



USAID
FROM THE AMERICAN PEOPLE



EARLY GRADE READING ASSESSMENT BASELINE REPORT

BALUCHISTAN PROVINCE

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 BALOCHISTAN

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 Directorate of Education (Schools) and the Government of Balochistan 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 Society for Community Strengthening and Promotion of Education Balochistan.

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	37
Annexes	40
Annex 1: Complete Item Statistics by Grade	41
Annex 2: Box Plots for Phonics and Reading-rate Fluency Tasks	42
Annex 3: Examples of Fluency Score Threshold Calculations	45
Annex 4: Distribution of Reading Fluency and Comprehension Scores using Fixed Intervals	47

List of Tables and Figures

Table 1: Timeline (January 2013 to May 2014)	11
Table 2: Schools by District, Treatment, Location, and Gender	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: Tasks Statistics (Full and Light Treatment Groups)	20
Table 9: Percent Correct Scores by Grade and Task (Full and Light Treatment Groups)	22
Table 10: Percent Correct Scores by Grade, Task, and Group	22
Table 11: Percent Correct Scores by Grade, Task, and Gender (Full and Light Treatment Groups)	24
Table 12: Percent Correct Scores by Group, Grade, Gender, and Task	25
Table 13: Baseline Maximum Scores on Fluency (Timed) Tasks (Full and Light Treatment Groups)	26
Table 14: Phonics and Reading-Rate Fluency Task Means by Grade (Full and Light Treatment Groups)	26
Table 15: Phonics and Reading-Rate Fluency Task Means by Grade and Group	27
Table 16: Phonics and Reading-Rate Fluency Task Means by Grade and Gender (Full and Light Treatment Groups)	27
Table 17: Phonics and Reading-Rate Fluency Task Means by Group, Grade, and Gender	28
Table 18: Percentage of students by Language Spoken at Home	29
Table 19: Summary Scores by Student Age	29
Table 20: Summary Scores by Reading the Quran at Home	29
Table 21: Summary Scores by the Presence of a Library at the School	30
Table 22: Summary Scores by the Presence of Newspapers at Home	30
Table 23: Summary Scores by the Presence of Magazines at Home	30
Table 24: Summary Scores by the Presence of Books at Home	31
Table 25: Summary Scores by Children Having Someone Read to Them at Home	31
Table 26: Summary Scores by Children Reading to Someone Else at Home	31
Table 27: Summary Scores by Children Reading Silently at Home	31
Table 28: Summary Scores by Teacher Academic Qualification	32

Table 29: Summary Scores by Teacher Professional Qualification	32
Table 30: Summary Scores by Teacher Age	32
Table 31: Summary Scores by Teacher Experience	33
Table 32: Summary Scores by Teacher In-Service Training	33
Table 33: Summary Scores by Head Teacher Academic Qualification.....	33
Table 34: Summary Scores by Head Teacher Professional Qualification.....	34
Table 35: Summary Scores by Head Teacher Experience.....	34
Table 36: Summary Scores by Head Teacher In-Service Training.....	34
Table 37: Summary Scores by Head Teacher Support of Teachers in Reading	35
Table 38: Summary Scores by Head Teacher Training in Teaching Reading	35
Table 39: Summary Scores by School Location.....	35
Table 40: Summary Scores by School Gender.....	36
Table 41: Summary Scores by PTA/SMC/PTSMC/PTC.....	36
Table 42: Summary Scores by Presence of a School Library.....	36
Table 43: Summary Scores by Infrastructure (Drinking Water, Electricity, Toilets)	36
Table A1: Complete Item Statistics by Grade.....	41
Table A2: Thresholds for CWPM with 80 Percent Comprehension	45
Table A3: Thresholds for CWPM with Fixed Intervals	46
Table A4: Grade 3 Reading Fluency and Comprehension.....	47
Table A5: Grade 5 Reading Fluency and Comprehension.....	48
Figure 1: Evaluation Design.....	9
Figure 2: Grade 3 Summary Scores.....	21
Figure 3: Grade 5 Summary Scores.....	21
Figure 4: Full Treatment Percent Correct Scores by Grade and Task	23
Figure 5: Light Treatment Percent Correct Scores by Grade and Task.....	23
Figure 6: Grade 3 Percent Correct Scores by Task and Gender (Full and Light Treatment Groups).....	24
Figure 7: Grade 5 Percent Correct Scores by Task and Gender (Full and Light Treatment Groups).....	25
Figure A1: Understanding Boxplots	42
Figure A2: Phonics and Reading-Rate Fluency Box Plots for Grade 3	43
Figure A3: Phonics and Reading-Rate Fluency Box Plots for Grade 5	44
Figure A4: Grade 3 Reading Fluency and Comprehension	48
Figure A5: Grade 5 Reading Fluency and Comprehension	49

ACRONYMS

AJK	Azad Jammu and Kashmir
B.A.	Bachelor of Arts
BEFARe	Basic Education for Awareness, Reforms and Empowerment
B.Sc.	Bachelor of Science
C.T.	Certificate of Teaching (Grade 12 plus FA/FSC Certificate)
DOE	Directorate of Education
EGRA	Early Grade Reading Assessment
F.A.	Fellow in Arts
FATA	Federally Administered Tribal Areas
F.Sc.	Fellow in Sciences
GB	Gilgit-Baltistan
ICT	Islamabad Capital Territory
KP	Khyber Pakhtunkhwa
M.A.	Master of Arts
Matric	Secondary School (Grade 10) Certificate (Matriculation)
M.Ed.	Master of Education
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
PTA	Parent Teacher Association
PTC	Parent Teacher Council
P.T.C.	Primary Teaching (Grade 12) Certificate
PTSMC	Parent Teacher School Management Committee
QCO	Quality Control Officer
SCSPEB	Society for Community Strengthening and Promotion of Education Balochistan
SPSS	Statistical Package for the Social Sciences
SRP	Sindh Reading Project
STS	School-to-School International
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

Overview

In 2013, Management Systems International (MSI) and School-to-School International (STS) conducted a baseline reading assessment for primary school children prior to the launching of two USAID-funded projects: the Pakistan Reading Project (PRP) and the Sindh Reading Program (SRP). PRP is targeting improved reading for 910,000 children in Azad Jammu and Kashmir (AJK), Balochistan, the Federally Administered Tribal Areas (FATA), Gilgit-Baltistan (GB), the Islamabad Capital Territory (ICT), Khyber Pakhtunkhwa (KP), and Sindh, while 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 Balochistan, which took place in October 2013. In May 2013, GB, AJK and ICT were part of Round 1 of the baseline data collection; Round 2 in KP and Sindh was completed in September 2013; and Round 3 in Balochistan, Punjab, and FATA was completed in October 2013. The following activities were carried out for all of the provinces, including Balochistan: 1) design, 2) sampling, 3) instrumentation, 4) planning, 5) training, 6) implementation, 7) analysis, and 8) reporting.

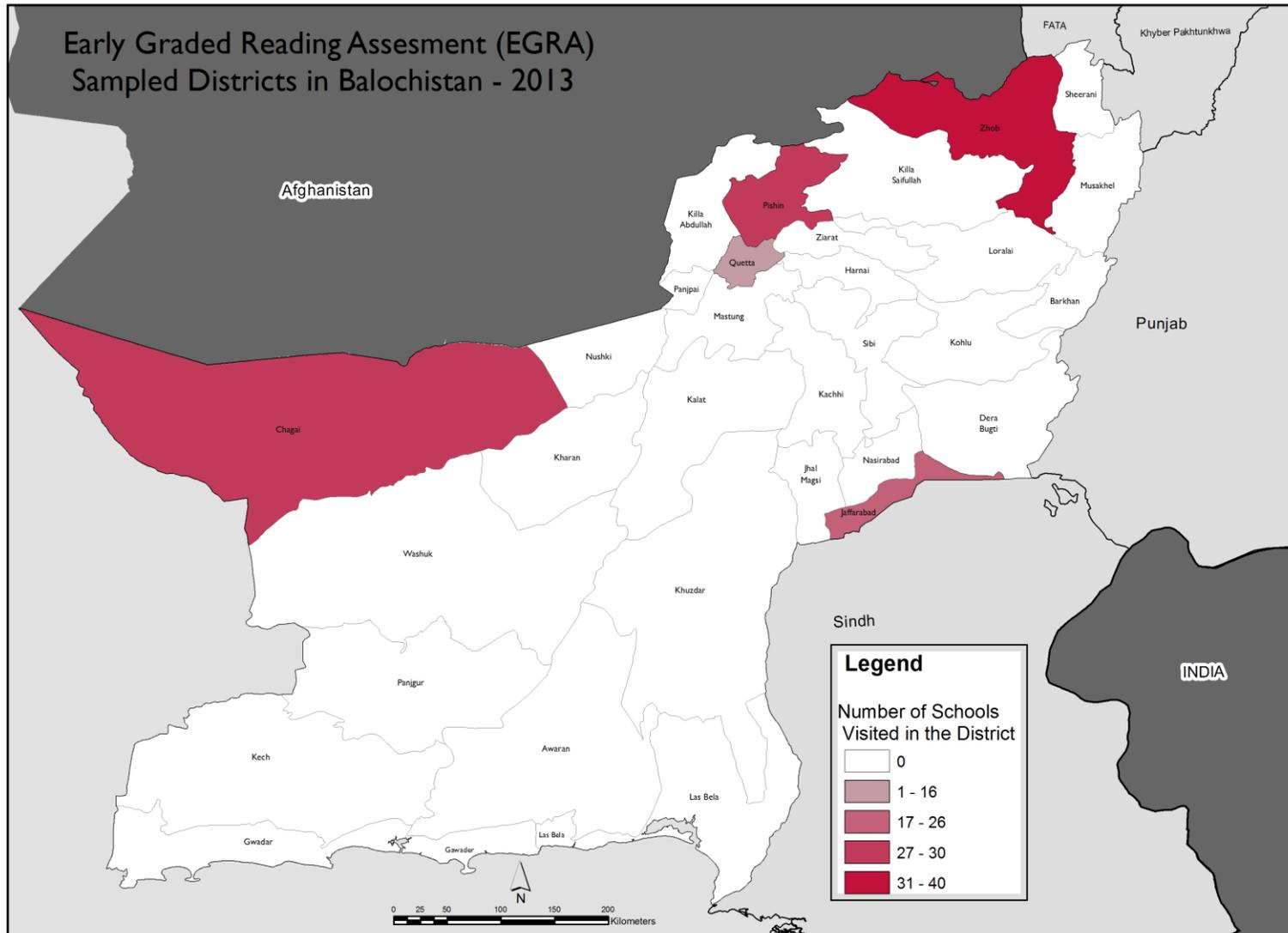
The external evaluation design, which was developed prior to the baseline assessment, was tailored to the implementation of the PRP and SRP in each province. In most of the provinces, a quasi-experimental design will be used, with two treatment groups: “full treatment” and “light treatment.” The full treatment group will receive support in two areas: 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 – were assessed at three time points: baseline, midline, and endline. An internationally accepted assessment tool, the Early Grade Reading Assessment (EGRA), was individually administered to a target sample of 33,000 children in 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 Balochistan, all activities were completed by the end of January 2014, including a draft report. The EGRA baseline results were presented and discussed at a consultative meeting in Karachi in April 2014. Representatives from the Directorate of Education (DOE), USAID, PRP, and the contractors (MSI and STS) attended the consultation. Revisions were then be 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 Balochistan are highlighted below:

Implementation

1. The Balochistan 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. Given the long-term design of this evaluation, this baseline report will not statistically test the differences between the groups' initial reading performance because doing so may confound the study by facilitating competition between the groups. The report will, however, present the baseline scores for each group. (Please see Figure 1 and the accompanying text for a fuller description of the evaluation design.)
2. District selection into full and light treatment groups was finalized following consultative meetings between the DOE and USAID in February 2013.
3. EGRA in Urdu was used in the Balochistan province. 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.
4. A total of 140 schools, with 70 schools from each group (full and light treatment) were selected for the baseline.
5. A simple random sample of three districts - Quetta, Pishin, and Jaffarabad – was taken for the full treatment, while two districts – Chagai and Zhob – were selected for the light treatment. Within these districts, a random sample of male and female schools were selected, followed by a random sample of grades 3 and 5 students within those schools. The number of schools in the districts and the number of samples from each district by gender is shown in Table 2.
6. The results from this sample are presented in this report as a generalized view of the reading levels for students in the Balochistan schools. 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 Balochistan was October 2013, but due to dissimilar district academic calendars, students were tested at different points along the grade 3 and grade 5 timeline. The full treatment groups in Quetta and Pishin, along with the light treatment district of Zhob, started school in March and had seven months of instruction prior to taking the EGRA. In contrast, Jaffarabad (full treatment) and Chagai (light treatment) districts started the academic year in August, affording two months of teaching prior to the assessment. Consequently, this lends additional support for maintaining a consistent testing calendar (October) at midline and endline and negating district comparisons.
8. The assessment tools were successfully administered in (with the percentage of the target reached in parentheses) 140 schools (100.0 percent) to 3,866 students (92.0 percent), 260 teachers (92.9 percent), and 140 head teachers (100.0 percent). The percent of teachers is low because some

teachers taught both grades and others did not indicate a grade. These responses were not counted in the survey results, which were analyzed by grade level.

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 and the piloting process. Reliability was assured through the high quality of the assessment tasks and the standardized administration of the tools. Reliability estimates (of internal consistency) were calculated using the coefficient alpha.
10. The data entry and data cleaning process followed international standards. All student data were entered twice into two separate databases. 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 tasks. These scores provide a comprehensive picture of student performance. Analysis of student, teacher, head teacher, and school characteristics was carried out using the summary scores.

Results

1. EGRA was administered to 1,985 grade 3 students and 1,881 grade 5 students. The reliability estimates were high for both grades ($\alpha = 0.88$ for grade 3 and 0.88 for grade 5), indicating that the items worked well in measuring reading constructs at each grade level.
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 posted the highest scores in orientation to print, followed by tasks (familiar word reading, passage reading, and letter name recognition) with less than 50 percent correct. The most difficult tasks for these students were comprehension (passage and listening) and phonics (non-word reading, letter sound knowledge, and phonemic awareness). At grade 5, the highest scores were in familiar word reading, passage reading, and orientation to print; whereas the most challenging tasks were comprehension (passage and listening) and letter sound knowledge.
4. There was substantial progression from grade 3 to grade 5 on the summary score (17 points) and on all tasks scores – the greatest gains were in familiar word, passage, and non-word reading. This progress was consistent across gender and treatment groups.
5. There were differences between boys and girls on the task and summary scores, but most of these differences were small. For example, the difference in the summary score was less than one point. The largest difference was in passage reading, favoring the girls (4.5 and 10.7 points for grades 3 and 5, respectively).
6. Students were timed on five tasks as they read words or passages. These tasks were categorized into phonics (letter name recognition, letter sound knowledge, and non-word reading) and reading-rate fluency (familiar word and passage reading). Students in both grades had lower phonics scores than reading-rate fluency scores. Moreover, gains from grade 3 to grade 5 were lower for phonics than for reading-rate fluency tasks. Although the passage was designed for grade 3, this difference shows that the reading-rate 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 and phonemic awareness 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 Balochistan.

7. The full treatment group had higher scores on half of the tasks, and the light treatment group had higher scores on the other half. The largest difference was in letter sound knowledge (11 points) favoring the light treatment group, but differences in all other tasks were small. This minor discrepancy will be corrected statistically at the midline and endline by analyzing the growth for each group from baseline to midline and endline. Because this is a baseline report, the group differences will not be statistically tested at this time.
8. The student questionnaire revealed three interesting findings. The first positive finding was that having reading materials and opportunities to read in the home seemed to have a positive effect on reading outcomes for both grades 3 and 5 students. Secondly, at grade 3, summary scores increased with relative age (younger than normal, normal, older than normal age); older students in the grade had higher reading scores. However, by grade 5 that advantage was no longer significant. Third, Balochistan students are performing well on the Urdu test considering only 1 percent of the students reported speaking Urdu as their primary language at home.
9. School, teacher, and head teacher questionnaire findings were mostly inconclusive, due to small sample sizes and the lack of variation in the scores that were related to their characteristics. For example, an analysis of student scores by teacher and head teacher education, certification, age, experience, and attendance at in-service trainings found no consistent patterns relating to lower or higher student scores. For the schools, better infrastructure was associated with better student reading scores.

Evaluation Recommendations

Given the success of the baseline assessment in Balochistan (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. Because of the variability among the districts' academic calendars, the instructional time from the start of school to the EGRA administration varied by five months among the districts in Balochistan. Since students can make great gains in reading during the primary grades, it is essential that testing occur at a consistent point in the academic year. Midline and endline testing in Balochistan should occur in October, thus matching the baseline timeframe and standardizing the instructional time across the study.

5. 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.
6. The data entry system should continue to be used, with the same systems for recruitment and training of data entry supervisors and operators, along with implementation through networked computers, double data entry, and reconciliation of errors.
7. The analysis should follow the same procedures with calculations of reliability, difficulty, task percent-correct scores, summary scores, and timed task scores. The baseline, midline, and endline scores should be computed using the same procedures so that improvements in students' reading can be accurately examined over time.
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 for collecting data on PRP- and SRP-supported interventions so that student scores can be correlated with these indicators.

CHAPTER I: INTRODUCTION

The Pakistan Reading Project (PRP) and the Sindh Reading Program (SRP) are two five-year initiatives funded by USAID. The projects/programs will cover over 40,000 government schools in Pakistan's eight provinces/areas/territories (hereafter referred to as provinces). PRP is targeting improved reading for 910,000 children in AJK, Balochistan, FATA, GB, ICT, Balochistan, 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. The goal of the evaluation is to conduct a long-term assessment for both groups in each province.

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

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 Balochistan, this was complemented by a quasi-experimental design with the two treatment groups (full and light).
2. Sampling – The sampling plan 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 30 districts in Balochistan, out of which 17 were full treatment and the remaining 13 were light treatment. Seven of the full treatment and eight of the light treatment districts were removed from the sample due to security reasons. A simple random sample of three districts - Quetta, Pishin, and Jaffarabad – was taken for full treatment and two districts – Chagai and Zhob – was taken for light treatment. Schools were then apportioned according to location and gender. As there are very few urban schools in Balochistan, balance for the location variable was not possible because too few urban schools (8) were in the representative sample. Therefore, it was not appropriate to fully investigate the EGRA differences between urban and rural schools in Balochistan. Conversely, half of the schools selected for the assessment were male and half were female, thus permitting some analysis at this level.
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 in Urdu and Sindhi. Model EGRA instruments were trans-adapted, piloted, revised, and finalized for use in Pakistan. The Urdu instruments were used in Balochistan.
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 they had clear comprehension and the adequate skills 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, and will be disseminated to the provincial education authorities. A template was developed according to guidelines from the USAID contract.

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.

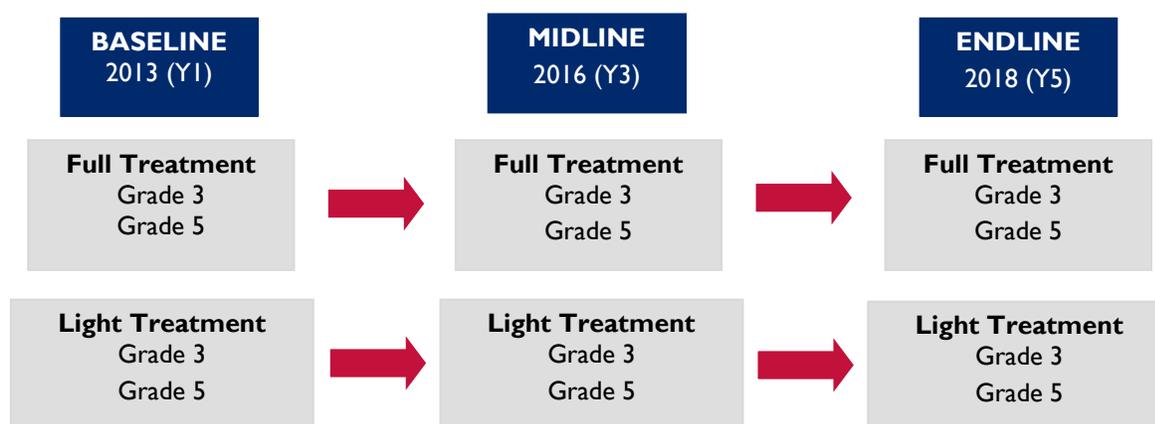
CHAPTER 2: DESIGN AND METHODOLOGY

This chapter presents the evaluation design and methodology, including the systems used for collecting the EGRA baseline data for schools in Balochistan. 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 Balochistan 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.

For the baseline assessment in Balochistan, a random selection from the full treatment districts as selected for the PRP interventions resulted in the choice of Quetta, Pishin, and Jaffarabad. Chagai and Zhob were randomly selected for the light intervention districts. For each treatment group and district, equal numbers of boys and girls schools were sampled for the EGRA testing. The sampling design met the USAID requirements of adequate sample size and equal gender representation (see the sampling section below).

In Balochistan, students were tested in Urdu, their main language of instruction. Some of the schools in the province use other languages during instruction (e.g., Balochi, Pashtun), though their materials (e.g., textbooks) are in Urdu.

In addition, some mixed schools were included in the sample, but only boys or girls were selected from these schools, and thus were considered as either male or female schools. Lists of the schools are safeguarded so that they can be used again in the midline and endline data collections.

Timeline

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

The process began with the planning and design of activities, including creating preliminary sampling designs, selecting model EGRA tasks, recruiting staff, and budgeting/contracting. From February to August, the EGRA team, with participation from Balochistan and other provinces, then prepared, piloted, and revised the EGRA tools and conducted the district/school sampling. The data collection in Balochistan took place in October 2013 and was followed by data entry, analysis, and reporting. Presentations to the DOE and USAID were made in April 2014. Note that the full treatment groups in Quetta and Pishin, along with the light treatment district of Zhob, started school in March and had seven months of instruction prior to testing. In contrast, Jaffarabad (full treatment) and Chagai (light treatment) districts started the academic year in August and had two months of teaching prior to testing.

TABLE 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															
Debrief to USAID and the MOE	X	X	X				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					
Produce draft reports												X	X				
Produce presentations															X		
Disseminate draft reports														X			
Make presentations															X		
Revise and finalize reports																X	X
Submit final reports to USAID																	X

Sampling

The sampling process started in January 2013 with the selection of the treatment districts by the provincial DOEs and USAID. The sampling for Balochistan was finalized in July 2013 following meetings with USAID. The EGRA team conducted the school sampling in July, which included developing the sampling requirements, verifying the sample in the field, and finalizing the sample.

There were a total of 30 districts in Balochistan, out of which 17 were full treatment and the remaining 13 were light treatment. Seven of the full treatment and eight of the light treatment districts were removed from the sample due to security reasons, leaving 10 full treatment and five light treatment districts. The districts included in the full treatment sampled population were Gwadar, Jaffarabad, Lasbela, Loralai, Nushki, Pishin, Quetta, Sibi, Qilla Saifullah, and Ziarat. The districts included in the light treatment sampled population were Chagai, Harnai, Qilla Abdullah, Sherani, and Zhob. A simple random sample of three districts - Quetta, Pishin, and Jaffarabad – was taken for full treatment. Two districts, Chagai and Zhob, were selected for light treatment. The sampling for Balochistan, as detailed in the sampling report for USAID¹, is briefly summarized in the following sub-sections 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 boys and girls schools (35 each) were selected.

Sampling Process and Field Verification

Due to the security concerns and other issues, not all districts in Balochistan were considered for PRP interventions or for the baseline assessment. From the chosen full treatment and light treatment districts, districts for the full and light treatment assessment groups were randomly selected. This resulted in a clustered sample. For the 35 boys and 35 girls schools in both the full and light treatment groups, the samples were divided among the selected districts according to the proportions of schools within those districts (stratified random sampling). An equal number of boys and girls schools were chosen within each group. For both groups, a second stratification was done at the “location” level, where schools were allocated by rural and urban according to their proportions in the National Education Management Information System (NEMIS). Few schools in Balochistan are located in urban areas, so the number of urban schools in the sample was relatively small. Table 2 shows the number of schools and replacement schools for both treatment groups per gender and location. 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*.

TABLE 2: SCHOOLS BY DISTRICT, TREATMENT, LOCATION, AND GENDER

District	Location	Schools	Percentage	Sample Schools		Replacement Schools	
				Boys	Girls	Boys	Girls
Full Treatment Group							
Jaffarabad	Rural	859	36	13	13	4	4
Jaffarabad	Urban	50	2	0	0	0	0
Pishin	Rural	813	34	13	13	3	3
Pishin	Urban	97	4	1	1	0	0
Quetta	Rural	373	16	5	5	2	2
Quetta	Urban	180	8	3	3	1	1
Sub-Total		2,372	100	35	35	10	10
Light Treatment Group							
Chagai	Rural	196	35	13	13	4	4
Chagai	Urban	33	6	2	2	0	0
Zhob	Rural	260	47	16	16	5	5
Zhob	Urban	67	12	4	4	1	1
Sub-Total		556	100	35	35	10	10
Total (both groups)			2,928	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 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

For the full treatment schools, five schools – three boys and two girls schools – were substituted with schools randomly selected from the “replacement schools” list. The schools were replaced due to lower than expected numbers of students in the original samples. For the light treatment schools, 20 schools, 10 girls and 10 boys schools, were replaced for the same reason. 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 Balochistan, 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 DOEs and teacher training institutes throughout Pakistan – including two subject specialists from Balochistan – 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, while the Sindhi tools were piloted in June in Sindh. 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 for the Urdu and English tools with a limited number of experts from the trans-adaptation workshop. The Sindhi tools were revised in July with Sindhi language experts. Changes were made to the instruments based on the pilot data and field observations. These changes were summarized in the piloting report. The team then finalized the four instruments for each language and submitted them to USAID. USAID made suggestions, particularly around the inclusion of reading- and library-related items in the questionnaires that would provide information for the PRP and SRP. The instruments were approved and then used in the training workshops in advance of Round 1 data collection in May. The final instruments consisted of the following:

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

These instruments are available for use by education officials.

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

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 August, the Society for Community Strengthening and Promotion of Education Balochistan (SCSPEB) was chosen for data collection activities, while the Basic Education for Awareness, Reforms, and Empowerment (BEFARe) was selected for data entry activities. MSI, STS, and SCSPEB collaborated on the data collection in Balochistan.

Data Collection

In September, EGRA senior managers trained MTs and QCOs during a two-week session in Islamabad. The MTs then spent one week in Islamabad training the SCSPEB data collection team, which was comprised of one regional coordinator, five field supervisors, and 64 enumerators. The QCOs, coordinators, supervisors, and enumerators organized the logistics for the data collection. Following the training and logistical preparations, the QCOs and field supervisors conducted a three-day refresher course for the enumerators in each district just prior to commencing data collection in the schools.

Over a 10-day period in October, the enumerators spent a day in each of the 140 schools to collect the baseline data in Balochistan. 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. At the end of data collection, all booklets were returned 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 September, the team trained the BEFARe data coordinator, four supervisors, and 36 data entry operators (DEOs) on the application. In October and November, the EGRA and BEFARe teams entered the data for over 21,000 student booklets, along with questionnaires for the teachers and head teachers (Rounds 2 and 3). This total included approximately 4,200 booklets and questionnaires for Balochistan.

Data Cleaning

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

Data Analysis

Methodology

In June, the EGRA statisticians and psychometricians 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 SPSS 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 for the student scores and the three sets of questionnaire data.

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

Validity and Reliability

Validity evidence for the tests was derived from previous experiences with EGRA in other developing countries, as well as through the trans-adaptation process in Pakistan. The test developers targeted grade 3 for the level of the tasks. The 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. Table 3 shows the reliability estimates for both grades were the same (0.88) in Balochistan. These reliabilities are excellent and lend credibility to the internal consistency of the tests, indicating that the items are generally measuring similar reading constructs for both grade levels.

TABLE 3: RELIABILITY ESTIMATES

Language	Grade Level	Tasks	N-count	Alpha
Urdu	Grade 3	9	1,985	0.88
	Grade 5	9	1,881	0.88

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

Score Calculation

The EGRA data were analyzed in 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 Balochistan. 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

The intended sample was 70 full and 70 light treatments schools. Within these schools, the target was to assess 15 students in each grade, totaling 4,200 students (i.e., 2,100 for each gender, treatment group, and grade). Table 7 shows the actual number of students in the sample by treatment, grade, and gender. The percentage of the target was higher for the full treatment (97.7) than light treatment (86.3). Grade 3 (94.5) was higher than grade 5 (89.6). The boys' percent (91.3) was slightly lower than the girls' (92.2).

A small number of students in grade 3 ($n = 6$) and grade 5 ($n = 6$) did not complete the gender item on the questionnaire. When analyzing the students by gender, the sample was 3,854 students (91.8 percent of the intended). However, when the data were not analyzed by gender, the total actual sample was 3,866 (92.0 percent of the intended).

During the field verification prior to the assessment, most of the schools reported having at least the minimum number of 15 students in each of grades 3 and 5; a few schools were kept in the sample even though, during the field verification, their actual numbers were below the target. The main reason, however, for the difference between the intended and actual samples was low student attendance on the survey date.

TABLE 7: ACTUAL STUDENT SAMPLE BY GRADE AND GENDER

Treatment	Grade Level	Sample	Boys	Girls	Missing	Total
Full Treatment	Grade 3	Students	515	520	4	1,035
		% of Target	98.1%	99.0%	--	98.6%
	Grade 5	Students	504	507	3	1,011
		% of Target	96.0%	96.6%	--	96.3%
	Total	Students	1,019	1,027	7	2,053
		% of Target	97.0%	97.8%	--	97.8%
Light Treatment	Grade 3	Students	459	485	2	946
		% of Target	87.4%	92.4%	--	90.1%
	Grade 5	Students	439	425	3	867
		% of Target	83.6%	81.0%	--	82.6%
	Total	Students	898	910	5	1,813
		% of Target	85.5%	86.7%	--	86.3%
Full and Light Treatment	Grade 3	Students	974	1,005	6	1,985
		% of Target	92.8%	95.7%	--	94.5%
	Grade 5	Students	943	932	6	1,881
		% of Target	89.8%	88.8%	--	89.6%
	Total	Students	1,917	1,937	12	3,866
		% of Target	91.3%	92.2%	0.3%	92.0%

Task and Item Statistics

Table 8 shows the statistics for the tasks for the Balochistan sample. Two 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 indicating that the item (or task) is of good quality.

In Table 8 below, the task p-values for grade 3 in Balochistan ranged from 0.13 to 0.50, thus providing a spread on the lower half of the difficulty spectrum. The p-values for grade 5 were higher, ranging from 0.34 to 0.70 or in the middle parts of the range. The level of difficulty for both grade levels was appropriate for this baseline measure because there will be enough room in the scale for capturing growth during the midline and endline assessments. For item-total correlations, a generally acceptable threshold is 0.20 and above. All of the task scores in grades 3 and 5 had item-total correlations greater than 0.30, indicating very good quality for these tasks. Complete item statistics are provided in Annex 1 at the end of this report.

TABLE 8: TASKS 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)	0.50	0.31	0.58	0.40
2. Letter name recognition (timed)	0.41	0.66	0.53	0.61
3. Phonemic awareness (untimed)	0.29	0.36	0.41	0.45
4. Letter sound knowledge (timed)	0.25	0.58	0.36	0.59
5. Familiar word reading (timed)	0.42	0.84	0.70	0.79
6. Non-word reading (timed)	0.26	0.79	0.48	0.77
7a. Passage reading (timed)	0.42	0.83	0.70	0.78
7b. Passage comprehension (untimed)	0.13	0.68	0.34	0.72
8. Listening comprehension (timed)	0.19	0.52	0.35	0.57

Task and Summary Scores

The next part of the analysis involved plotting the summary scores. Histograms of the summary scores (Figures 2 and 3) show that the distributions are moving from left to 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

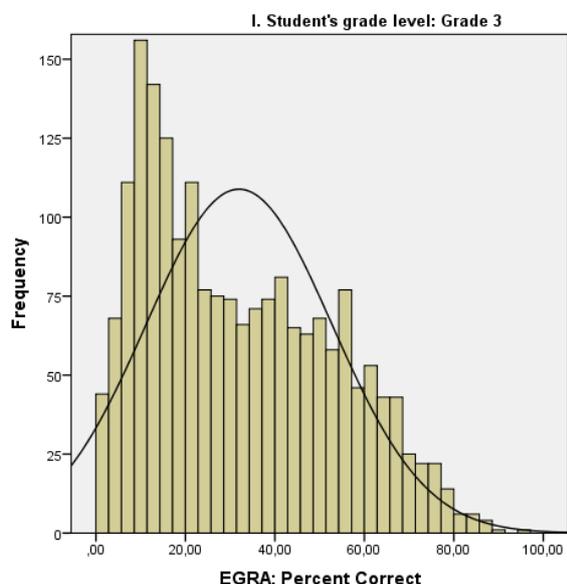


FIGURE 3: GRADE 5 SUMMARY SCORES

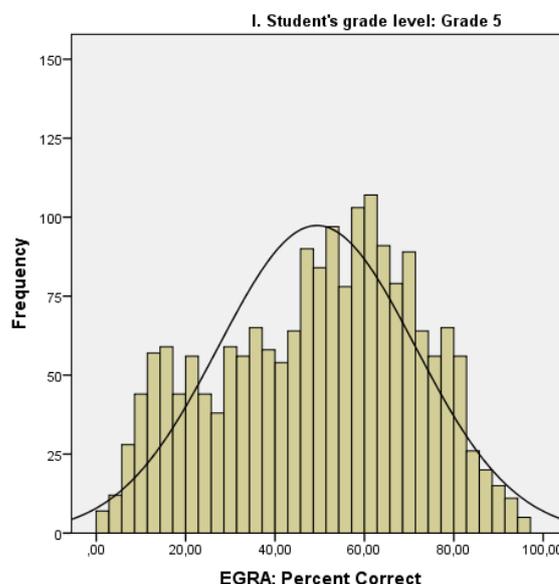


Table 7 and 8 and Figure 4 provide the average scores by task using percent correct scores. The score for each task was calculated using the total number correct and dividing by the number of items. For instance, a student who scored 3 out of 5 on Task 1 would receive a score of 60 percent. Averages were then calculated for all students on Task 1, which in Balochistan was 50.4 percent for grade 3 and 57.6 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., 57.6 percent minus 50.4 percent equals 7.2 percentage points.

Table 9 provides the scores using the percent correct metric by task, and for the summary (or grand mean) for all grade 3 and 5 students (i.e., for the full and light treatment groups combined). Grade 3 posted the highest scores in orientation to print, followed by familiar word reading, passage reading, and letter name recognition. The most difficult tasks for these students were comprehension (passage and listening) and phonics (non-word reading, letter sound knowledge, and phonemic awareness). At grade 5, the highest scores were in familiar word reading, passage reading, and orientation to print; the most challenging tasks were comprehension (passage and listening) and letter sound knowledge.

There was also substantial progression from grade 3 to grade 5 on the summary score (17 points). The greatest gains were in passage reading (28 points), familiar word reading (28 points), and non-word reading (23 points). In areas where there are small differences between the 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	50.4%	57.6%	7.2% points
2. Letter name recognition	40.7%	53.2%	12.5% points
3. Phonemic awareness	29.4%	40.7%	11.3% points
4. Letter sound knowledge	25.2%	35.8%	10.6% points
5. Familiar word reading	42.2%	70.0%	27.8% points
6. Non-word reading	25.7%	48.3%	22.6% points
7a. Passage reading	41.7%	69.9%	28.2% points
7b. Passage comprehension	13.5%	33.7%	20.2% points
8. Listening comprehension	19.1%	35.4%	16.3% points
Reading Summary Score	32.0%	49.4%	17.4% points

As seen in Table 10, the light group had higher scores on five tasks for grade 3, and four tasks for grade 5. These discrepancies are generally small and will be corrected statistically at midline and endline by analyzing the growth for each group from baseline to midline and endline. Because this is a baseline report, the group differences were not statistically tested at this time.

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

Task (Subtest)	Full		Light	
	Grade 3	Grade 5	Grade 3	Grade 5
1. Orientation to print	48.2%	56.4%	53.0%	59.2%
2. Letter name recognition	40.7%	54.2%	40.9%	52.2%
3. Phonemic awareness	26.1%	37.1%	33.2%	44.9%
4. Letter sound knowledge	21.9%	35.5%	30.0%	36.5%
5. Familiar word reading	39.7%	68.9%	45.1%	71.1%
6. Non-word reading	23.6%	49.0%	28.2%	47.8%
7a. Passage reading	39.6%	69.7%	44.4%	70.5%
7b. Passage comprehension	11.8%	32.3%	15.6%	35.5%
8. Listening comprehension	16.2%	33.5%	22.4%	37.4%
Reading Summary Score	29.6%	48.4%	34.6%	50.6%

FIGURE 4: FULL TREATMENT PERCENT CORRECT SCORES BY GRADE AND TASK

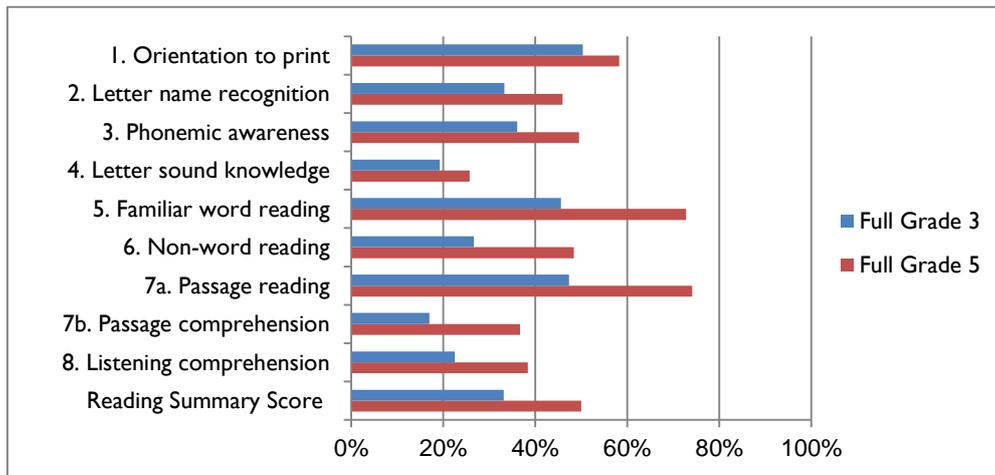
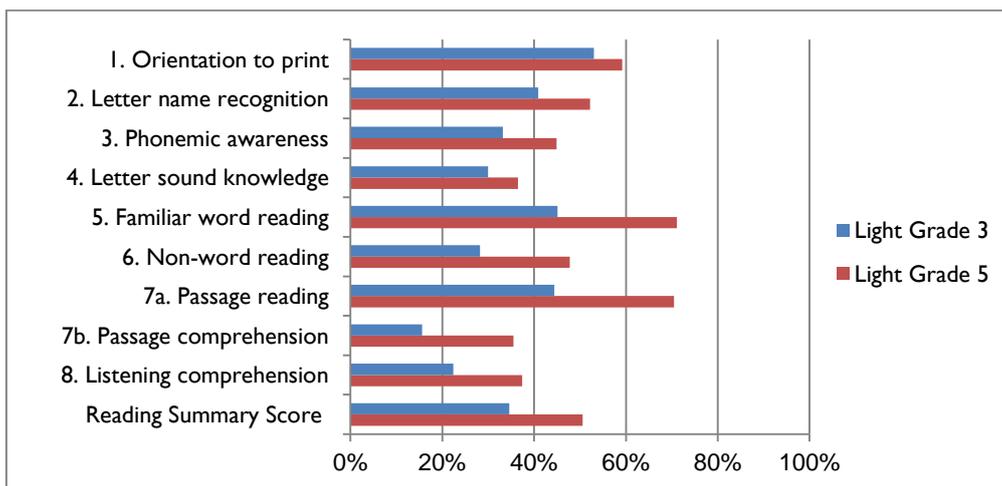


FIGURE 5: LIGHT TREATMENT PERCENT CORRECT SCORES BY GRADE AND TASK



Percent correct scores by grade and gender are presented in Table 11 and in Figures 6 and 7. At grade 3, the boys and girls performed best on orientation to print, followed by tasks that had less than 50 percent correct: letter name recognition, familiar word reading, and passage reading. For both genders, the most challenging tasks were in comprehension (passage and listening) and phonics (letter sound knowledge and non-word reading). Similarly, at grade 5, girls and boys scored well on familiar word reading, passage reading, and orientation to print, and were challenged by the two comprehension tasks and letter sound knowledge. Grade 3 girls had statistically higher scores in passage reading and passage comprehension ($p < 0.01$), while grade 5 girls had statistically higher scores than boys on phonemic awareness and passage comprehension ($p < 0.05$).

TABLE II: PERCENT CORRECT 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	50.1%	50.8%	57.4%	58.1%
2. Letter name recognition	41.2%	40.4%	53.9%	52.9%
3. Phonemic awareness	30.4%	28.4%	38.7%	42.8%*
4. Letter sound knowledge	25.9%	24.6%	36.5%	35.7%
5. Familiar word reading	40.6%	43.8%	71.0%	69.1%
6. Non-word reading	25.5%	26.0%	50.5%	46.6%*
7a. Passage reading	38.9%	43.3%*	69.3%	68.7%
7b. Passage comprehension	11.3%	15.8%*	28.6%	39.3%*
8. Listening comprehension	20.0%	18.3%	34.5%	36.4%
Reading Summary Score	31.5%	32.2%	48.8%	49.7%

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

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

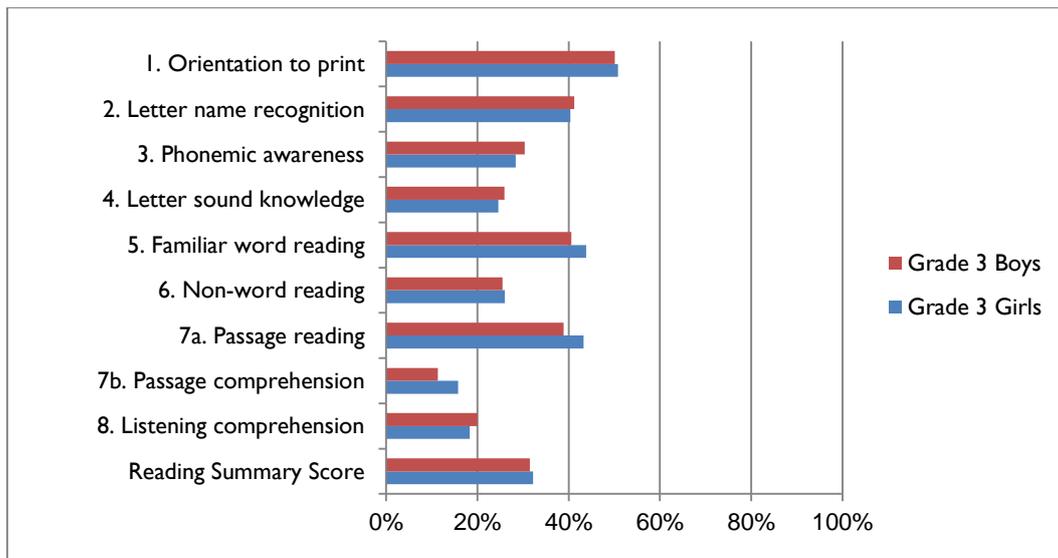
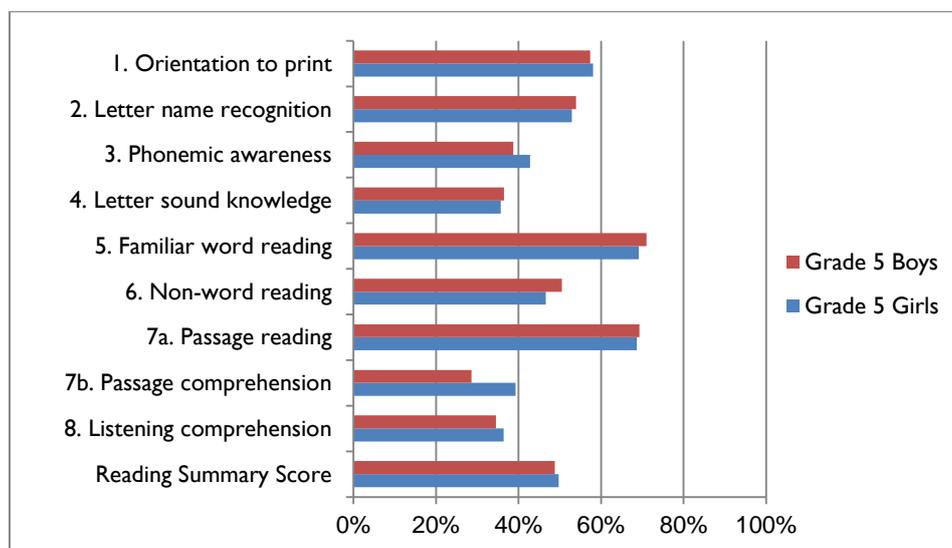


FIGURE 7: GRADE 5 PERCENT CORRECT SCORES BY TASK AND GENDER (FULL AND LIGHT TREATMENT GROUPS)



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, GENDER, AND TASK

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	48.5%	47.8%	55.4%	57.4%	52.0%	54.0%	59.7%	59.0%
2. Letter name recognition	38.2%	43.2%	52.9%	55.5%	44.5%	37.4%	55.0%	49.7%
3. Phonemic awareness	24.4%	27.7%	33.7%	40.4%	37.1%	29.2%	44.4%	45.7%
4. Letter sound knowledge	20.1%	23.6%	34.6%	36.3%	32.4%	25.6%	38.6%	34.9%
5. Familiar word reading	35.3%	44.0%	68.0%	69.8%	46.6%	43.5%	74.4%	68.2%
6. Non-word reading	21.2%	25.9%	50.4%	47.6%	30.2%	26.2%	50.6%	45.3%
7a. Passage reading	34.6%	44.6%	67.5%	71.8%	45.2%	43.5%	73.9%	67.4%
7b. Passage comprehension	7.4%	16.2%	23.7%	40.8%	15.6%	15.5%	34.1%	37.5%
8. Listening comprehension	13.3%	19.1%	29.8%	37.1%	27.6%	17.4%	40.0%	35.6%
Reading Summary Score	26.9%	32.2%	46.2%	50.5%	36.8%	32.4%	52.2%	49.1%

Timed Tasks: Phonics and Reading-Rate Fluency 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 13 to 17 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 15 to 17 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	180	330
4. Letter sound knowledge	110	154
6. Non-word reading	94	131
Reading-Rate Fluency Subtest	Grade 3	Grade 5
5. Familiar word reading	250	167
7a. Passage reading	225	240

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 14 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	42.1	54.8	12.7 points
4. Letter sound knowledge	33.0	42.1	9.1 points
6. Non-word reading	21.1	34.0	12.9 points
Reading-Rate Fluency Subtest	Grade 3	Grade 5	Difference (G5 – G3)
5. Familiar word reading	33.5	57.0	23.5 points
7a. Passage reading	44.6	79.0	34.4 points

In comparing the treatment groups (Table 15), there were differences in the fluency tasks between the two groups favoring the light treatment group. However, both groups showed similar patterns in fluency. Non-word reading was the most challenging and passage reading was the most fluent. Again, these differences will be corrected statistically in the midline and endline evaluations.

TABLE 15: PHONICS AND READING-RATE FLUENCY TASK MEANS BY GRADE AND GROUP

Phonics Fluency Subtest	Grade 3		Grade 5	
	Full	Light	Full	Light
2. Letter name recognition	42.1	42.3	53.4	55.9
4. Letter sound knowledge	33.4	32.6	39.3	44.8
6. Non-word reading	18.1	25.7	29.8	38.5
Reading-Rate Fluency Subtest	Grade 3		Grade 5	
	Full	Light	Full	Light
5. Familiar word reading	31.4	36.1	53.6	60.3
7a. Passage reading	39.8	51.2	73.8	84.2

Boys and girls fluency rates were very similar for both grades (Table 16). Passage reading was the only statistically significant difference, where third grade girls performed better than the third grade boys ($p < 0.01$).

TABLE 16: PHONICS AND READING-RATE FLUENCY TASK MEANS BY GRADE AND GENDER (FULL AND LIGHT TREATMENT GROUPS)

Phonics Fluency Subtest	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
2. Letter name recognition	42.5	41.9	55.5	54.1
4. Letter sound knowledge	32.4	33.6	43.0	41.1
6. Non-word reading	21.9	20.4	34.6	33.5
Reading-Rate Fluency Subtest	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
5. Familiar word reading	31.6	35.1	56.9	57.0
7a. Passage reading	41.3	47.2*	76.9	81.0

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

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	38.4	43.3	53.6	56.7	44.6	37.5	56.2	50.0
4. Letter sound knowledge	19.9	23.7	35.3	36.3	32.4	25.6	38.7	34.7
6. Non-word reading	11.4	13.9	29.4	27.8	15.5	14.5	29.0	27.1
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	20.8	26.9	50.7	53.2	27.3	29.1	53.8	52.2
7a. Passage reading	25.5	35.0	64.7	72.3	32.0	38.1	70.1	73.3

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.

Student Questionnaires

One survey question asked the students what language was spoken in the home. Both grades showed similar patterns in the primary language spoken at home (Table 18). Most families spoke Pashto (50 percent), followed by Balochi (25 percent), other (15 percent), Sindhi (6 percent), Punjabi (1 percent), and Urdu (1 percent). Although the assessments were in Urdu, only 1 percent of Balochistan students spoke Urdu as their primary language in the home.

TABLE 18: PERCENTAGE OF STUDENTS BY LANGUAGE SPOKEN AT HOME

Language	Grade 3		Grade 5	
	n-count	Percent	n-count	Percent
Urdu	25	1.3%	27	1.4%
Sindhi	122	6.1%	115	6.1%
Pashto	1,022	51.5%	953	50.7%
Punjabi	22	1.1%	24	1.3%
Balochi	498	25.1%	476	25.3%
Other or Missing	296	14.9%	286	15.2%
Total	1,985	100.0%	1,881	100.0%

Table 19 has 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 significant differences among the age groups; younger students had significantly lower average summary scores than the two older groups. Conversely, at grade 5, the scores were not significantly different among age groups, eradicating the older-student advantage by grade 5.

TABLE 19: SUMMARY SCORES BY STUDENT AGE

Age Group	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Younger than normal age	51	26.1%*	35	48.5%
Normal age	542	31.2%*	480	52.4%
Older than normal age	1,381	32.6%*	1,351	48.5%
Missing	11	--	5	--
Total	1,985	32.0%	1,881	49.4%

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

Table 20 shows the summary scores according to whether the student reads the Quran at home. There were significant differences in both grades favoring students who read the Quran.

TABLE 20: SUMMARY SCORES BY READING THE QURAN AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	150	21.2%	107	36.9%
Yes	1,810	32.8%*	1,751	50.3%*
Missing	25	--	23	--
Total	1,985	32.0%	1,881	49.4%

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

Table 21 depicts the differences in scores based on whether there is a library at the school. Students reporting the presence of a library had significantly higher summary scores for grade 3. Conversely, at grade 5, students indicating the lack of a library had higher summary scores. Please note the high percentage (16 percent and 13 percent for grades 3 and 5, respectively) of students who had missing responses or did not know if the school had a library.

TABLE 21: 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	1,080	31.9%	1,055	52.6%*
Yes	590	37.7%*	577	48.3%
Missing	315	--	249	--
Total	1,985	32.0%	1,881	49.4%

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

In Tables 22 to 24, the data showed that the existence of newspapers, magazines, and books generally made a difference in the reading scores for both grades.

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

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	1,315	28.5%	992	44.0%
Yes	670	38.9%*	889	55.4%*
Missing	0	-	0	-
Total	1,985	32.0%	1,881	49.4%

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

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

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	1,917	31.6%	1,758	48.6%
Yes	68	34.8%*	123	60.6%*
Missing	0	--	0	--
Total	1,985	32.0%	1,881	49.4%

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

TABLE 24: SUMMARY SCORES BY THE PRESENCE OF BOOKS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	1,522	31.0%	1,393	49.6%
Yes	463	35.3%*	488	48.8%
Missing	0	--	0	--
Total	1,985	32.0%	1,881	49.4%

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

The final set of student questions (in Tables 25 to 27) pertained to children’s reading habits at home. In general, these habits made a difference in student scores in all cases for both grades. Having someone read to children at home, having children read to someone else at home, and children reading silently at home was related to higher reading scores.

TABLE 25: 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	894	28.2%	774	46.6%
Yes	1,041	35.8%*	1,074	51.8%*
Missing	50	--	33	--
Total	1,985	32.0%	1,881	49.4%

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

TABLE 26: SUMMARY SCORES BY CHILDREN READING TO SOMEONE ELSE AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	1,001	28.0%	818	46.5%
Yes	941	36.7%*	1,035	52.2%*
Missing	43	--	28	--
Total	1,985	32.0%	1,881	49.4%

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

TABLE 27: SUMMARY SCORES BY CHILDREN READING SILENTLY AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	421	25.2%	365	44.9%
Yes	1,520	34.3%*	1,466	51.0%*
Missing	44	--	50	--
Total	1,985	32.0%	1,881	49.4%

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

Teacher Questionnaires

With the smaller sample size, the analysis of the teacher questionnaires was limited to descriptive statistics, i.e., no group comparisons. A total of 131 teachers said that they taught grade 3 and 129 taught grade 5. Tables 28 and 29 provide information on teacher academic and professional qualifications, neither of which showed consistent patterns in the student scores.

TABLE 28: 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.	21	32.6%	31	50.5%
B.A./B.Sc.	48	35.9%	55	51.6%
F.A./F.Sc.	38	30.4%	28	51.5%
Matric	22	31.1%	15	46.0%
Missing	2	--	0	--
Total	131	32.0%	129	49.4%

TABLE 29: 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	45.3%	9	47.3%
B.Ed.	42	36.7%	50	53.2%
C.T.	1	31.2%	4	48.4%
P.T.C.	70	30.7%	61	49.3%
Missing	4	--	0	--
Total	131	32.0%	129	49.4%

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 lower or higher student scores (Tables 30 and 31). Again, small teacher sample sizes made drawing conclusions difficult.

TABLE 30: SUMMARY SCORES BY TEACHER AGE

Age Group in Years	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
40 and less	69	31.8%	70	49.2%
Between 41 and 50	48	33.1%	49	52.5%
51 and more	14	37.8%	7	50.7%
Missing	0	--	3	--
Total	131	32.0%	129	49.4%

TABLE 31: SUMMARY SCORES BY TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
10 or less	35	29.9%	26	52.2%
Between 11 and 20	44	32.7%	47	52.3%
Between 21 and 30	45	34.6%	45	49.5%
31 or more	4	50.5%	2	67.0%
Missing	3	--	9	--
Total	131	32.0%	129	49.4%

There were no significant differences in summary scores at grade 3 or grade 5 among teachers who attended or did not attend in-service training sessions (Table 32). Once more, any differences should be interpreted with caution due to the small sample size.

TABLE 32: SUMMARY SCORES BY TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	87	34.1%	79	51.7%
One time	31	30.7%	39	49.1%
Two times	6	24.5%	4	46.4%
Three times	4	38.2%	7	51.6%
Missing	3	--	0	--
Total	131	32.0%	129	49.4%

Head Teacher Questionnaires

Similar to the teachers, the sample size for the head teacher questionnaires was small, so data interpretations should be treated with caution. Tables 33 and 34 show reading scores by the head teachers' academic background. In general, the results show that higher teacher academic and professional qualifications may be related to student scores, but this observation was not statistically significant and could be due to the small sample sizes in F.A./F.Sc and Matriculation groups. In terms of academic and professional qualification, no discernible pattern relationships were found with summary reading scores.

TABLE 33: 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.	79	34.4%	79	50.7%
B.A./B.Sc.	41	30.0%	41	50.9%
F.A./F.Sc.	8	21.6%	8	42.9%
Matric	12	23.9%	12	41.6%
Missing	6	--	6	--
Total	140	32.0%	140	49.4%

TABLE 34: 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.	55	34.2%	55	51.7%
B.Ed.	49	35.9%	49	42.2%
C.T.	0	--	0	--
P.T.C.	30	23.9%	30	42.0%
Missing	6	--	6	--
Total	140	32.0%	140	49.4%

Tables 35 and 36 provide information on head teachers' experience and in-service training. For both grades, no discernible pattern was revealed in the head teachers' years of experience. Teachers attending 6-10 trainings had slightly higher scores than the other groups. Again, any differences should be interpreted with caution due to the small sample size.

TABLE 35: SUMMARY SCORES BY HEAD TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
2 or less	30	30.0%	30	50.7%
3 to 5	29	30.6%	29	44.4%
6 to 10	17	35.5%	17	53.5%
11 or more	52	31.2%	52	51.5%
Missing	0	--	0	--
Total	140	32.0%	140	49.4%

TABLE 36: SUMMARY SCORES BY HEAD TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	97	33.8%	97	49.9%
One time	27	29.8%	27	47.0%
Two times	10	30.0%	10	45.3%
More than two times	3	19.9%	3	42.6%
Missing	3	--	3	--
Total	140	32.0%	140	49.4%

Tables 37 and 38 provide data on head teachers' support to teachers in reading and the training that head teachers received in teaching reading. There were too few head teachers who reported not supporting teachers in reading (11); therefore the sample size is too small to make valid conclusions. Given that disclaimer, students of head teachers who did not receive training on teaching reading had slightly higher scores than of those who did receive the trainings.

TABLE 37: SUMMARY SCORES BY HEAD TEACHER SUPPORT OF TEACHERS IN READING

Support to Teachers	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	11	27.3%	11	41.0%
Yes	127	32.9%	127	50.3%
Missing	2	--	2	--
Total	140	32.0%	140	49.4%

TABLE 38: 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	97	33.8%	97	49.9%
Yes	40	29.1%	40	48.2%
Missing	3	-	3	--
Total	140	32.0%	140	49.4%

School Characteristics

The final section provides information on school characteristics (from the head teacher questionnaires) by student summary scores. As with the teacher and head teacher characteristics, most patterns appeared to be inconclusive (Tables 39 to 43). The few rural schools had higher reading scores than those in urban settings, and the boys schools performed better than the girls and mixed-gender schools. Schools with parent-teacher organizations posted higher scores for both grades. Schools with libraries (only 17 percent) had slightly higher scores at grade 3, but even scores at grade 5. Lastly, better infrastructure was related to higher student reading scores; scores increased with the addition of water, electricity, or toilets.

TABLE 39: SUMMARY SCORES BY SCHOOL LOCATION

School Gender	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Rural school	8	39.7%	8	60.8%
Urban school	62	28.4%	62	50.0%
Missing	0	--	0	--
Total	70	32.0%	70	49.4%

TABLE 40: SUMMARY SCORES BY SCHOOL GENDER

School Gender	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Boys school	34	33.3%	34	52.1%
Girls school	26	26.2%	26	44.7%
Mixed Gender	8	30.4%	8	52.5%
Missing	2	--	2	--
Total	70	32.0%	70	49.4%

TABLE 41: SUMMARY SCORES BY PTA/SMC/PTSMC/PTC

Parent Teacher Committee	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	29	25.7%	29	41.8%
Yes	39	32.3%	39	53.4%
Missing	2	--	2	--
Total	70	32.0%	70	49.4%

TABLE 42: SUMMARY SCORES BY PRESENCE OF A SCHOOL LIBRARY

School Library	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	58	30.0%	58	48.5%
Yes	12	33.3%	12	48.7%
Missing	0	--	0	--
Total	70	32.0%	70	49.4%

TABLE 43: 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	10	24.6%	10	45.5%
1	21	25.5%	21	46.4%
2	26	31.0%	26	48.5%
3	13	37.9%	13	54.6%
Missing	0	--	0	--
Total	70	32.0%	70	49.4%

CHAPTER 4: CONCLUSIONS AND RECOMMENDATIONS

This final chapter provides conclusions and recommendations from the Balochistan 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. The province also has one instructional language, Urdu. In addition, Balochistan 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, the lack of the requisite number of students per grade level was an issue. The actual sample of schools was 100 percent, and the actual sample of students reached 92 percent of the intended sample.
3. The EGRA test in Urdu administered in Balochistan was of good quality. 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 tests were such that it should be a strong measure of potential progress over time due to project-led interventions. Given the baseline scores, histograms, and box plots, the EGRA is expected to accurately measure the higher reading abilities that are expected at midline and endline. However, 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 the low actual enrollment of students in some schools. There was a high level of standardization reported by the quality control officers, which they attributed to the effective training process administered 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 Balochistan 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 Balochistan. 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 very good, showing that the EGRA tasks and items worked well in measuring reading constructs at both grade levels. The task and item statistics showed that EGRA discriminates well between low- and high-achieving students in both grades. They also showed that there is adequate room for growth by students in each grade level.
2. Grade 3 posted the highest scores in orientation to print, followed by familiar word reading, passage reading, and letter name recognition. The most difficult tasks for these students were comprehension (passage and listening) and phonics (non-word reading, letter sound knowledge, and phonemic awareness). At grade 5, the highest scores were in familiar word reading, passage reading, and orientation to print; whereas the most challenging tasks were comprehension (passage and listening) and letter sound knowledge.

3. There was substantial progression from grade 3 to grade 5 on the summary score (17 points) and on all task scores – the greatest gains were in familiar word, passage, and non-word reading. This progress was consistent across gender and treatment groups.
4. There were differences between boys and girls on the task and summary scores, but most of these differences were small. For example, the difference in the summary score was less than one point. The largest differences were in passage reading, favoring the girls (4.5 and 10.7 points for grades 3 and 5, respectively).
5. The full treatment group had higher scores on half of the tasks, with the light treatment group having higher scores on the other half. The largest difference was in letter sound knowledge (11 points) favoring the light group, but differences in all other tasks were small. This minor discrepancy will be corrected statistically at midline and endline by analyzing the growth for each group from baseline to midline and endline. Because this is a baseline report, the group differences will not be statistically tested at this time.
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 Balochistan.
7. The student questionnaire revealed three interesting findings. The first positive finding was that having reading materials and opportunities to read in the home seemed to have a positive effect on reading outcomes for both grades 3 and 5 students. Second, at grade 3, summary scores increased with relative age (younger than normal, normal, older than normal age); older students in the grade had higher reading scores. However, by grade 5 that advantage was no longer significant. Third, Balochistan students are performing well on the Urdu test considering only 1 percent of the students reported speaking Urdu as their primary language at home.
8. School, teacher, and head teacher questionnaire findings were mostly inconclusive, due to small sample sizes and the lack of variation in the scores that were related to their characteristics. For example, an analysis of student scores by teacher and head teacher education, certification, age, experience, and attendance at in-service trainings found no consistent patterns relating to lower or higher student scores. For the schools, better infrastructure was associated with better student reading scores.

Evaluation Recommendations

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

1. The EGRA instruments proved to be of high quality, and equivalent versions of those tools should be developed – through trans-adaptation, piloting, and revision – for the midline and endline assessments so that progress can be accurately measured over time.
2. The EGRA items and tasks had good reliability values and covered the low-to-middle difficulty range. At baseline, the reading scores were relatively low for both grades, and show room for growth. In addition, histograms and box plots provided evidence that the tool is expected to measure higher levels of reading that are anticipated due to project-led interventions. Therefore,

the baseline data indicates that the EGRA is appropriate for measuring increases in reading ability at midline and endline.

3. The sampling was reasonable in terms of finding a balance between the resources available, the required sample size, and the geographic coverage. It should be maintained in the midline and endline, i.e., keep the same districts and schools, along with the methods at the school level.
4. Because of the variability among the districts' academic calendars, the instructional time from the start of school to the EGRA administration varied by five months among the districts in Balochistan. Students make great gains in reading during the primary grades. 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 Balochistan should occur in October, thus matching the baseline timeframe and standardizing the instructional time across the study.
5. The systems for field data collection should be replicated, with the same systems for recruitment and training for the master trainers, field supervisors, QCOs, and enumerators as used in the baseline.
6. The data entry system should continue to be used, with the same systems for recruitment and training of data entry supervisors and operators, along with implementation through networked computers, double data entry, and reconciliation of errors.
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 Balochistan baseline was successful in providing accurate data on which to base decisions for implementation of the PRP interventions, and also for tracking student reading progress over time. It provides a solid foundation for the midline and endline assessments.

ANNEXES

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

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

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

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

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

Annex I: Complete Item Statistics by Grade

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

Previously, we presented task statistics (Chapter 3, Table 8) with explanations of how they are calculated. These item statistics are calculated in the same way. They show the difficulty and quality of the items. Recall that when constructing a test, we strive for tasks and items that have difficulty values (p-values) that are spread across the range from about 0.05 to 0.90 and quality values (item-total correlations) of at least 0.20. The difficulty values ranged from 0.06 to 0.73 for grade 3 and 0.12 to 0.78 for grade 5, indicating an adequate range of item difficulties. A total of 20 and 21 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 A1: COMPLETE ITEM STATISTICS BY GRADE

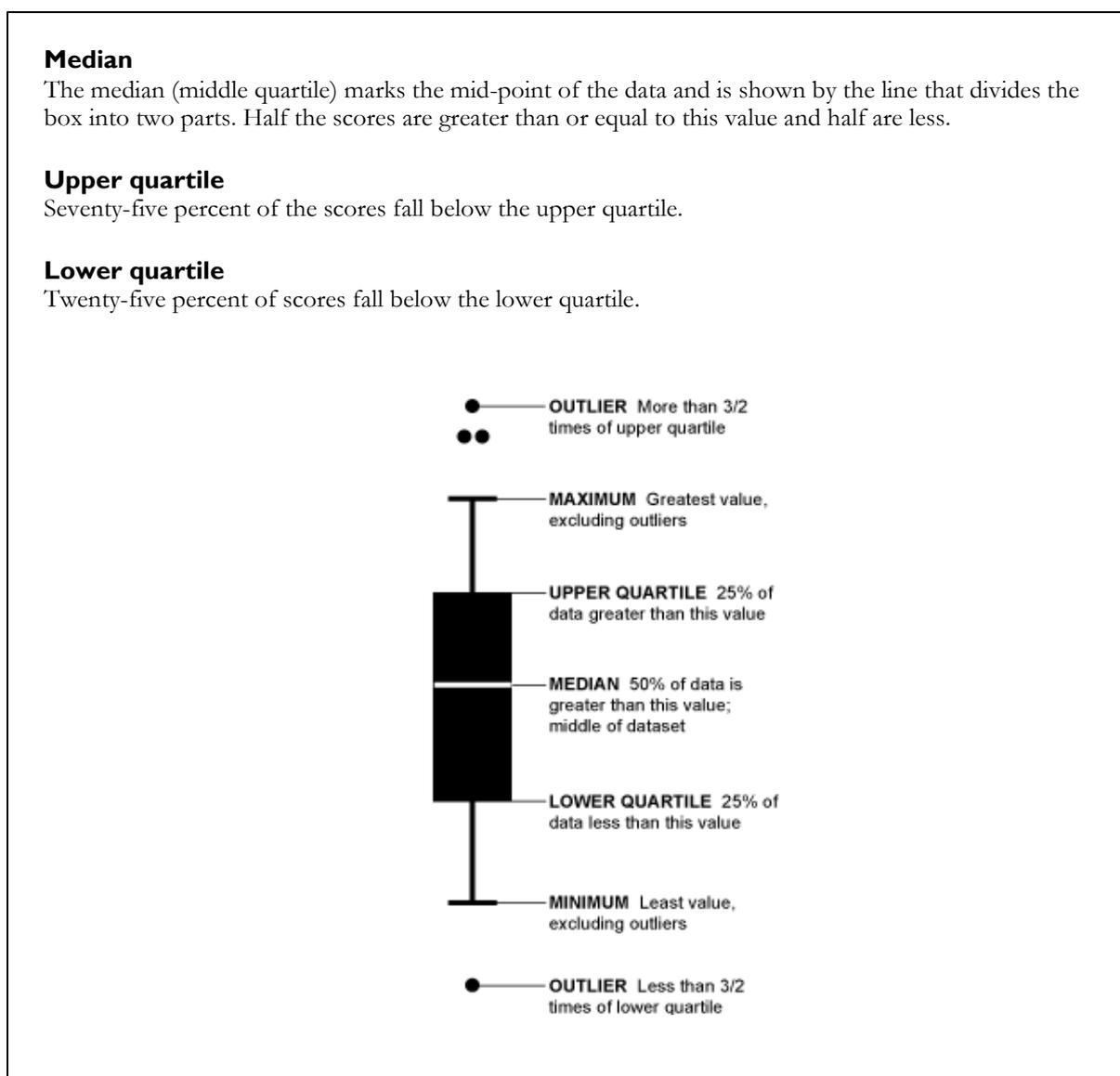
Task (Subtest)	Item	Grade 3		Grade 5	
		P-Value	Item-Total	P-Value	Item-Total
1. Orientation to print (untimed)	Q1	0.73	0.33	0.78	0.28
	Q2	0.72	0.41	0.71	0.26
	Q3	0.40	0.19	0.45	0.16
	Q4	0.11	0.06	0.26	0.11
	Q5	0.56	0.18	0.69	0.20
2. Letter name recognition (timed)	--				
3. Phonemic awareness (untimed)	Q1	0.46	0.45	0.61	0.43
	Q2	0.20	0.38	0.32	0.45
	Q3	0.27	0.39	0.39	0.43
	Q4	0.21	0.33	0.33	0.39
	Q5	0.33	0.40	0.43	0.42
	Q6	0.41	0.42	0.51	0.40
	Q7	0.17	0.32	0.27	0.39
	Q8	0.24	0.41	0.37	0.45
	Q9	0.22	0.38	0.31	0.43
	Q10	0.43	0.48	0.54	0.44
4. Letter sound knowledge (timed)	--				
5. Familiar word reading (timed)	--				
6. Non-word reading (timed)	--				
7a. Passage reading (timed)	--				
7b. Passage comprehension (untimed)	Q1	0.12	0.59	0.36	0.60
	Q2	0.12	0.50	0.255	0.51
	Q3	0.06	0.36	0.18	0.43
	Q4	0.18	0.61	0.44	0.60
	Q5	0.18	0.61	0.47	0.61
8. Listening comprehension (untimed)	Q1	0.22	0.44	0.39	0.37
	Q2	0.06	0.32	0.12	0.27
	Q3	0.30	0.42	0.56	0.34

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



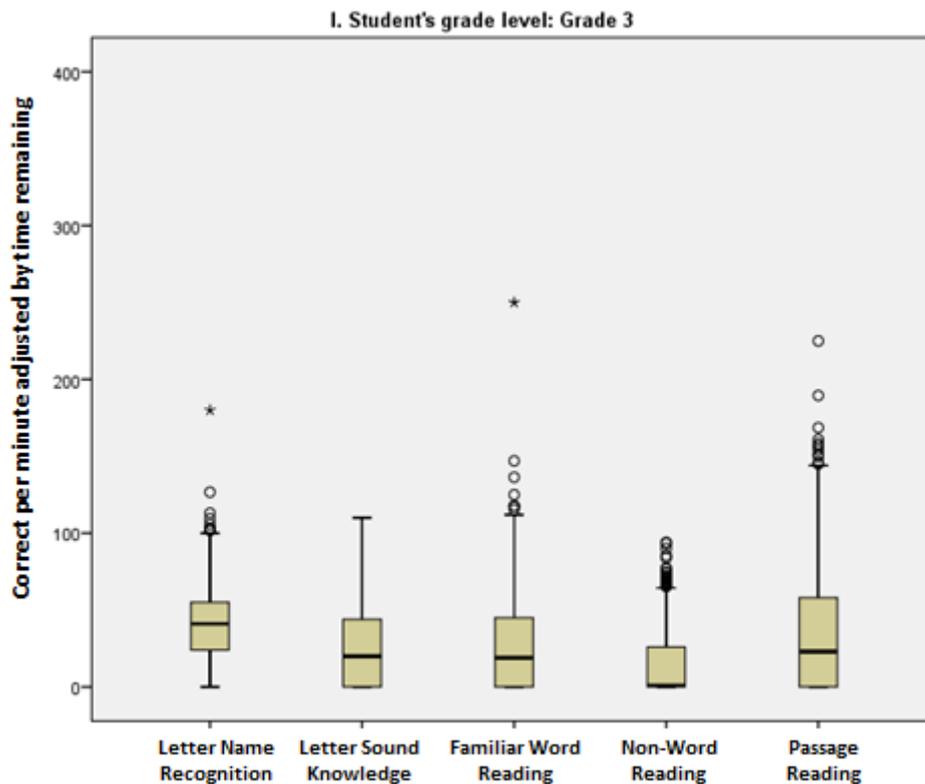
Box plots are presented below (Figures A2 and A3) for the results by grade level on the five fluency (timed) tasks: letter name recognition (task 2), letter sound knowledge (task 4), familiar word reading (task 5), non-word reading (task 6), and passage reading (task 7a).

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 30 (letter name recognition) items per minute. It shows that the students had better knowledge of letter names than grapheme-morpheme correspondence.

The variation (i.e., the range of scores, without outliers) for each of the tasks varied from about 60 (non-word reading) to about 130 (passage reading). It shows that the scores were more spread out when reading connected words than sounding out pseudo-words.

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



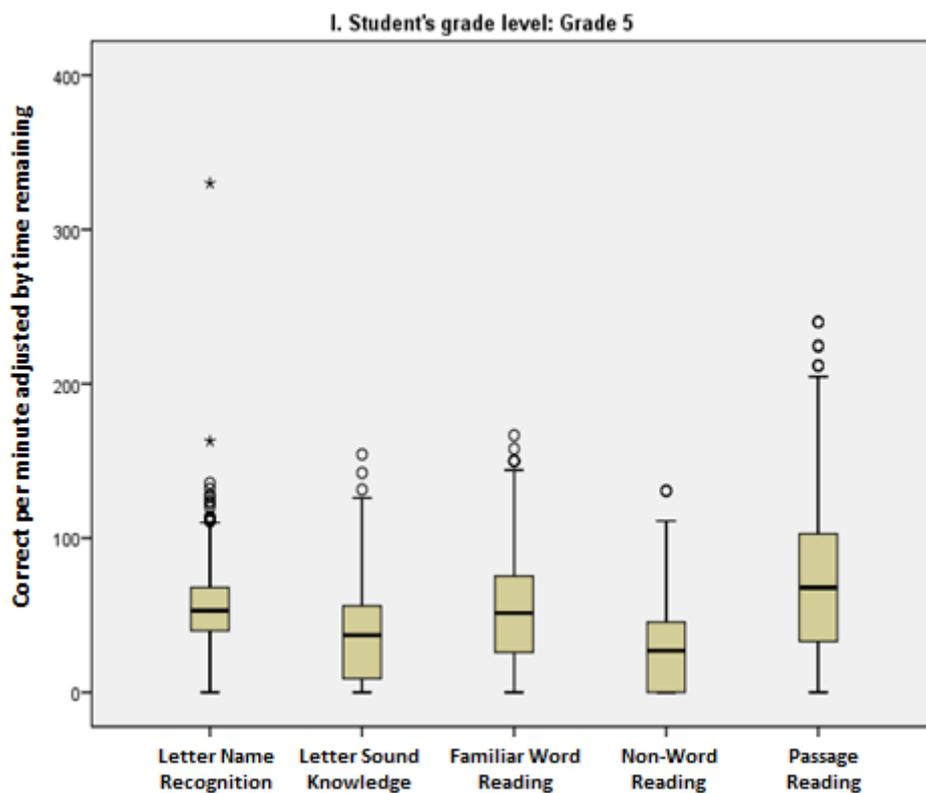
Grade 5

For grade 5, the central tendency (the median speed) for each of the tasks ranged from about 30 (letter sound knowledge) to about 80 (passage reading) items per minute. It shows that the students had more fluency with reading connected words than with phonics.

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

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

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



Annex 3: Examples of Fluency Score Threshold Calculations

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

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

Fluency using an 80 percent comprehension threshold

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

At grade 3, the mean reading speed on the passage reading task (Task 7a) for students who scored 80 percent on the passage comprehension task (Task 7b) was 83.5 (rounded to 84). With this method, 84 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 106.0 (or 106). Then 106 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 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	41.0%	0	17.6%
Non-Fluent Reader	1 to 83	49.2%	1 to 105	58.7%
Fluent Reader	84 and above	9.8%	106 and above	23.7%
Total	--	100.0% ¹	--	100.0%

Note that the around half 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	41.0%	17.6%
Early Reader	1 to 40	21.0%	10.9%
Intermediate Reader	41 to 80	26.9%	30.9%
Fluent Reader	81 to 120	8.7%	26.1%
Advanced Reader	121 and above	2.3%	14.6%
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 7 percent of the advanced readers have 100 percent comprehension. Also, at grade 3, 31 percent (24 percent + 7 percent) of the advanced readers have comprehension levels of 80 percent or above.

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	78%	16%	5%	1%	0%	0%	100%
Intermediate Reader	41 to 80	41%	16%	18%	14%	8%	2%	100%
Fluent Reader	81 to 120	18%	14%	27%	23%	14%	4%	100%
Advanced Reader	121 and above	17%	11%	20%	22%	24%	7%	100%

FIGURE A4: GRADE 3 READING FLUENCY AND COMPREHENSION

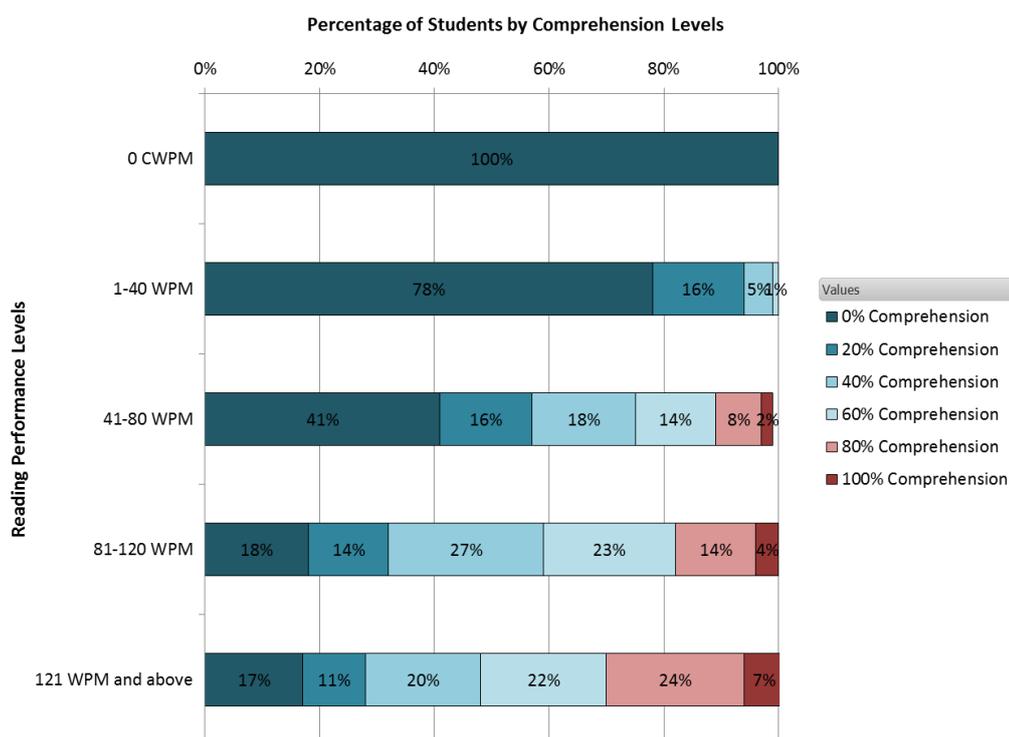
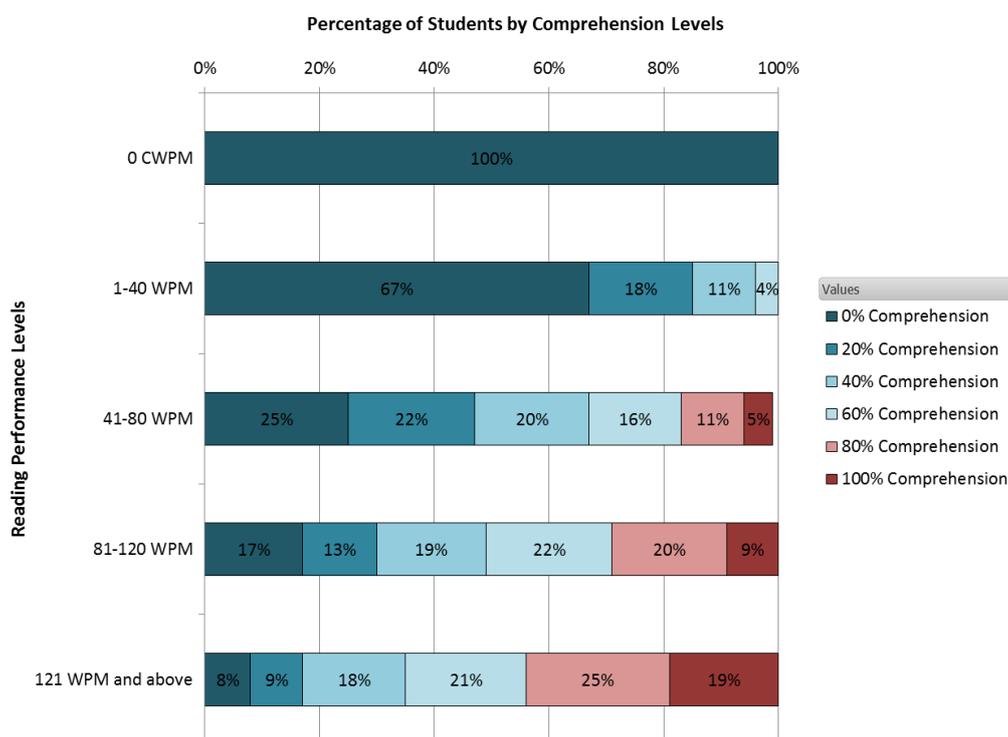


TABLE A5: GRADE 5 READING FLUENCY AND COMPREHENSION

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	67%	18%	11%	4%	0%	0%	100%
Intermediate Reader	41 to 80	25%	22%	20%	16%	11%	5%	100%
Fluent Reader	81 to 120	17%	13%	19%	22%	20%	9%	100%
Advanced Reader	121 and above	8%	9%	18%	21%	25%	19%	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 (78 percent at grade 3 and 67 percent at grade 5) had 0 percent comprehension. None of them achieved 80 percent comprehension.
- Intermediate Readers (41-80 CWPM) – A substantial proportion of the intermediate readers (41 percent at grade 3 and 25 percent at grade 5) had 0 percent comprehension. A minority of them (10 percent at grade 3 and 16 percent at grade 5) achieved at least 80 percent comprehension.
- Fluent Readers (81-120 CWPM) – About one out of every five fluent readers at each grade level had 0 percent comprehension (18 percent at grade 3 and 17 percent at grade 5). Less than one-third of them (18 percent at grade 3 and 29 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. Fewer than half of them (31 percent at grade 3 and 44 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.