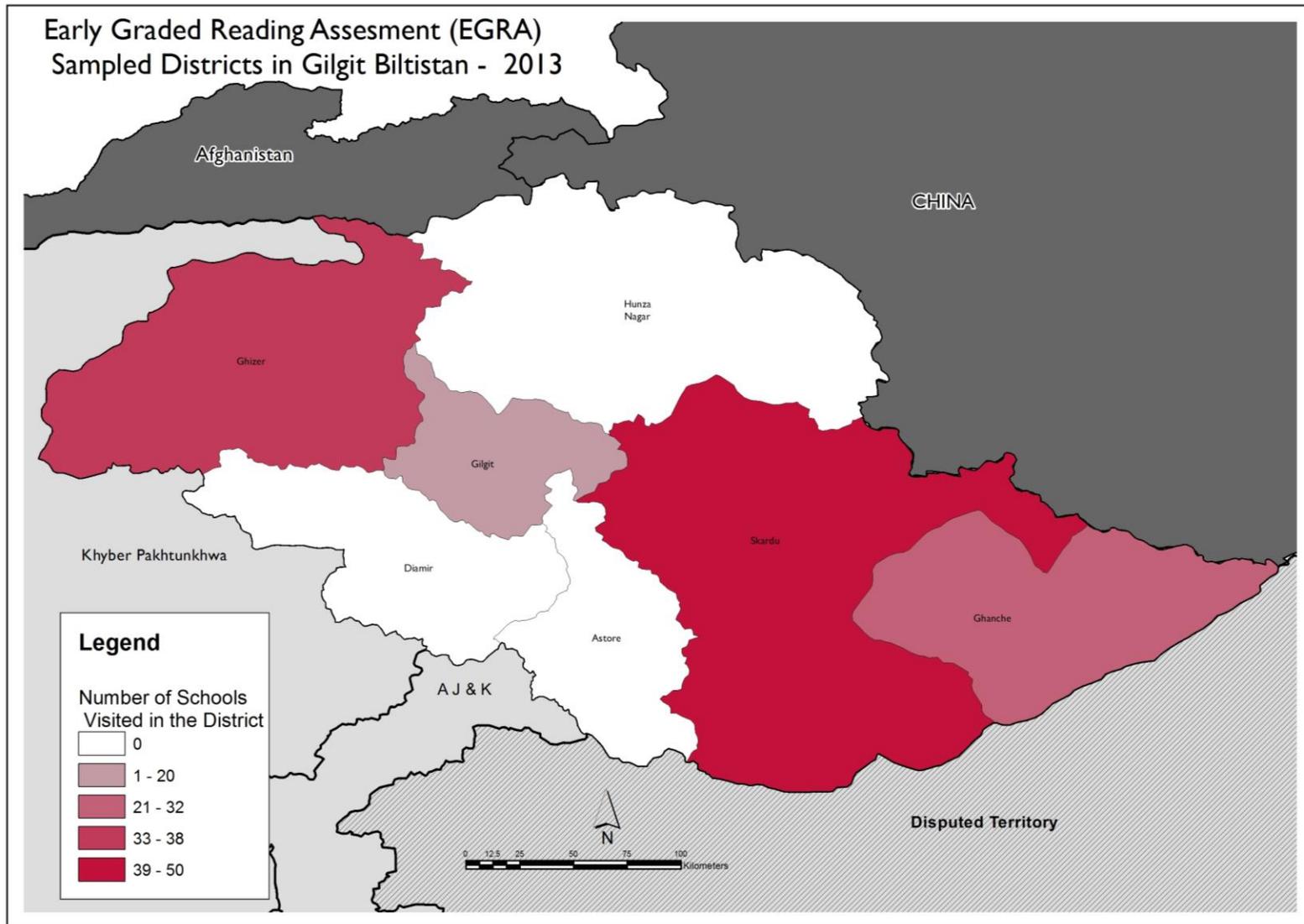
The external evaluation design, which was developed prior to the baseline assessment, was tailored to the implementation of the PRP and SRP in each province. In most of the provinces, a quasi-experimental design will be used, with two treatment groups: “full treatment” and “light treatment.” The full treatment group will receive both the first and second 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 accordance with the USAID evaluation guidelines, students at two selected grade levels – grades 3 and 5 – will be assessed at 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 33,600 children in over 1,120 schools throughout the country. Over the course of the projects, the evaluators will compare the baseline results with those at the midline and endline to examine success in improving children's reading levels 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 GB, all activities were completed by the end of September 2013, including a draft report. The EGRA baseline results were presented and discussed at a consultative meeting in Islamabad on September 25, 2013. Representatives from the provincial Ministry of Education, USAID, PRP, and the contractors (MSI and STS) attended the consultation. Revisions were then made to this report based on the discussions between the stakeholders.

Map of Sampled Districts



Key Points

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

Implementation

1. The GB evaluation involves two kinds of comparisons: 1) a comparison of full and light treatment groups to determine the effects of full treatment above and beyond that of the light treatment, and 2) a comparison of each group to itself at the baseline, midline, and endline. (Please see Figure 1 and the accompanying text for a fuller description of the evaluation design.)
2. Five districts of GB were selected for “full treatment” during the initial consultative meetings between the DOE and USAID in January 2013 (i.e., Gilgit, Skardu, Diamir, Astor, and Hunza Nagar). The other two districts were selected as “light treatment.” Two of the five full treatment districts (i.e., Gilgit and Skardu) were randomly sampled for the baseline assessment along with both of the light treatment districts (i.e., Ghanche and Ghizer).
3. Since the majority of the schools in GB are Urdu-medium, only EGRA in Urdu was used. 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 version of the tools was piloted in AJK, ICT, and KP.
4. A total of 140 schools, with 70 schools from each group (full and light treatment), were randomly selected for the baseline. These schools were selected on a proportional basis from the two full treatment sample districts and from the two light treatment sample districts.
5. The baseline data were collected in the schools in the four sampled districts in GB. A random sample of male and female schools was selected, followed by a random sample of grades 3 and 5 students within those schools.
6. The results from this representative sample are presented in this report as a generalized view of the reading levels for students in GB in Urdu. Please note that district comparisons are not possible because the districts were not evenly sampled; the number of sampled schools varied by district, and the sample sizes are limited for each district. Moreover, the gap between the start of the school year and the EGRA administration fluctuated by district, thereby altering the amount of instructional time students received and potentially affecting the reading performance levels students achieved across the districts.
7. The EGRA testing window for GB was May 2013. All schools were reached during this time period.
8. The assessment tools were successfully administered in the schools in the zones as follows (with a percentage of the target reached in parentheses): 140 schools (100.0 percent) to 3,725 students (88.7 percent), 225 teachers (80.3 percent – with some teachers covering both grades 3 and 5, thus reducing the total number of teachers in the survey) and 140 head teachers (100.0 percent).
9. 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. 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.

10. 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, with a resulting discrepancy rate of less than 1 percent. All data were reconciled across the two databases and with the assessment booklets. A clean data file was produced for analysis.
11. 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. (The calculation of these scores is fully explained in the analysis section of this report.) These scores provide a comprehensive picture of student performance. Contextual analysis of student, teacher, head teacher, and school characteristics was carried out using the summary scores.

Results

1. The EGRA was administered to 1,827 grade 3 students and 1,898 grade 5 students. The reliability was good for both grades ($\alpha = 0.82$ for grade 3 and 0.85 for grade 5). These reliabilities indicate that the items worked well in measuring reading constructs at both grade levels.
2. The task and item statistics showed that the EGRA discriminates well between low- and high-achieving students in both grades. The task p-values for grade 3 provided a spread on the lower to lower-middle section of the difficulty range, while p-values for grade 5 were higher and covered the upper-lower half to the high-middle parts of the spectrum. All task scores at grades 3 and 5 had item-total correlations equal to or greater than 0.30, indicating good discrimination quality for these tasks. (Complete item statistics are listed in Annex 1.)
3. Grade 3 children did relatively better on the orientation to print, letter name recognition, and phonemic awareness tasks, though all of the scores were below 50 percent. The lowest scores were in phonics (non-word reading and letter sound knowledge) and comprehension (passage and listening). At grade 5, their best scores were in the two reading tasks (familiar word and passage). They still had relatively low scores in phonics (non-word reading, letter sound knowledge) and comprehension (passage and listening). There was also substantial progression from grade 3 to grade 5 on some of the tasks scores – especially in familiar word and passage reading.
4. Scores for male and female students were similar, especially at grade 5. By task, the boys tended to perform better in orientation to print and listening comprehension. The girls did better in phonemic awareness, familiar word reading, and passage reading and comprehension. However, the scores for the boys and girls were close to each other on all tasks.
5. 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. Moreover, gains from grade 3 to grade 5 were lower for phonics than for reading-rate fluency tasks. Passage reading (fluency) was approximately 30 points higher in grade 5 than in grade 3. Although the passage was designed for grade 3, this difference shows that the reading levels in grade 3 are low, but that children can make substantial progress in the early grades if expectations are high enough and if they are provided with the opportunity to learn. Specifically mastery of phonics, such as letter sound knowledge, phonemic awareness, and non-word reading, should help the students become better overall readers. It is clear that these types of knowledge and skills are not receiving an appropriate emphasis in GB schools.

6. The summary score for students from full treatment schools was about four points higher in grade 3 and six points higher in grade 5 than in light treatment schools. For each task, students from full treatment schools show higher scores, except for orientation to print and listening comprehension. The differences in mean scores were to be expected since the assignment of districts into full and light treatment was not random. The group scores will require a statistical correction at the midline to ensure equivalence at baseline.
7. Questionnaire findings were mostly inconclusive, due to small sample sizes and the lack of variation in the scores that were related to the student, teacher, and head teacher characteristics. For the students, one of the positive findings was that having reading materials and opportunities to read in the home seemed to have a positive effect on reading outcomes. For the teachers, higher qualifications, both academic and professional, were associated with higher student reading scores. For head teachers, providing support to teachers in reading instruction tended to relate to higher reading scores for students. For the schools, the presence of a library and better infrastructure were associated with better student reading scores.

Evaluation Recommendations

Given the success of the baseline assessment in GB (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 districts and schools along with the sampling methods at the school level.
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 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 students' 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 to collect 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 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 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 schools within districts will receive the same type of treatment.

To measure results from PRP and SRP, a rigorous external evaluation is being conducted. The evaluation baseline took place in 2013, prior to the launch of the reading interventions. In accordance with USAID program evaluation guidelines, samples of students in two selected grade levels – grade 3 and grade 5 – were assessed throughout Pakistan so that independent baselines can be established in each province. Students 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 Gilgit-Baltistan (GB). Along with AJK and ICT, GB was part of the baseline data collection in May 2013; data from Pakistan's other five provinces were collected in September and October 2013. The following activities were planned for all of the provinces, including GB:

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, including GB, this was complemented by a quasi-experimental design with the two treatment groups (full and light).
2. Sampling – Schools were selected from the full and light treatment districts. The sample enabled the collection of student reading assessment data that were representative of the treatment groups, grade levels, gender, and urban/rural zones. There were a total of seven districts in GB, out of which five were full treatment and the remaining two were light treatment. A simple random sample of two of the full treatment districts – Gilgit and Skardu – were taken for full treatment assessment along with the two light treatment districts – Ghanche and Ghizer. Schools were then apportioned according to location and gender. Most of the sampled schools were in rural areas to reflect the preponderance of rural schools in GB. For gender, half of the schools selected for the assessment were male and half were female.
3. Instrumentation – EGRA tools were developed, with tests at the grade 3 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 master trainers, supervisors, enumerators, and QCOs. Enumerators and supervisors were observed to ensure clear comprehension and 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 spreadsheet (Excel) and statistical (SPSS) software. 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.

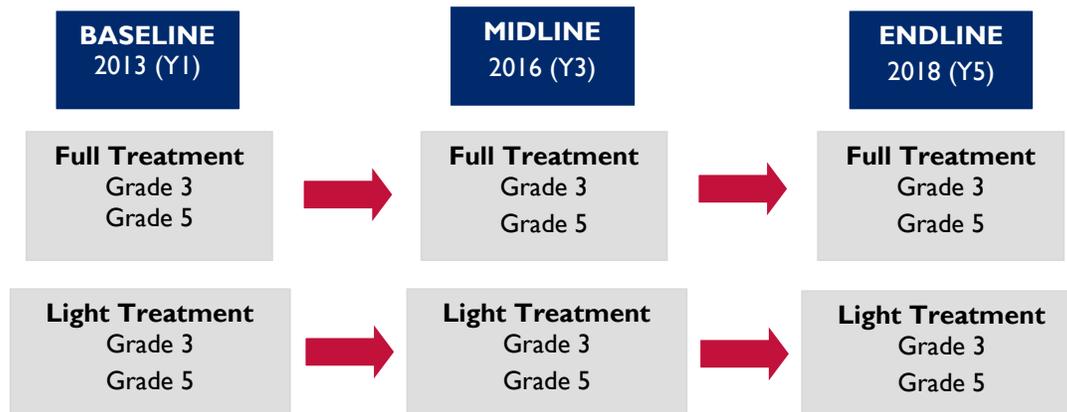
CHAPTER 2: DESIGN AND METHODOLOGY

This chapter presents the evaluation design and methodology, including the 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 for the “full” and “light” treatment groups were pre-selected by the DOE and USAID for GB in January and February 2013. Since district-level selection for the two groups was not random, equivalence at baseline of the two treatment groups cannot be assured, and a quasi-experimental design was selected. In this design, any differences in scores at baseline (and midline and endline) will be statistically removed in the analysis, i.e., the two groups will be made statistically equivalent even though their average scores may be different. This will ensure fairness in the comparison of the full and light treatment groups. In addition, scores between the groups will not be statistically tested at baseline because the goal of the evaluation is to compare the long-term progress of both groups. Providing group comparisons at baseline may introduce potential competition between the groups and invalidate the experimental design.

In addition, while most districts have the two treatment groups, two of the provinces – AJK and ICT – will receive full treatment across all districts, and another province – FATA – will have full treatment in some districts but no treatment (and no data collection) in the others. In GB, five of the districts will be covered by the PRP full reading intervention (i.e., Astor, Diamir, Gilgit, Hunza Nagar, and Skardu), and the other two will be light treatment districts (i.e., Ghanche and Ghizer).

The GB students were tested in Urdu, their main languages of instruction. Equal numbers of male and female schools, i.e., 35 male and 35 female schools per treatment group, 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 GB baseline, like the other provinces for Round 1, was conducted according to a timeline that started in January and ended in September 2013, with submission of draft reports to USAID in October. This final report may be distributed to the provincial DOEs and other stakeholders as appropriate. (See Table 1 below.)

The process began in January with the planning and design of activities, including the creation of 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 GB. From February to April, the EGRA team, with participation from GB and other provinces, then prepared, piloted, and revised the EGRA tools and conducted the district/school sampling. The data collection in GB took place in May, and was followed by the data entry, analysis, and reporting from June to September, including the presentations to GB and USAID in late September. The draft report for GB was submitted in October and it was finalized in May 2014.

TABLE I: ROUND I TIMELINE (JANUARY 2013 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	X													
Administer EGRA					X												
Enter data						X	X										
Analyze baseline data							X	X	X								
Produce draft reports								X	X								
Produce presentations									X								
Disseminate draft reports									X								
Make presentations									X								
Revise and finalize reports																	X
Submit reports to USAID																	X

Sampling

The sampling for Round 1 started in January with the selection of the treatment districts by the provincial DOEs and USAID. The EGRA team conducted the school sampling in March and April. This included developing the sampling requirements, verifying the sample in the field, and finalizing the sample. As mentioned above, five of the districts in GB were full treatment and the other two were light treatment. The sampling for GB, as detailed in the sampling report for USAID¹, is briefly summarized in the following subsections of this report.

Sampling Requirements

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 treatment groups (full and light), equal numbers of male and female schools (35 each) were selected.

Sampling Process and Field Verification

Due to the need to balance representativeness, logistics, and resources, two of the five full treatment districts were randomly selected for the survey, along with both of the light treatment districts. This resulted in a clustered sample. For the 35 male and 35 female schools in each of the two groups, the samples were divided among the selected districts according to the proportions of schools within those districts (stratified random sampling). A second stratification was done at the “location” level, where schools were allocated by rural and urban. As seen in Tables 2 and 3 below, there were relatively few urban schools in GB (about 10 percent). After sampling the 140 schools in GB, an additional 10 male and 10 female schools per group were selected as replacements. 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).

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

TABLE 2: SAMPLE SCHOOLS BY DISTRICT, GENDER, AND LOCATION FOR FULL TREATMENT

District	Location	Schools	Pct.	Sample Schools		Replacement Schools	
				Boys	Girls	Boys	Girls
Full Treatment Group							
Gilgit	Rural	292	20	7	7	2	2
Gilgit	Urban	127	9	3	3	1	1
Skardu	Rural	941	66	23	23	6	6
Skardu	Urban	78	5	2	2	1	1
Total		1,438	100	35	35	10	10

District	Location	Schools	Pct.	Sample Schools		Replacement Schools	
				Boys	Girls	Boys	Girls
Light Treatment Group							
Ghanche	Rural	325	49	17	17	5	5
Ghanche	Urban	29	4	2	2	0	0
Ghizer	Rural	265	40	14	14	4	4
Ghizer	Urban	48	7	2	2	1	1
Total		667	100	35	35	10	10

Total (both groups)		2,105		70	70	20	20
----------------------------	--	--------------	--	-----------	-----------	-----------	-----------

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 National Education Management Information System (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 student numbers were near the minimum.

Intended and Actual Samples

Seven schools – two male and five female – were replaced due to lower than expected numbers of children in the original samples. 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 GB, were presented in a report to USAID.² This report is available to provincial education officials.

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

Trans-adaptation

In February, the EGRA team used tasks from the EGRA core instrument along with additional tasks used in instruments 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 Ministries of Education (MOEs) and Teacher Training Institutes (TTIs) throughout Pakistan – including two subject specialists from GB – 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 2013, the EGRA English and Urdu tools were piloted in selected schools in AJK, ICT, and KP provinces (with the Sindhi tools piloted in June). Four tools were included in the pilot: 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 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 with a limited number of experts from the trans-adaptation workshop. 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 in April. USAID made suggestions, particularly around the inclusion of reading- and library-related items into the questionnaires that would provide baseline information for the PRP and SRP. The instruments were approved and then used in the training workshops in advance of the Round 1 data collection in May. The final instruments were comprised of the following:

- Students: 16 informational items; 8 tasks (one of which has 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.

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

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 data entry. In April, the Institute for Social and Applied Policy Studies (I-SAPS) was chosen for both activities. A joint team from MSI, STS, and I-SAPS collaborated on the data collection in GB.

Data Collection

In April 2013, EGRA senior managers (from STS and MSI) trained MTs and QCOs during a two-week session in Islamabad. The MTs then spent one week, also in Islamabad, training the I-SAPS GB data collection team, which was comprised of a regional coordinator, four field supervisors, and 64 enumerators. The GB team was trained alongside the teams from AJK and ICT. The QCOs, coordinator, supervisors, and enumerators organized the logistics for the data collection. Following the training and logistical preparations in Islamabad, the QCOs and field supervisors conducted a two-day refresher course for the enumerators in Gilgit just prior to commencing data collection in the schools.

Over a 10-day period in May, the enumerators spent a day in each of the 140 schools to collect the baseline data in GB. The enumerators were in regular communication with 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.

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 June, the team trained the I-SAPS data coordinator, two supervisors, and 30 data entry operators (DEOs) on the application, with additional hands-on training using actual data from AJK, GB, and ICT. In June and July, the EGRA and I-SAPS teams did the data entry for over 10,000 student booklets, along with the questionnaires for the students, teachers, and head teachers. This included approximately 4,200 booklets for GB.

Data Cleaning

In July, the EGRA and I-SAPS 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, the EGRA statisticians and psychometrician (from STS) 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 by grade level and gender, and for the three sets of questionnaire data.

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 in GB for the tests in Urdu.

TABLE 3: RELIABILITY ESTIMATES

Language	Grade Level	Tasks	N-count	Alpha
Urdu	Grade 3	9	1,827	0.82
	Grade 5	9	1,898	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 was 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 Balochistan students' mastery of the tasks, and third, Balochistan 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 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 GB. There are sections on the student sample, task and item statistics, score calculation, task and summary scores, timed task scores, and questionnaire findings.

Student Sample

Table 7 shows the number of students in the sample by gender. For grades 3 and 5, the actual samples were 87.0 and 90.4 percent of the intended sample, respectively. For boys, the actual sample was 92.3 percent and for girls, the actual sample was 85.0 percent. A small number of students in grade 3 (n = 2) and grade 5 (n = 1) did not complete the gender item on the questionnaire. The total actual sample in GB was 3,725 students, or 88.7 percent of the intended sample.

TABLE 7: ACTUAL STUDENT SAMPLE BY GRADE AND GENDER

Treatment	Grade Level	Sample	Boys	Girls	Missing	Total
Full Treatment	Grade 3	Students	470	509	0	979
		% of Target	89.5%	97.0%	--	93.2%
	Grade 5	Students	509	471	1	981
		% of Target	97.0%	89.7%	--	93.4%
	Total	Students	979	981	1	1,960
		% of Target	99.70%	98.00%	--	93.3%
Light Treatment	Grade 3	Students	473	373	2	848
		% of Target	90.1%	71.0%	--	80.8%
	Grade 5	Students	486	431	0	917
		% of Target	92.6%	82.1%	--	87.3%
	Total	Students	959	804	2	1,765
		% of Target	99.70%	98.00%	--	84.0%
Full and Light Treatment	Total	Students	1,938	1,784	3	3,725
		% of Target	92.3%	85.0%	--	88.7%

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 good quality.

In GB, the task p-values for grade 3 ranged from 0.05 to 0.44, thus providing a spread on the lower half of the difficulty spectrum. The p-values for grade 5 ranged from 0.24 to 0.61, or in the lower and middle parts of the spectrum. All of the task scores in grades 3 and 5 had item-total correlations of greater than 0.20, indicating good quality for these tasks.

TABLE 8: TASK STATISTICS (FULL AND LIGHT TREATMENT GROUPS)

Task (Subtest)	Grade 3		Grade 5	
	P-Value	Item-Total	P-Value	Item-Total
1. Orientation to print (untimed)	0.44	0.22	0.51	0.34
2. Letter name recognition (timed)	0.35	0.56	0.50	0.59
3. Phonemic awareness (untimed)	0.33	0.29	0.41	0.31
4. Letter sound knowledge (timed)	0.20	0.59	0.34	0.56
5. Familiar word reading (timed)	0.22	0.75	0.61	0.76
6. Non-word reading (timed)	0.10	0.73	0.30	0.71
7a. Passage reading (timed)	0.26	0.76	0.61	0.77
7b. Passage comprehension (untimed)	0.05	0.60	0.24	0.68
8. Listening comprehension (timed)	0.16	0.33	0.35	0.40

The full item statistics for each of the items in the untimed tasks are provided in Annex 1 at the end of this report. Note that the “items” in the timed tasks (e.g., letters, sounds, and words) are not appropriate for these types of calculations for two reasons: 1) they are too numerous (up to 100 letters or sounds within a task), and 2) data were not collected on the individual letters, sounds, or words within the timed tasks.

Task and Summary Scores

The next part of the analysis involved plotting the scores. Since the idea with plotting the scores is to show the distributions at baseline, the scores for the full and light treatment groups are plotted together. 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. In addition, as with the task and item statistics, it also shows that there is room for growth at each grade level. The main goal of the intervention is to see movement of the score distributions to the right within the same grade level (i.e., grades 3 and 5) from the baseline to midline to endline.

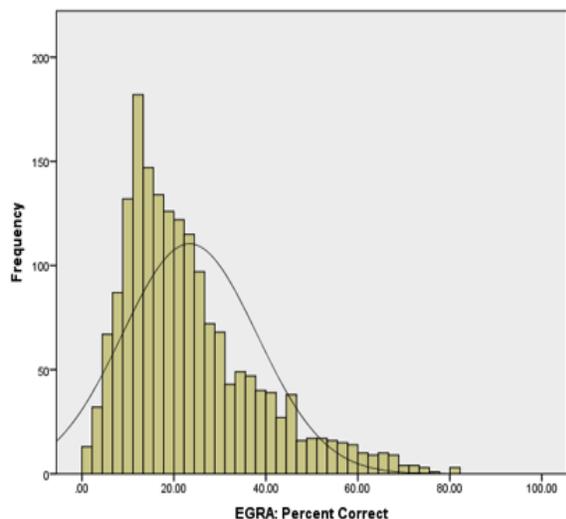
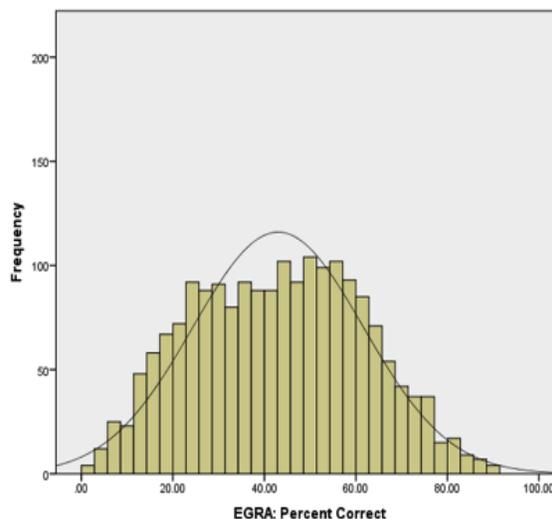
FIGURE 2: GRADE 3 SUMMARY SCORES**FIGURE 3: GRADE 5 SUMMARY SCORES**

Table 9 and 10 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 Gilgit-Baltistan was 43.9 percent for grade 3 and 50.9 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., 50.9 percent minus 43.9 percent equals 7.1 percentage points.

Grade 3 posted the highest scores in orientation to print, followed by letter name recognition and phonemic awareness. The most difficult tasks for these students were comprehension (passage and listening) and phonics (non-word reading and letter sound knowledge). At grade 5, the highest scores were in familiar word reading and passage reading, followed by orientation to print and letter name recognition; the most challenging tasks were comprehension (passage and listening), non-word reading, and letter sound knowledge.

There was also substantial progression from grade 3 to grade 5 on the summary score (20 points). The greatest gains were in familiar word reading (38 points) and passage reading (35 points). In areas where there are small differences between scores in grades 3 and 5, interventions at grade 3 could have particularly large effects in accelerating children's learning.

TABLE 9: PERCENT CORRECT SCORES BY GRADE AND TASK (FULL AND LIGHT TREATMENT GROUPS)

Task (Subtest)	Grade 3	Grade 5	Difference (G5 – G3)
1. Orientation to print	43.9%	50.9%	7.1% points
2. Letter name recognition	35.1%	49.8%	14.7% points
3. Phonemic awareness	32.4%	40.9%	8.5% points
4. Letter sound knowledge	19.9%	34.2%	14.3% points
5. Familiar word reading	22.2%	60.5%	38.3% points
6. Non-word reading	10.1%	30.2%	20.1% points
7a. Passage reading	25.9%	60.7%	34.8% points
7b. Passage comprehension	4.8%	24.1%	19.3% points
8. Listening comprehension	15.6%	35.1%	19.5% points
Reading Summary Score	23.4%	42.9%	19.6% points

Table 10 and Figures 4 and 5 provide the scores of the full and light treatment students by task, and for the summary (or grand mean), using the percent correct metric for all grade 3 and 5 students (i.e., for the full and light treatment groups combined). For both groups, the students in grade 3 demonstrated relatively better skills in orientation to print, letter name recognition, and phonemic awareness. They had lower skills in phonics (e.g., letter sound knowledge, non-word reading) and comprehension (reading and listening). Grade 5 students showed the strongest increases in the reading areas – familiar words and passages. They still had relatively low scores in letter sound knowledge, non-word reading, and comprehension. The full and light treatment groups had similar scores, though full treatment was a few points higher; this will be corrected statistically at the midline.

TABLE 10: SCORES BY GRADE, TASK, AND GROUP

Task (Subtest)	Full Treatment		Light Treatment	
	Grade 3	Grade 5	Grade 3	Grade 5
1. Orientation to print	43.3%	51.3%	44.4%	50.5%
2. Letter name recognition	37.5%	53.1%	32.7%	46.5%
3. Phonemic awareness	36.0%	46.1%	28.8%	35.6%
4. Letter sound knowledge	21.1%	37.4%	18.7%	30.9%
5. Familiar word reading	25.6%	64.1%	18.7%	56.8%
6. Non-word reading	11.8%	34.4%	8.4%	26.0%
7a. Passage reading	29.1%	64.6%	22.6%	56.7%
7b. Passage comprehension	5.6%	25.6%	4.0%	22.5%
8. Listening comprehension	15.5%	36.9%	15.7%	33.3%
Reading Summary Score	25.1%	45.9%	21.6%	39.9%

FIGURE 4: FULL TREATMENT SCORES BY GRADE AND TASK

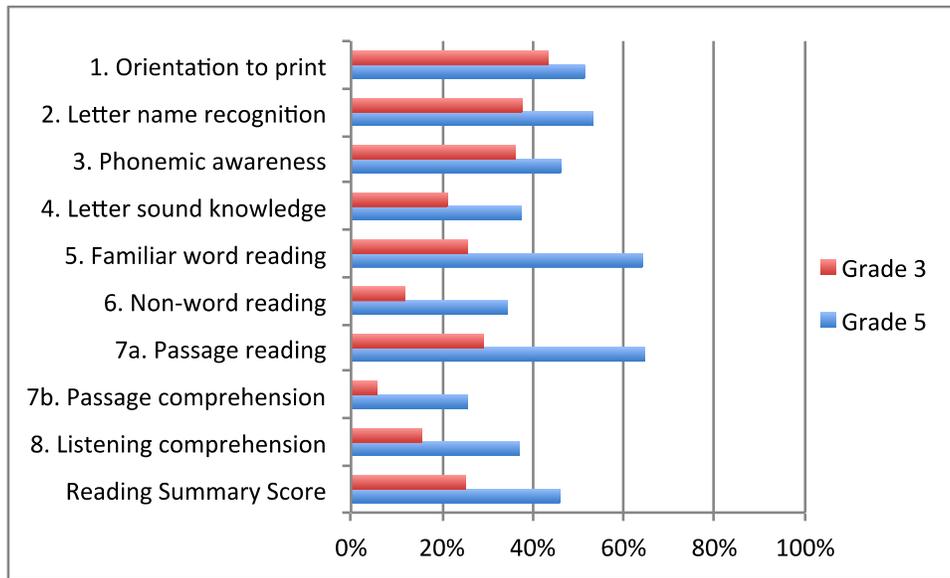
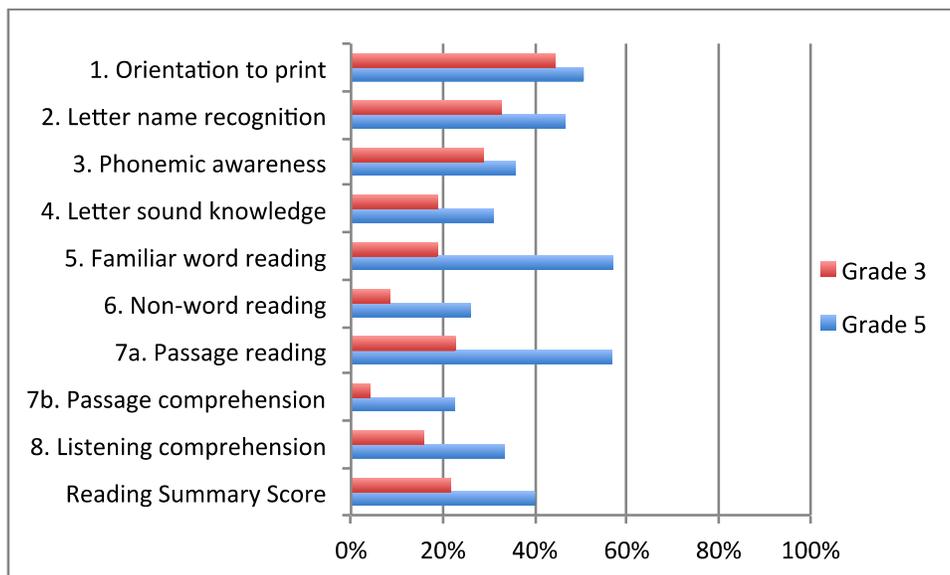


FIGURE 5: LIGHT TREATMENT SCORES BY GRADE AND TASK



When the scores were disaggregated by gender (Table 11 and Figures 6 and 7), again the scores across groups are combined to show the general situation at baseline. There were differences between boys and girls on the task and summary scores, but most of these differences were small. There were larger differences in favor of the girls in familiar word reading and passage reading at grade 3, and in passage comprehension at grade 5. The gains from grade 3 to grade 5 were about the same for the boys and girls, with each increasing by about 20 points.

TABLE 11: SCORES BY GRADE, TASK, AND GENDER (FULL AND LIGHT TREATMENT GROUPS)

Task (Subtest)	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
1. Orientation to print	45.9%	41.6%	52.8%	48.9%
2. Letter name recognition	34.2%	35.9%	48.8%	51.0%
3. Phonemic awareness	31.7%	33.0%	41.8%	39.9%
4. Letter sound knowledge	19.8%	20.2%	35.2%	33.0%
5. Familiar word reading	18.9%	25.7%	58.2%	63.0%
6. Non-word reading	8.9%	11.5%	30.1%	30.4%
7a. Passage reading	22.8%	29.2%	58.7%	62.9%
7b. Passage comprehension	3.6%	6.1%	21.0%	27.5%
8. Listening comprehension	16.1%	15.2%	35.8%	34.2%
Reading Summary Score	22.4%	24.2%	42.5%	43.4%

FIGURE 6: GRADE 3 SCORES BY TASK AND GENDER (FULL AND LIGHT TREATMENT GROUPS)

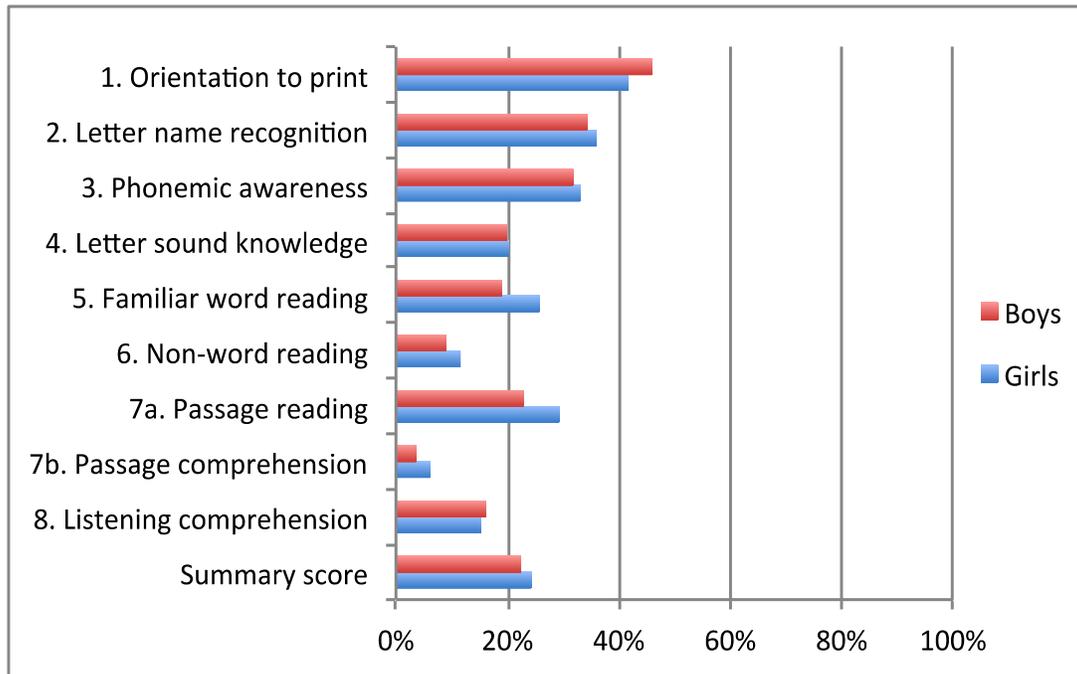
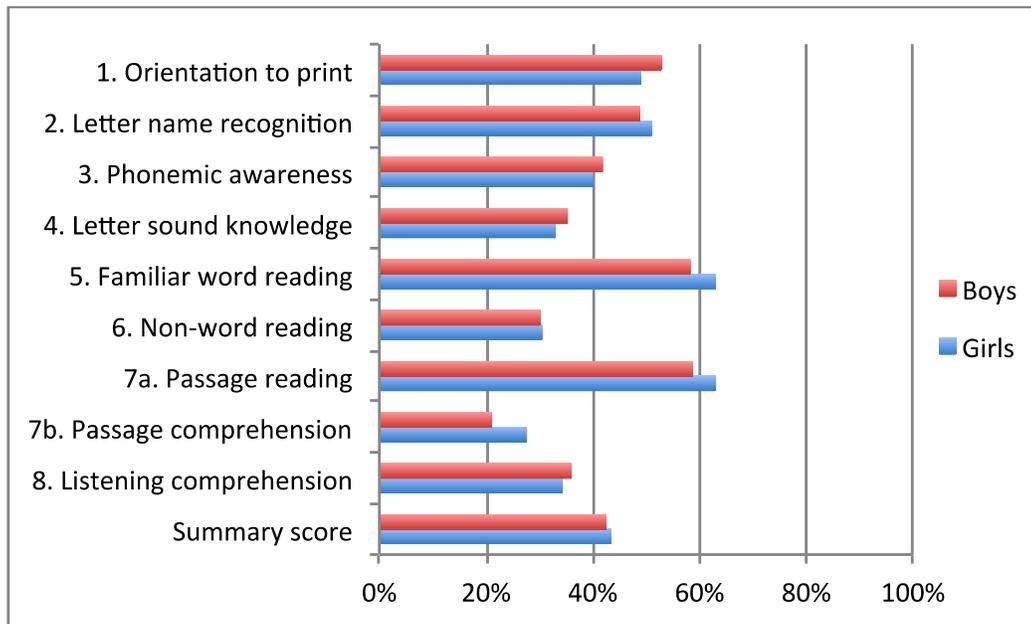


FIGURE 7: GRADE 5 SCORES BY TASK AND GENDER (FULL AND LIGHT TREATMENT GROUPS)



The final table in this section (Table 12) further disaggregates the scores by treatment group, grade level, and gender. As seen in the tables above, the light treatment group scored higher on some of the tasks, which will be statistically corrected at the midline and endline. There were some variations in the scores by gender and treatment group. For instance, on many of the tasks, the girls scored higher than the boys in the full treatment group, but the boys scored higher than the girls in the light treatment group. Further investigation would be required to determine the reasons for this trend.

TABLE 12: PERCENT CORRECT SCORES BY GROUP, GRADE, AND GENDER

Task (Subtest)	Full Treatment				Light Treatment			
	Grade 3		Grade 5		Grade 3		Grade 5	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
1. Orientation to print	44.6%	42.2%	53.4%	49.1%	47.1%	41.0%	52.1%	48.7%
2. Letter name recognition	36.0%	38.8%	50.8%	55.6%	32.4%	33.0%	46.7%	46.3%
3. Phonemic awareness	34.4%	37.4%	45.8%	46.5%	28.9%	28.7%	37.7%	33.3%
4. Letter sound knowledge	21.9%	20.3%	38.1%	36.7%	17.7%	20.0%	32.4%	29.2%
5. Familiar word reading	22.2%	28.7%	62.2%	66.1%	15.5%	22.6%	54.1%	59.8%
6. Non-word reading	10.8%	12.5%	34.2%	34.6%	6.9%	10.4%	25.9%	26.1%
7a. Passage reading	25.7%	32.2%	62.7%	66.6%	19.8%	26.1%	54.6%	59.1%
7b. Passage comprehension	3.8%	7.2%	21.3%	30.3%	3.3%	4.9%	20.6%	24.7%
8. Listening comprehension	16.1%	14.9%	36.3%	37.5%	16.0%	15.4%	35.3%	30.9%
Reading Summary Score	24.0%	26.0%	45.0%	47.0%	20.8%	22.4%	39.9%	39.8%

Timed Task Scores

Fluency is a measure of reading efficiency. On the Pakistan EGRA Baseline, 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 14 below show scores in terms of raw scores (instead of the percent correct scores on the previous tables). Table 13 has the maximum raw scores attained by students on each task at each grade level. Tables 12 to 14 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). Because these calculations are different from percent correct, the maximum scores are higher (see Figures A1 and A2 in Annex 2). Table 13 provides the baseline maximum scores at grade 3 and 5 for the five timed tasks.

TABLE 13: BASELINE MAXIMUM SCORES ON FLUENCY (TIMED) TASKS (FULL AND LIGHT TREATMENT GROUPS)

Phonics Fluency Subtest	Grade 3	Grade 5
2. Letter name recognition	99	107
4. Letter sound knowledge	85	121
6. Non-word reading	112	82
Reading-Rate Fluency Subtest	Grade 3	Grade 5
5. Familiar word reading	111	150
7a. Passage reading	203	205

As shown in Table 14, students at grade 3 and 5 showed similar patterns in phonics and reading-rate fluency. In terms of phonics, students read more letters than non-words in the given time frame. The phonics fluency tasks were more challenging than the reading-rate fluency tasks. The non-word rates were lower than the familiar word and the passage reading rates. The greatest gains from grade 3 to 5 were in reading-rate fluency tasks. In contrast, the lack of growth in phonics fluency should be a target for instruction. Table 12 also shows the difference between grades, i.e., the progression from grade 3 to grade 5. The general term “points” was used to designate letters, sounds, words, or non-words.

TABLE 14: PHONICS AND READING-RATE FLUENCY TASK MEANS BY GRADE (FULL AND LIGHT TREATMENT GROUPS)

Phonics Fluency Subtest	Grade 3	Grade 5	Difference (G5 – G3)
2. Letter name recognition	35.1	49.9	14.8 points
4. Letter sound knowledge	19.9	34.4	14.5 points
6. Non-word reading	5.3	16.5	11.2 points
Reading-Rate Fluency Subtest	Grade 3	Grade 5	Difference (G5 – G3)
5. Familiar word reading	16.8	41.7	24.9 points
7a. Passage reading	18.3	51.4	33.1 points

For both full and light treatment (Table 15), the highest scores on the timed tasks at grade 3 were in letter name recognition. The other tasks had lower scores. At grade 5, the highest scores were in letter name recognition and passage reading. The areas of greatest progress from grade 3 to grade 5 were two of the reading tasks: familiar word reading and passage reading. Non-word reading did not show much progress from grade 3 to grade 5.

TABLE 15: TIMED TASK SCORES BY GRADE AND GROUP

Phonics Fluency Subtest	Grade 3		Grade 5	
	Full	Light	Full	Light
2. Letter name recognition	37.5	32.7	53.1	46.7
4. Letter sound knowledge	21.1	18.7	37.7	31.0
6. Non-word reading	6.0	4.5	18.6	14.3
Reading-Rate Fluency Subtest	Grade 3		Grade 5	
	Full	Light	Full	Light
5. Familiar word reading	14.6	11.2	44.3	37.7
7a. Passage reading	20.5	16.3	56.1	46.7

By gender (Table 16), the scores across groups are again combined to show the general situation at baseline. The differences between boys and girls were small. The grade 3 girls had higher scores on familiar word reading and passage reading. The other tasks showed similar scores by gender. The grade 5 girls had higher scores in passage reading, while the scores on the other tasks were similar.

TABLE 16: TIMED TASK SCORES BY GRADE AND GENDER (FULL AND LIGHT TREATMENT GROUPS)

Phonics Fluency Subtest	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
2. Letter name recognition	34.2	36.4	49.0	51.5
4. Letter sound knowledge	19.7	20.1	35.3	33.5
6. Non-word reading	4.5	6.1	16.1	16.9
Reading-Rate Fluency Subtest	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
5. Familiar word reading	10.4	15.5*	39.9	43.8
7a. Passage reading	15.5	21.4*	47.6	55.9*

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

The final table in this section (Table 17) further disaggregates the scores by treatment group, grade level, and gender. As with the percent correct scores, the light treatment group scored higher on some of the tasks, which will be statistically corrected at the midline and endline.

TABLE 17: PHONICS AND READING-RATE FLUENCY TASK MEANS BY GROUP, GRADE, AND GENDER

Phonics Fluency Subtest	Full Treatment				Light Treatment			
	Grade 3		Grade 5		Grade 3		Grade 5	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
2. Letter name recognition	36.0	38.9	51.0	56.0	32.4	33.1	46.9	46.5
4. Letter sound knowledge	21.9	20.3	38.1	37.2	17.6	19.9	32.4	29.4
6. Non-word reading	5.5	6.4	18.2	18.9	3.5	5.6	13.9	14.7
Reading-Rate Fluency Subtest	Full Treatment				Light Treatment			
	Grade 3		Grade 5		Grade 3		Grade 5	
	Boys	Girls	Boys	Girls	Boys	Girls	Boys	Girls
5. Familiar word reading	12.3	16.6	42.2	46.4	8.5	13.9	34.8	41.0
7a. Passage reading	17.6	23.0	52.4	59.8	13.3	19.3	42.6	51.4

Questionnaire Findings

Selected results are presented below, including for those characteristics or items that showed significant differences in student scores. Due to the students having the same language, the results were combined for the full and light treatment groups to increase the sample size and more accurately detect effects between the categories. 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 those who responded.

Since these are baseline data, reporting on the full and light treatment groups together will not affect the analyses at midline and endline. We combined the survey data for the groups since some of the questions led to reporting by relatively small categories (e.g., for teacher qualifications) and we wanted to know whether the survey results were associated with the student scores in general.

In addition, since the samples were by treatment group, the results will be generalized to the populations for each group. This will be done prior to the midline. The results will be generalized to by calculating sampling weights, applying the weights to the results, and then generalizing to the population by treatment group. We will also do this for the midline and endline. The current analyses only apply to the sampled districts.

Statistical significance was determined based on *t*-tests for indicators with two categories and analyses of variance for indicators with three or more categories (with post hoc pairwise comparisons).

Student Questionnaires

Table 18 shows the 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. At grade 3, there were no significant differences based on student age group; at grade 5, the scores for the normal age students were significantly higher than the scores for the older students. (Note that there was no significant difference between the younger and older students due to a small sample size for the younger students.)

TABLE 18: SUMMARY SCORES BY STUDENT AGE

Age Group	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Younger than normal age	32	21.0%	41	45.9%
Normal age	506	24.1%	442	45.4%
Older than normal age	1,283	23.2%	1,411	42.2%
Missing	6	--	4	--
Total	1,827	23.4%	1,898	43.0%*

* Indicates that the performance of a group was significantly higher, $p < 0.05$ level.

Table 19 shows the summary scores according to whether the student reads the Quran at home. There were significant differences in both grades in favor of students who read the Quran.

TABLE 19: SUMMARY SCORES BY READING THE QURAN AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	213	19.7%	138	32.6%
Yes	1,610	23.9%*	1,760	43.8%*
Missing	4	--	0	--
Total	1,827	23.4%	1,898	43.0%

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

Table 20 shows the differences in scores based on whether there is a library at the school. While the scores were slightly higher at grade 3 for those students who said that there is a library at their school, there were no significant differences at either grade level.

TABLE 20: 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	1,152	22.9%	1,182	43.1%
Yes	499	24.5%	580	42.9%
Missing	176	--	136	--
Total	1,827	23.4%	1,898	43.0%

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

In Tables 21 to 23, the data showed that the existence of newspapers and magazines generally made a difference in reading scores in most cases. On the other hand, the effect of the presence of books at home on scores was positive but not statistically significant.

TABLE 21: SUMMARY SCORES BY THE PRESENCE OF NEWSPAPERS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	1,449	22.9%	1,302	41.4%
Yes	378	25.5%*	596	46.7%*
Missing	0	--	0	--
Total	1,827	23.5%	1,898	43.0%

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

TABLE 22: SUMMARY SCORES BY THE PRESENCE OF MAGAZINES AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	1,759	23.3%	1,801	42.6%
Yes	68	26.6%	97	49.9%*
Missing	0	--	0	--
Total	1,827	23.5%	1,898	43.0%

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

TABLE 23: SUMMARY SCORES BY THE PRESENCE OF BOOKS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	955	23.1%	1,131	42.2%
Yes	872	23.9%	767	44.2%
Missing	0	--	0	--
Total	1,827	23.5%	1,898	43.0%

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

The final set of student questions (in Tables 24 to 26) pertained to children's reading habits at home. In general, these habits made a difference in student scores in all cases in grade 5, but only for silent reading at home in grade 3. Having someone read to children at home and having children read to someone else at home made a difference at grade 5 but not at grade 3. Having children read silently at home made a difference at grade 3 and at grade 5.

TABLE 24: 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	813	22.8%	857	42.0%
Yes	977	24.1%	1,030	44.1%*
Missing	37	--	11	--
Total	1,827	23.6%	1,898	43.1%

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

TABLE 25: SUMMARY SCORES BY CHILDREN READING TO SOMEONE ELSE AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	854	23.0%	769	41.8%
Yes	939	24.1%	1,118	44.0%*
Missing	34	--	11	--
Total	1,827	23.6%	1,898	43.1%

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

TABLE 26: SUMMARY SCORES BY CHILDREN READING SILENTLY AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	348	22.0%	376	41.5%
Yes	1,455	23.9%*	1,515	43.5%*
Missing	24	--	7	--
Total	1,827	23.5%	1,898	43.1%

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

Teacher Questionnaires

With the smaller sample size, the full and light treatment groups were again combined and the analysis of the teacher questionnaires was limited to descriptive statistics, i.e., no group comparisons. Tables 27 and 28 provide information on teacher academic and professional qualifications, neither of which showed consistent patterns in the student scores.

TABLE 27: SUMMARY SCORES BY TEACHER ACADEMIC QUALIFICATION

Academic Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.A./M.Sc./M.Phil.	13	24.4%	14	45.1%
B.A./B.Sc.	64	23.5%	63	43.8%
F.A./F.Sc.	20	24.6%	25	42.9%
Matric	15	19.0%	9	38.2%
Missing	1	--	1	--
Total	113	23.2%	112	43.3%

TABLE 28: SUMMARY SCORES BY TEACHER PROFESSIONAL QUALIFICATION

Professional Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.Ed./M.A.	5	25.8%	9	44.9%
B.Ed.	45	24.1%	42	42.7%
C.T.	29	22.1%	33	44.9%
P.T.C.	9	25.6%	5	48.0%
Missing	25	--	22	--
Total	113	23.7%	112	44.0%

In an analysis of student scores by teacher age and experience, there were no consistent patterns of younger or older teachers, or those with less or more experience, relating to higher or lower student scores (Tables 29 and 30). Again, small teacher sample sizes made drawing conclusions difficult.

TABLE 29: SUMMARY SCORES BY TEACHER AGE

Age Group in Years	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
40 and less	84	23.5%	79	43.2%
Between 41 and 50	20	24.6%	19	44.1%
51 and more	7	21.5%	12	41.5%
Missing	2	--	2	--
Total	113	23.6%	112	43.1%

TABLE 30: SUMMARY SCORES BY TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
10 or less	74	23.5%	66	43.8%
Between 11 and 20	26	22.7%	26	42.2%
Between 21 and 30	8	23.9%	13	40.5%
31 or more	5	24.8%	7	47.1%
Missing	0	--	0	--
Total	113	23.3%	112	43.0%

For frequency of in-service training, there were also no clear patterns (Table 31). Again, any differences should be interpreted with caution due to the small sample size.

TABLE 31: SUMMARY SCORES BY TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	67	23.6%	71	41.8%
One time	36	22.4%	28	45.8%
Two times	6	24.6%	10	45.9%
Three times	4	28.1%	2	50.5%
Missing	0	--	1	--
Total	113	23.3%	112	43.2%

Head Teacher Questionnaires

Similarly to the teachers, the sample size for the head teacher questionnaires was small, so data interpretations should be treated with caution. Tables 32 and 33 show head teacher academic and professional qualifications. In general, the results show that higher teacher academic and professional qualifications are related to better student scores.

TABLE 32: 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./M.Phil.	52	24.1%	52	44.2%
B.A./B.Sc.	64	23.6%	64	42.4%
F.A./F.Sc.	16	20.3%	16	40.6%
Matric	7	17.5%	7	38.4%
Missing	1	--	1	--
Total	140	23.1%	140	42.7%

TABLE 33: 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.	22	25.4%	22	45.8%
B.Ed.	86	23.9%	86	43.2%
C.T.	15	20.1%	15	39.6%
P.T.C.	6	18.7%	6	40.3%
Missing	11	--	11	--
Total	140	23.5%	140	43.1%

Tables 32 and 33 provide information on head teachers' experience and in-service training. The relationships between experience, training, and reading scores were inconsistent.

TABLE 34: SUMMARY SCORES BY HEAD TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
2 or less	52	23.5%	52	44.4%
3 to 5	25	22.9%	25	42.6%
6 to 10	23	23.4%	23	43.1%
11 or more	37	22.8%	37	40.3%
Missing	3	--	4	--
Total	140	23.2%	140	42.8%

TABLE 35: SUMMARY SCORES BY HEAD TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	85	22.7%	85	41.1%
1 time	42	23.0%	42	44.5%
2 times	7	22.9%	7	41.8%
More than 2 times	4	26.2%	4	48.2%
Missing	2	--	2	--
Total	140	22.9%	140	42.4%

Tables 36 and 37 provide data on head teachers' support to teachers in reading and the training that head teachers received in teaching reading. The data consistently associated higher reading scores with more head teacher support to teachers in reading and to head teacher training in teaching reading.

TABLE 36: 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	12	16.2%	12	34.2%
Yes	128	23.7%	128	43.4%
Missing	0	--	0	--
Total	140	23.1%	140	42.6%

TABLE 37: 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	58	20.8%	58	38.7%
Yes	81	24.8%	81	45.3%
Missing	1	--	1	--
Total	140	23.1%	140	42.6%

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, most patterns appeared to be inconclusive (Tables 38 to 41). The male and female schools performed better than the mixed schools. Better infrastructure seemed to have a positive relationship with student scores.

TABLE 38: SUMMARY SCORES BY SCHOOL GENDER

School Gender	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Male school	39	25.1%	39	45.4%
Female school	26	24.5%	26	44.4%
Mixed school	74	21.6%	74	40.5%
Missing	1	--	1	--
Total	140	23.1%	140	42.6%

TABLE 39: SUMMARY SCORES BY PTA/SMC/PTSMC/PTC

Parent Teacher Committee	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	9	19.5%	9	41.8%
Yes	131	23.4%	131	42.7%
Missing	0	--	0	--
Total	140	23.1%	140	42.6%

TABLE 40: SUMMARY SCORES BY PRESENCE OF A SCHOOL LIBRARY

School Library	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	106	22.5%	106	41.5%
Yes	34	25.1%	34	46.1%
Missing	0	--	0	--
Total	140	23.1%	140	42.6%

TABLE 41: 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	16	20.2%	16	37.5%
1	33	21.4%	33	41.0%
2	37	21.9%	37	41.7%
3	54	25.8%	54	45.5%
Missing	0	--	0	--
Total	140	23.1%	140	42.6%

CHAPTER 4: CONCLUSIONS AND RECOMMENDATIONS

This final chapter provides conclusions and recommendations from the GB EGRA baseline. The conclusions are 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. This will allow for an examination of the progress of students in grades 3 and 5 over the life of the PRP. In addition, GB has two treatment groups: full and light. This will allow for an evaluation of the full treatment effects above and beyond those of the light treatment.
2. The sampling issues were addressed as well as could have been expected. In a limited number of schools, there was an issue of a lack of the requisite number of students per grade level. However, the actual sample of schools was 100 percent and the actual sample of students reached 88.7 percent of the intended sample.
3. The EGRA test was of good quality. The reliability estimates were in the high part of 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 an acceptable level of quality. The characteristics of the test were such that it should be sensitive to potential 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 difficult terrain in GB. There was a high level of standardization reported by the quality control officers, 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

The GB evaluation involves two kinds of analyses: 1) a comparison of full and light treatment groups to determine the effects of full treatment above and beyond that of the light treatment, and 2) a comparison of each group to itself at the baseline, midline, and endline.

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

1. EGRA was administered to a robust sample at each grade level (3 and 5) and in each group (full and light treatment). Test reliabilities were 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. The students in grade 3 were strongest in orientation to print, letter name recognition, and phonemic awareness. Their scores were relatively low in phonics (non-word reading and letter sound knowledge) and comprehension (passage reading and listening). In grade 5, the students' best scores were on the two reading tasks (familiar word and passage). Scores in phonics (non-word reading and letter sound knowledge) were still low at grade 5.

3. On most tasks, however, there was substantial progress from grade 3 to grade 5, particularly in the areas of familiar word reading and passage reading.
4. Male and female students had similar scores in GB, both on the tasks and the summary scores. This was the pattern across the two grade levels. The girls were slightly higher than the boys on the main reading tasks – familiar word reading and passage reading. The boys were slightly higher on orientation to print and listening comprehension. The summary scores for the two groups were similar.
5. Scores for the full and light treatment groups were somewhat different, most likely due to preselecting the districts that comprised the two treatment groups rather than assigning the districts to treatment group at random. The differences in scores will need to be corrected statistically during the midline analysis.
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. Although the passage was designed for grade 3, this difference shows that the fluency levels in grade 3 are low, but that students can make substantial progress in the early grades if expectations are high enough and if they are provided with the opportunity to learn. Specifically, mastery of phonics, such as letter sound knowledge and non-word reading, should help the students become better overall readers. It is clear that these types of knowledge and skills are not receiving an appropriate emphasis in schools in GB.
7. Questionnaire findings were mostly inconclusive, due to small sample sizes and the lack of differences in responses within the student, teacher, and head teacher samples. For the students, attending a grade at an appropriate age, or even younger, seemed to have a positive effect on reading outcomes. In terms of the home environment, the presence of reading materials seemed to have a small positive effect on children’s reading levels. It was the same with having a person to read with, with more of a positive effect for the younger children.
8. The teacher, head teacher, and school variables were often not related to student scores, although there were some exceptions. For the students, one of the positive findings was that having reading materials and opportunities to read in the home seemed to have a positive effect on reading outcomes. For the teachers, higher qualifications, both academic and professional, were associated with higher student reading scores. For head teachers, providing support to teachers in reading instruction tended to relate to higher reading scores for students. For the schools, the presence of a library and better infrastructure were associated with better student reading scores.

Evaluation Recommendations

Given the success of the baseline assessment in GB (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 instrument development and trans-adaptation process was comprehensive and resulted in high quality EGRA tools. This should be repeated as soon as possible with the tasks that need to be changed for the midline and endline tools (to minimize test-retest effects and security breaches), so that reading 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. To accurately measure these gains in the future, the testing needs to occur at a consistent point in the academic year. Midline and endline testing in GB should occur in May, thus matching the baseline timeframe and standardizing the instructional time across the study.
5. The systems developed for field data collection should be repeated. The different layers of management, coordination, supervision, and quality control contributed to successful planning, implementation, and problem solving. The quality control officers were particularly important in maintaining standards and providing support for the local subcontractors.
6. 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.
7. 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.
8. 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.
9. 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 GB 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) spread across the range from about 0.05 to 0.90 and quality values (item-total correlations) of at least 0.20. The difficulties ranged from 0.02 to 0.62 for grade 3 and 0.10 to 0.69 for grade 5, indicating item difficulties slightly below the recommended range. A total of 21 and 22 items for grades 3 and 5 respectively out of the 23 items had item-total correlations of at least 0.20, indicating high quality items.

TABLE AI: COMPLETE ITEM STATISTICS BY GRADE

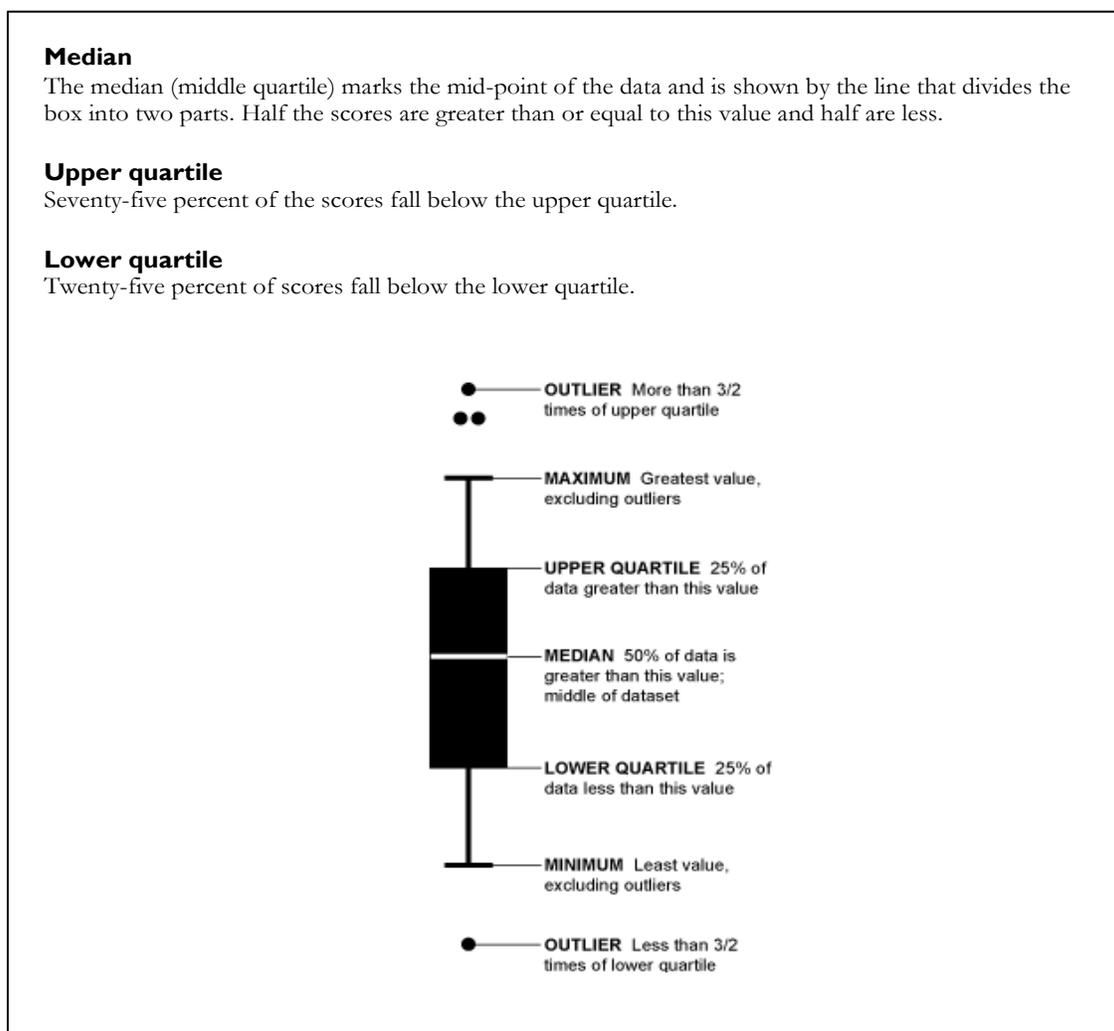
Task (Subtest)	Item	Grade 3		Grade 5	
		P-Value	Item-Total	P-Value	Item-Total
1. Orientation to print (untimed)	Q1	0.61	0.37	0.65	0.34
	Q2	0.62	0.40	0.69	0.36
	Q3	0.38	0.29	0.39	0.23
	Q4	0.08	0.09	0.18	0.15
	Q5	0.50	0.29	0.64	0.31
2. Letter name recognition (timed)	--				
3. Phonemic awareness (untimed)	Q1	0.52	0.35	0.63	0.39
	Q2	0.28	0.33	0.42	0.43
	Q3	0.31	0.29	0.38	0.28
	Q4	0.28	0.29	0.35	0.32
	Q5	0.30	0.27	0.42	0.33
	Q6	0.45	0.33	0.54	0.37
	Q7	0.17	0.23	0.21	0.31
	Q8	0.29	0.36	0.35	0.37
	Q9	0.19	0.27	0.23	0.28
	Q10	0.48	0.35	0.56	0.38
4. Letter sound knowledge (timed)	--				
5. Familiar word reading (timed)	--				
6. Non-word reading (timed)	--				
7a. Passage reading (timed)	--				
7b. Passage comprehension (untimed)	Q1	0.04	0.41	0.37	0.40
	Q2	0.04	0.37	0.38	0.36
	Q3	0.02	0.32	0.32	0.34
	Q4	0.06	0.47	0.48	0.48
	Q5	0.03	0.43	0.40	0.35
8. Listening comprehension (untimed)	Q1	0.13	0.30	0.28	0.31
	Q2	0.03	0.15	0.10	0.23
	Q3	0.31	0.32	0.67	0.31

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



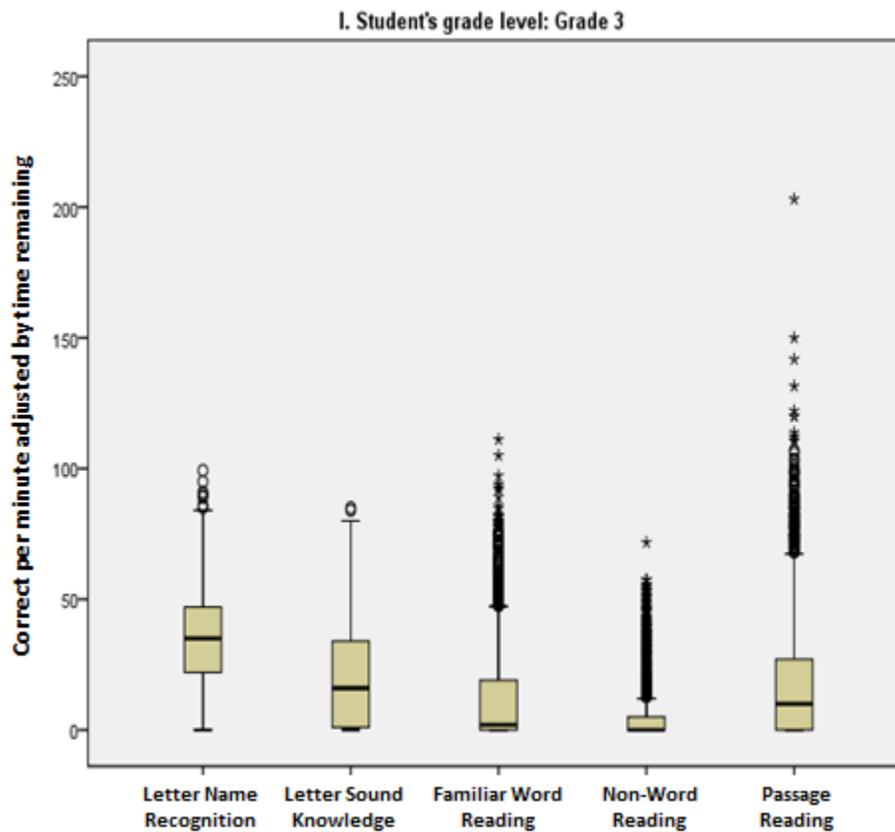
Box plots are presented below (Figures A2 and A3) for the results by grade level on the five fluency (timed) tasks: letter name recognition (task 2), letter sound knowledge (task 4), familiar word reading (task 5), non-word reading (task 6), and passage reading (task 7a).

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 0 (non-word reading) to about 40 (letter name recognition) items per minute. It shows that the students had much better knowledge of letter names than of grapheme-phoneme correspondence.

The variation (i.e., the range of scores, without outliers) for each of the tasks varied from about 10 (non-word reading) to about 70 (letter name recognition). It shows that the scores were more spread out when recognizing letter names than sounding out pseudo-words.

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



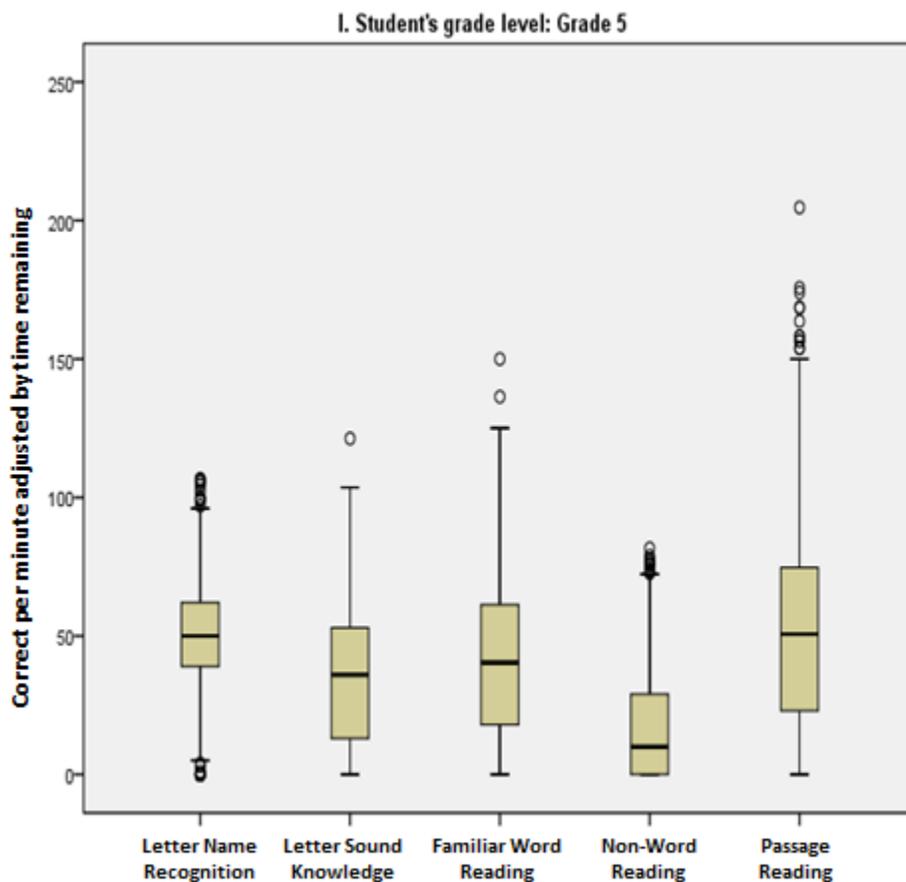
Grade 5

For grade 5, the central tendency (the median speed) for each of the tasks ranged from about 10 (non-word reading) to about 50 (passage reading) items per minute. It shows that the students had more fluency with reading connected words than conducting grapheme-morpheme correspondence.

The variation (range of scores) for each of the tasks varied from about 80 (non-word reading) to about 150 (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 71.4 (rounded to 71). With this method, 71 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 89.8 (rounded to 90). Then 90 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	36.7%	0	11.2%
Non-Fluent Reader	1 to 70	58.6%	1 to 89	73.3%
Fluent Reader	71 and above	4.7%	90 and above	15.5%
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	36.7%	11.2%
Early Reader	1 to 40	47.3%	29.6%
Intermediate Reader	41 to 80	13.1%	38.7%
Fluent Reader	81 to 120	2.7%	17.5%
Advanced Reader	121 and above	0.3%	3.0%
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 6 percent of the fluent readers have 80 percent comprehension. Also, at grade 3, 14 percent (6 percent + 8 percent) of the fluent readers have at least 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	89%	9%	2%	0%	0%	0%	100%
Intermediate Reader	41 to 80	53%	18%	16%	9%	4%	1%	100%
Fluent Reader	81 to 120	27%	15%	21%	23%	6%	8%	100%
Advanced Reader	121 and above	20%	40%	20%	20%	0%	0%	100%

FIGURE A4: GRADE 3 READING FLUENCY AND COMPREHENSION

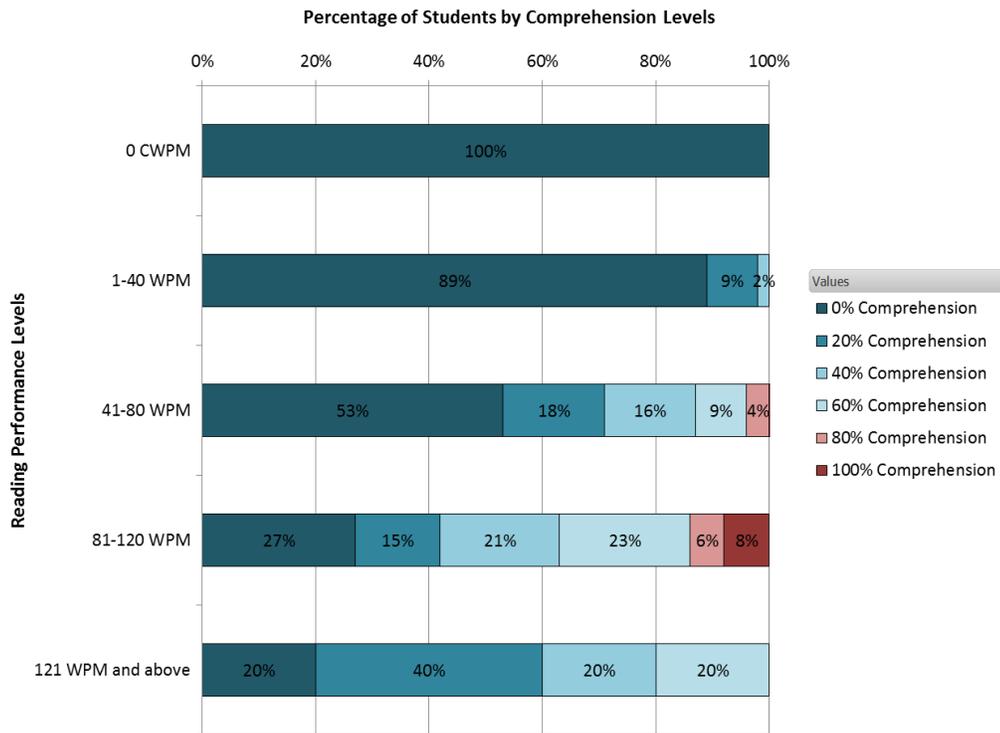
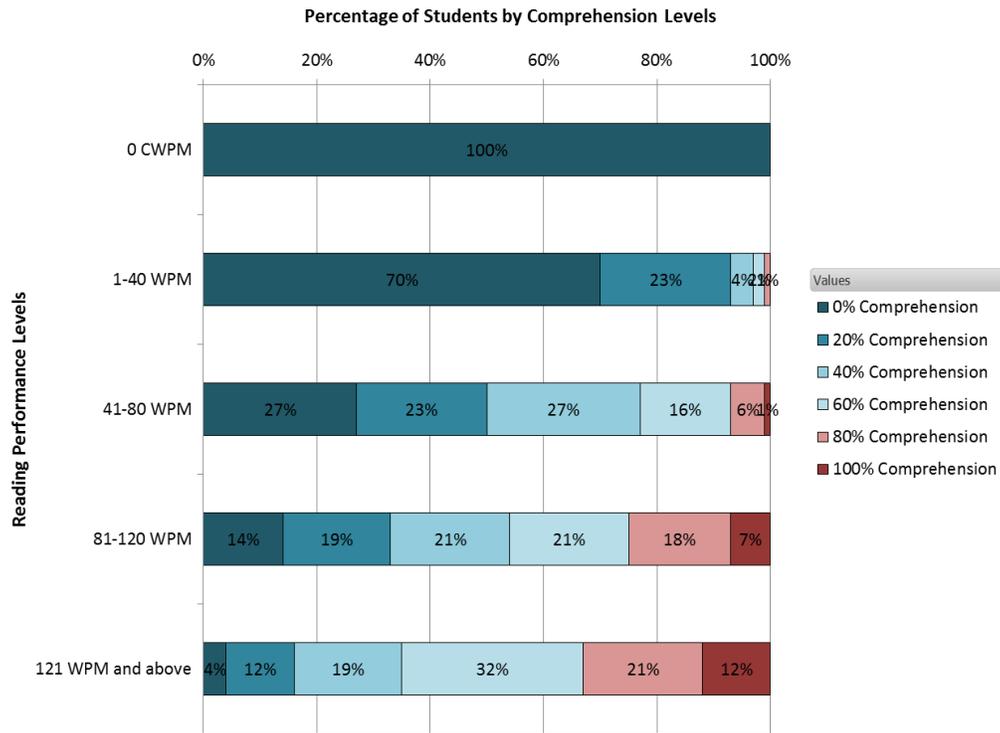


TABLE A5: GRADE 5 READING FLUENCY AND COMPREHENSION

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	70%	23%	4%	2%	1%	0%	100%
Intermediate Reader	41 to 80	27%	23%	27%	16%	6%	1%	100%
Fluent Reader	81 to 120	14%	19%	21%	21%	18%	7%	100%
Advanced Reader	121 and above	4%	12%	19%	32%	21%	12%	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 (89 percent at grade 3 and 70 percent at grade 5) had 0 percent comprehension. None of them achieved 80 percent comprehension.
- Intermediate Readers (41-80 CWPM) – A substantial percentage of intermediate readers (53 percent at grade 3 and 27 percent at grade 5) had 0 percent comprehension. A minority of them (5 percent at grade 3 and 7 percent at grade 5) achieved at least 80 percent comprehension.
- Fluent Readers (81-120 CWPM) – Among fluent readers, 27 percent in grade 3 and 14 percent in grade 5 had 0 percent comprehension. Less than one-quarter of them (14 percent at grade 3 and 25 percent at grade 5) achieved at least 80 percent comprehension.
- Advanced Readers (121 CWPM and above) – A small percentage of the advanced readers had 0 percent comprehension (20 percent in grade 3 and 4 percent in grade 5). None of the third graders and about a third of the fifth graders (33 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.