



USAID
FROM THE AMERICAN PEOPLE



EARLY GRADE READING ASSESSMENT BASELINE REPORT

ISLAMABAD CAPITAL TERRITORY

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 ISLAMABAD CAPITAL TERRITORY

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

Monitoring and Evaluation Program (MEP)

DISCLAIMER

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

ACKNOWLEDGEMENTS

We would like to thank the Education team of USAID/Pakistan for their forward planning to be able to collect baseline data before the roll out of the two important reading programs. Their support and responsiveness under a demanding timeline made this study possible. We would also like to thank the Government of Pakistan, the Ministry of Capital Administration and Development, and the Federal Directorate of Education for their support of this activity. Finally, this effort would not have been possible without the dedication of our field teams of quality control officers and our local data collection partner, the Institute for Social and Applied Policy Studies (I-SAPS).

CONTENTS

Executive Summary	1
Chapter 1: Introduction	6
Chapter 2: Design and Methodology	8
Chapter 3: Findings and Results	17
Chapter 4: Conclusions and Recommendations	44
Annexes	47
Annex 1: Complete Item Statistics by Grade	48
Annex 2: Box Plots for Phonics and Reading-rate Fluency Tasks	50
Annex 3: Examples of Fluency Score Threshold Calculations.....	55
Annex 4: Distribution of Reading Fluency and Comprehension Scores using Fixed Intervals.....	58

List of Tables and Figures

Table 1: Round 1 Timeline (January 2103 to May 2014).....	10
Table 2: Sample Schools by Gender and Location.....	11
Table 3: Reliability Estimates	14
Table 4: EGRA Score Ranges and Calculations	15
Table 5: Example of EGRA Percent Correct and Summary Scores	16
Table 6: Example of EGRA Timed Task Scores.....	16
Table 7: Actual Student Sample by Grade and Gender.....	17
Table 8: English Task Statistics	18
Table 9: Urdu Task Statistics.....	18
Table 10: English Scores by Grade and Task.....	20
Table 11: Urdu Scores by Grade and Task.....	21
Table 12: English Scores by Grade and Gender.....	22
Table 13: Urdu Scores by Grade and Gender	23
Table 14: English Baseline Maximum Scores on Fluency (Timed) Tasks	25
Table 15: Urdu Baseline Maximum Scores on Fluency (Timed) Tasks	25
Table 16: English Phonics and Reading-Rate Fluency Task Means by Grade	26
Table 17: English Phonics and Reading-Rate Fluency Task Means by Grade and Gender	26
Table 18: Urdu Phonics and Reading-Rate Fluency Task Means by Grade.....	26
Table 19: Urdu Phonics and Reading-Rate Fluency Task Means by Grade and Gender	27
Table 20: English Summary Scores by Student Age	28
Table 21: Urdu Summary Scores by Student Age.....	28
Table 22: English Summary Scores by Reading the Quran at Home.....	28
Table 23: Urdu Summary Scores by Reading the Quran at Home.....	29
Table 24: English Summary Scores by the Presence of a Library at the School	29
Table 25: Urdu Summary Scores by the Presence of a Library at the School.....	29
Table 26: English Summary Scores by the Presence of Newspapers at Home	30
Table 27: Urdu Summary Scores by the Presence of Newspapers at Home	30
Table 28: English Summary Scores by the Presence of Magazines at Home	30
Table 29: Urdu Summary Scores by the Presence of Magazines at Home.....	30
Table 30: English Summary Scores by the Presence of Books at Home.....	31
Table 31: Urdu Summary Scores by the Presence of Books at Home.....	31
Table 32: English Summary Scores by Children Having Someone Read to Them at Home.....	31
Table 33: Urdu Summary Scores by Children Having Someone Read to Them at Home.....	32
Table 34: English Summary Scores by Children Reading to Someone Else at Home	32

Table 35: Urdu Summary Scores by Children Reading to Someone Else at Home	32
Table 36: English Summary Scores by Children Reading Silently at Home.....	32
Table 37: Urdu Summary Scores by Children Reading Silently at Home.....	33
Table 38: English Summary Scores by Children Having a Computer at Home.....	33
Table 39: Urdu Summary Scores by Children Having a Computer at Home.....	33
Table 40: English Summary Scores by Teacher Academic Qualification.....	34
Table 41: Urdu Summary Scores by Teacher Academic Qualification	34
Table 42: English Summary Scores by Teacher Professional Qualification.....	34
Table 43: Urdu Summary Scores by Teacher Professional Qualification.....	35
Table 44: English Summary Scores by Teacher Age.....	35
Table 45: Urdu Summary Scores by Teacher Age.....	35
Table 46: English Summary Scores by Teacher Experience.....	36
Table 47: Urdu Summary Scores by Teacher Experience.....	36
Table 48: English Summary Scores by Teacher In-Service Training.....	36
Table 49: Urdu Summary Scores by Teacher In-Service Training.....	37
Table 50: English Summary Scores by Head Teacher Academic Qualification	37
Table 51: Urdu Summary Scores by Head Teacher Academic Qualification.....	37
Table 52: English Summary Scores by Head Teacher Professional Qualification	38
Table 53: Urdu Summary Scores by Head Teacher Professional Qualification	38
Table 54: English Summary Scores by Head Teacher Experience	38
Table 55: Urdu Summary Scores by Head Teacher Experience	39
Table 56: English Summary Scores by Head Teacher In-Service Training.....	39
Table 57: Urdu Summary Scores by Head Teacher In-Service Training.....	39
Table 58: English Summary Scores by Head Teacher Support to Teachers in Reading.....	40
Table 59: Urdu Summary Scores by Head Teacher Support to Teachers in Reading	40
Table 60: English Summary Scores by Head Teacher Training in Teaching Reading.....	40
Table 61: Urdu Summary Scores by Head Teacher Training in Teaching Reading.....	40
Table 62: English Summary Scores by School Gender	41
Table 63: Urdu Summary Scores by School Gender.....	41
Table 64: English Summary Scores by School Location	41
Table 65: Urdu Summary Scores by School Location	42
Table 66: English Summary Scores by PTA/SMC/PTSMC/PTC	42
Table 67: Urdu Summary Scores by PTA/SMC/PTSMC/PTC.....	42
Table 68: English Summary Scores by Presence of a School Library	42
Table 69: Urdu Summary Scores by Presence of a School Library	43
Table 70: English Summary Scores by Infrastructure (Drinking Water, Electricity, Toilets).....	43
Table 71: Urdu Summary Scores by Infrastructure (Drinking Water, Electricity, Toilets).....	43
Table A1: English Item Statistics by Grade.....	48
Table A2: Urdu Item Statistics by Grade.....	49
Table A3: English Thresholds for CWPM with 80 Percent Comprehension.....	55
Table A4: Urdu Thresholds for WCPM with 80 Percent Comprehension.....	56
Table A5: English Thresholds for CWPM with Fixed Intervals.....	56
Table A6: Urdu Thresholds for CWPM with Fixed Intervals.....	57
Table A7: Grade 3 Reading Fluency and Comprehension, English.....	58
Table A8: Grade 5 Reading Fluency and Comprehension, English.....	59
Table A9: Grade 3 Reading Fluency and Comprehension, Urdu	60
Table A10: Grade 5 Reading Fluency and Comprehension, Urdu	61
Figure 1: Evaluation Design.....	8
Figure 2: English Grade 3 Summary Scores	19
Figure 3: Urdu Grade 3 Summary Scores	19

Figure 4: English Grade 5 Summary Scores	19
Figure 5: Urdu Grade 5 Summary Scores	19
Figure 6: English Scores by Grade and Task.....	20
Figure 7: Urdu Scores by Grade and Task.....	21
Figure 8: English Grade 3 Scores by Task and Gender.....	22
Figure 9: English Grade 5 Scores by Task and Gender.....	23
Figure 10: Urdu Grade 3 Scores by Task and Gender.....	24
Figure 11: Urdu Grade 5 Scores by Task and Gender.....	24
Figure A1: Understanding Boxplots	50
Figure A2: Phonics and Reading-Rate Fluency Box Plots for Grade 3, English.....	51
Figure A3: Phonics and Reading-Rate Fluency Box Plots for Grade 5	52
Figure A4: Phonics and Reading-Rate Fluency Box Plots for Grade 3, Urdu.....	53
Figure A5: Phonics and Reading-Rate Fluency Box Plots for Grade 5, Urdu.....	54
Figure A6: Grade 3 Reading Fluency and Comprehension, English.....	59
Figure A7: Grade 5 Reading Fluency and Comprehension, English.....	60
Figure A8: Grade 3 Reading Fluency and Comprehension, Urdu.....	61
Figure A9: Grade 5 Reading Fluency and Comprehension, Urdu.....	62

ACRONYMS

AJK	Azad Jammu and Kashmir
B.A.	Bachelor of Arts
B.Sc.	Bachelor of Science
C.T.	Certificate of Teaching (Grade 12 plus FA/FSC Certificate)
EGRA	Early Grade Reading Assessment
F.A.	Intermediate College (Grade 12) Certificate in Arts
FATA	Federally Administered Tribal Areas
F.Sc.	Intermediate College (Grade 12) Certificate in Sciences
GB	Gilgit-Baltistan
ICT	Islamabad Capital Territory
I-SAPS	Institute for Social and Applied Policy Studies
KP	Khyber Pakhtunkhwa
M.A.	Master of Arts
Matric	Secondary School (Grade 10) Certificate (Matriculation)
M.Ed.	Master of Education
M.Sc.	Master of Science
MOE	Ministry of Education
MSI	Management Systems International
MT	Master Trainers
NEAS	National Education Assessment System
NEMIS	National Education Management Information System
PRP	Pakistan Reading Project
P.T.C.	Primary Teaching (Grade 12) Certificate
QCO	Quality Control Officer
SPSS	Statistical Package for the Social Sciences
SQL	Structured Query Language
SRP	Sindh Reading Project
STS	School-to-School International
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

Overview

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

To measure results from PRP and SRP, a rigorous external evaluation is being conducted. This report covers the baseline assessment in the Islamabad Capital Territory. In May 2013, ICT, along with AJK and GB, was part of Round 1 of the baseline data collection; data from Pakistan's other five provinces/areas/territories (hereafter referred to as provinces) were collected in Rounds 2 and 3 in September and October 2013, respectively. The following activities were carried out for all of the provinces, including ICT: 1) design, 2) sampling, 3) instrumentation, 4) planning, 5) training, 6) implementation, 7) analysis, and 8) reporting.

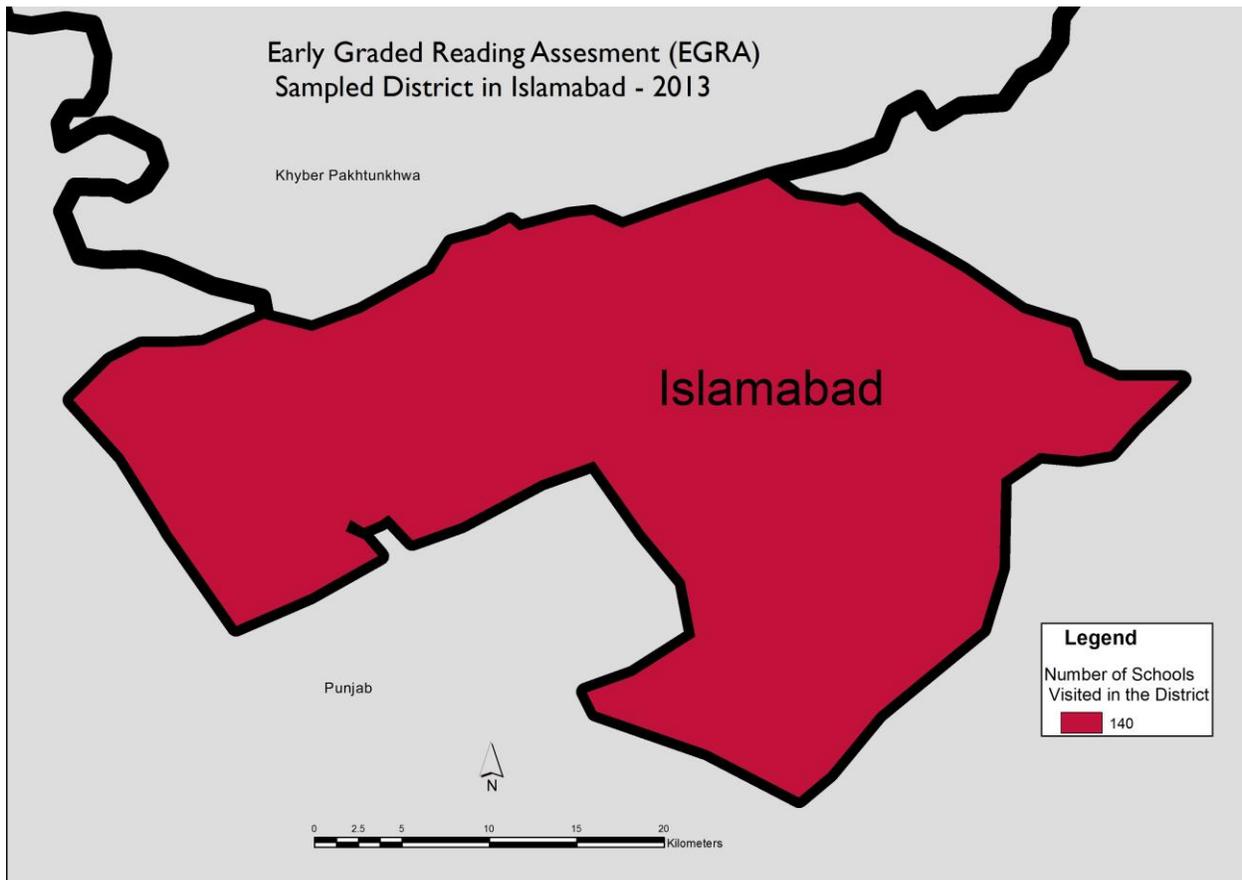
The external evaluation design, which was developed prior to the baseline assessment, was tailored to the implementation of the PRP and SRP in each province. In most of the provinces, a quasi-experimental design will be used, with two treatment groups: "full treatment" and "light treatment." The full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

In accordance with the USAID evaluation guidelines, students at two selected grade levels – grades 3 and 5 – were assessed at three time points: baseline, midline, and endline. An internationally accepted assessment tool, the Early Grade Reading Assessment (EGRA), will be individually administered to over 30,000 students in over 1,000 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 students' 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 ICT, all activities were completed by the end of September, including a draft report. The results were presented and discussed at a consultative meeting in Islamabad on September 24, 2013. Representatives from the provincial Ministry of Education (MOE), USAID, PRP, and the contractors (MSI and STS) attended the consultation. Revisions were then made to this report based on the discussions between the stakeholders.

Map of Sampled Sector



Key Points

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

Implementation

1. All five sectors of ICT were selected for “full treatment” during the initial consultative meetings between the MOE and USAID in January 2013. In ICT, there will not be a comparison of groups to determine the effects of full treatment above and beyond those of the “light treatment”; rather, the results at the baseline will be compared against those in the midline and endline for the full treatment group only.
2. Since ICT has a reasonably large percentage of students attending English-medium schools, the ICT baseline will have two groups: English-medium and Urdu-medium. Separate samples of 70 schools each will be taken from these two groups. The groups will not be formally compared to each other due to differences in language structure (and perhaps language difficulty).
3. The baseline data were collected in a random sample of schools across ICT’s five sectors (districts). A random sample of male and female schools was selected, followed by a random sample of grades 3 and 5 students within those schools.
4. The EGRA tools, which have been administered in various forms in over 40 countries, were successfully adapted for use in Pakistan. These included individually administered reading tests for

students, along with questionnaires for students, teachers, and head teachers. The Urdu version of the tools was piloted in AJK, ICT, and KP. The English version was piloted in ICT.

5. A total of 140 schools, with 70 schools from each language (English and Urdu) were selected for the baseline.
6. The results from this sample are presented in this report as a generalized view of the reading levels for students in the ICT schools. Please note that sector comparisons are not possible because the sectors were not evenly sampled; the number of sampled schools varied by sector, and the sample sizes are limited for each sector.
7. The EGRA testing window for ICT was May 2013, and all schools were covered during this time period.
8. The assessment tools were successfully administered in the schools in the sectors as follows (with a percentage of the target reached in parentheses): 140 schools (100.0 percent) to 4,105 students (97.7 percent), 246 teachers (87.9 percent), and 139 head teachers (99.3 percent).
9. The validity and reliability of the tools was acceptable. Validity was assured through the adaptation process, which involved 17 educationists from throughout the country who participated in a workshop in Islamabad. Reliability was assured through the high quality of the assessment tasks and the standardized administration of the tools. Reliability estimates (of internal consistency) were calculated using the coefficient alpha.
10. The data entry and data cleaning process followed international standards. All student data were entered twice into two separate databases. These databases were then compared, with a resulting discrepancy rate of less than 1 percent. All data were reconciled across the two databases and with the assessment booklets. A clean data file was produced for analysis.
11. In the analysis phase, scores were calculated in three ways: 1) percentage correct scores for the reading tasks, 2) average percentage correct (grand means) for reading summary scores, and 3) adjusted raw scores for the timed reading tasks. 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 2,021 grade 3 students and 2,084 grade 5 students. The reliability estimates were acceptable for both grades (English: alpha = 0.79 for grade 3 and 0.77 for grade 5; Urdu: alpha = 0.83 for grade 3 and 0.82 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. In each language, 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 but one of the task scores at grades 3 and 5 (orientation to print) had item-total correlations equal to or greater than 0.20, indicating good discrimination quality for these tasks. (Complete item statistics are listed in Annex 1.)
3. Students had the most difficulty with phonics-related tasks such as letter sound knowledge and non-word reading. Passage and listening comprehension were also areas of weakness. On the other hand, students did relatively well on familiar word reading and passage reading (fluency). There was also substantial progression from grade 3 to grade 5 on some of the tasks and for the summary scores.

4. Female students had higher scores, in general, than did their male counterparts. Areas such as letter name recognition, familiar word reading, passage reading, and passage comprehension were areas of particular strength for the females over the males in ICT. Boys tended to perform better than girls only in orientation to print.
5. Students were timed on five tasks as they read words or passages. These tasks were categorized into phonics (letter name recognition, letter sound knowledge, and non-word reading) and reading-rate fluency (familiar word and passage reading). Students in both grades had lower phonics scores than reading-rate fluency scores. Moreover, gains from grade 3 to grade 5 were lower for phonics than for reading-rate fluency tasks. Passage reading in English was nearly 40 points higher in grade 5 than in grade 3, and nearly 45 points higher in Urdu. Although the passage was designed for the grade 3 level, this difference shows that the reading 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.
6. Mastery of phonics, such as letter sound knowledge, phonemic awareness, and non-word reading, should help the students become better overall readers. It is clear that these types of knowledge and skills are not receiving an appropriate emphasis in ICT schools.
7. Questionnaire findings were mostly inconclusive, due to small sample sizes and the lack of variation in the scores that were related to the student, teacher, and head teacher characteristics. For the students, one of the positive findings was that attending a grade at an appropriate age seemed to have a positive effect on reading outcomes. In terms of the home environment, the presence of reading materials and the availability of a reading companion had some effects on outcomes, though they were limited.

Evaluation Recommendations

Given the success of the baseline assessment in ICT (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 sectors and schools, along with the sampling methods at the school level.
4. The systems for field data collection should be replicated, with the same systems for recruitment and training for the master trainers (MTs), field supervisors, quality control officers (QCOs), and enumerators as used in the baseline.
5. The data entry system should continue to be used, with the same systems for recruitment and training of data entry supervisors and operators, along with implementation through networked computers, double data entry, and reconciliation of errors.

6. The analysis should follow the same procedures with calculations of 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.
7. Reading proficiency levels should be created to provide educators and other stakeholders with meaningful results. Most parents and educators better understand reading achievement in useful terms or levels, such as emerging, proficient, or advanced, rather than interpreting a percent-correct test score that may differ by test or reading passage difficulty. Education officials are encouraged to select specific EGRA scores to serve as levels of reading proficiency for both grades. Percent correct for each task, summary score, as well as fluency rates are recommended for this purpose. The baseline EGRA data can be used for establishing these reading proficiency levels.
8. Finally, it may be advisable to add items to the student, teacher, and head teacher questionnaires for collecting data on PRP- and SRP-supported interventions so that student scores can be correlated with these indicators.

CHAPTER I: INTRODUCTION

The Pakistan Reading Project (PRP) and the Sindh Reading Program (SRP) are two five-year initiatives funded by USAID. The projects/programs will cover over 40,000 government schools in Pakistan's eight provinces/areas/territories (hereafter referred to as provinces). PRP is targeting improved reading for 910,000 children in AJK, Balochistan, FATA, GB, ICT, KP, and Sindh, while the SRP is targeting improved reading and mathematics for 750,000 children in Sindh. Targets will be achieved through support for 1) improved policies, laws, and guidelines for teachers and educational administrators, and 2) improved reading instruction for children in primary grades. Some districts 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 is taking 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 – are being assessed throughout Pakistan so that independent baselines can be established in each province. Students at the same grade levels will be assessed at the midline and endline time points to evaluate the success of the interventions, taking into account the two treatment groups.

This report covers ICT. ICT, along with AJK and GB, was part of Round 1 of the baseline data collection in May 2013; data from Pakistan's other five provinces were collected in September (Round 2) and October (Round 3) 2013. The following activities were planned for all of the provinces, including ICT:

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, this was complemented by a quasi-experimental design with the two treatment groups.
2. Sampling – The sampling plan for ICT enabled the collection of student reading assessment data that were representative of the language, grade levels, gender, and urban/rural zones. During the consultation process, the MOE and USAID decided to intervene in all five of ICT's sectors, so schools were selected throughout the territory. Schools were apportioned according to language of instruction (Urdu/English), location (urban/rural) and gender (boys/girls). While the samples were selected by language (70 Urdu-medium schools and 70 English-medium schools) and then by gender within each language (35 boys and 35 girls schools), it was not possible to have an equal distribution by location since most of the English-medium schools were urban and almost all of the Urdu-medium schools were rural.
3. Instrumentation – EGRA tools were developed, with tests at the grade 3 level in English, Sindhi, and Urdu, and questionnaires for teachers, head teachers and students. Model EGRA instruments were trans-adapted, piloted, revised, and finalized for use in Pakistan.
4. Planning – A field administration plan was developed for the baseline administration that would ensure the reliability of the data collected. The plan specified the timeline, training, logistics, field activities, supervision, data entry, analysis, reporting, and quality control.
5. Training – Workshops were conducted to train all master trainers, supervisors, enumerators, and QCOs. Enumerators and supervisors were observed to ensure clear comprehension and skills adequate to implement the EGRA tools.
6. Implementation – The baseline survey was implemented according to the plan. It ensured that all of the field activities took place in a standardized manner, as verified by the QCOs. The fieldwork was followed by data entry and preparation of a clean data file.

7. Analysis – Data were analyzed using spreadsheet (Excel) and statistical (SPSS) software. Experienced statisticians/psychometricians conducted the analysis, produced data tables and graphs, and ensured quality control.
8. Reporting – Provincial level reports were produced. A reporting template was developed according to guidelines from the USAID contract. These reports will be disseminated to the provincial education authorities.

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

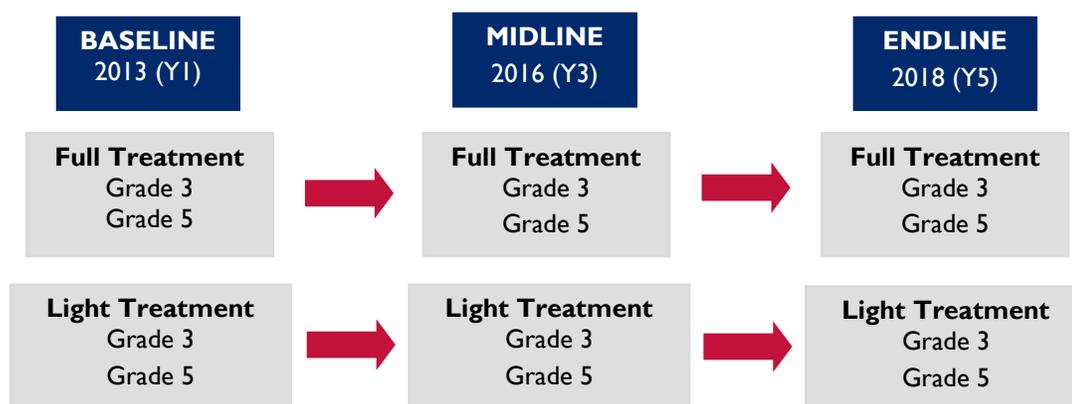
CHAPTER 2: DESIGN AND METHODOLOGY

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

Evaluation Design

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

FIGURE 1: EVALUATION DESIGN



Districts for the “full” and “light” treatment groups were pre-selected by the provincial MOE and USAID during consultations 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 will be used. In this design, any differences in scores at baseline (and midline) 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, while most districts have the two treatment groups, two of the provinces – AJK and ICT – will receive full treatment across all districts (sectors), and another province – FATA – will have full treatment in some districts but no treatment (and no data collection) in the others. In ICT, all five of the sectors will be covered by the PRP full treatment reading intervention. With this design, there will be no counterfactual (i.e., light treatment) for the ICT reading interventions.

In ICT, students were tested in English and Urdu, their main languages of instruction. Equal numbers of male and female schools, i.e., 35 male and 35 female schools per language, were sampled for the EGRA testing; some mixed schools were included in the sample, but only boys or girls were selected from these schools, and thus they were considered as either male or female schools. The sampling design met the USAID requirements of adequate sample size and equal gender representation (see the sampling section below).

Timeline

The ICT baseline, like the other provinces for Round 1, was conducted according to a timeline that started in January and ended in September 2013, with final submission of reports to USAID in October. The reports may then be distributed to the provincial MOEs and other stakeholders as appropriate.

The process began with the planning and design of activities, including creating preliminary sampling designs, selecting model EGRA tasks, recruiting staff, and budgeting/contracting. This was followed in February by provincial consultations, including those for ICT. In February to April, the EGRA team, with participation from ICT and other provinces, then prepared, piloted, and revised the EGRA tools and conducted the sector/school sampling. The data collection in ICT took place in May, followed by the data entry, analysis, and reporting in June to September, including the presentations to the MOE and USAID in late September. The final report for ICT was submitted in May 2014 (see Table 1 below).

TABLE I: ROUND I TIMELINE (JANUARY 2103 TO MAY 2014)

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

Sampling

The sampling for Round 1 started in January with the selection of the treatment sectors by the provincial MOE and USAID. The EGRA team conducted the school sampling in March and April. This included developing the sampling requirements, verifying the sample in the field, and finalizing the sample. The findings were provided in the sampling report for USAID.¹ As mentioned above, in ICT, all sectors were selected for full treatment. The sampling for ICT, as detailed in the sampling report, is briefly summarized in the following sub-sections of this report.

Sampling Requirements

The sampling frame for ICT included all five of the sectors. During the consultation process, the MOE and USAID decided to intervene in all of ICT's sectors, so schools were selected throughout the territory.

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

Sampling Process and Field Verification

Given the relatively small size of ICT, there was no need to divide the sample into zones. Besides having separate samples by medium of instruction, the territory was stratified at the "location" level, i.e., schools were allocated by rural and urban. As seen in Table 2 below, there were relatively few urban schools in the area of Urdu-medium schools, but over three-fourths of the English-medium schools were in urban areas. After sampling the 140 schools in ICT, an additional 20 male and 20 female schools were selected as replacements, if needed. Note that mixed schools may have been selected for some replacement schools due to not having enough options for replacement schools of strictly one gender. However, only students from the respective genders were included in those samples (i.e. if a mixed school was selected to replace a female school, only females were sampled).

TABLE 2: SAMPLE SCHOOLS BY GENDER AND LOCATION

Language	Location	Schools	Percentage	Sample Schools		Replacement Schools	
				Boys	Girls	Boys	Girls
English	Rural	23	24	8	8	2	2
English	Urban	74	76	27	27	8	8
English Total		97	100	35	35	10	10
Urdu	Rural	225	96	34	34	10	10
Urdu	Urban	9	4	1	1	0	0
Urdu Total		234	100	35	35	10	10
Total		331	100	70	70	20	20

Once the schools were sampled, the QCOs, supplemented by EGRA senior managers, verified the samples in the field. This step was necessary due to two factors: 1) some inaccuracies in the National Education Management Information System (NEMIS) data, and 2) changes in student numbers since the time period when the schools had submitted their data to NEMIS. If the original schools had fewer than 15 students in

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

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

In ICT, 20 schools from the original sample had to be replaced due to conflicting data gathered during the verification process. In the Urdu school sample, two schools were replaced, while 18 schools were replaced in the English school sample. The numbers of schools in Table 2 above was finalized through the field verification and the data collection. 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 ICT, were presented in a report to USAID.²

Trans-adaptation

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

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

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

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

Piloting

In March, the EGRA English tools were piloted in ICT while the Urdu tools were piloted in selected schools in AJK, ICT, and KP provinces. 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

Revision and Finalization

The EGRA team held a revision workshop in March with a limited number of experts from the trans-adaptation workshop. Changes were made to the instruments based on the pilot data and field observations. These changes were summarized in the piloting report. The EGRA team then finalized the four instruments for each language and submitted them to USAID in April. USAID made suggestions, particularly around the

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

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

inclusion of reading- and library-related items into 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 the Round 1 data collection in May. The final instruments consisted of the following:

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

These instruments are available for use by education officials.

Data Collection

Subcontractor Selection

The EGRA team, with the participation of USAID, issued a request for proposals and followed a set of criteria to select local sub-contractors for the field data collection and for data entry. In April, the Institute for Social and Applied Policy Studies (I-SAPS) was chosen for both activities (data collection and data entry). MSI, STS, and I-SAPS collaborated on the data collection in ICT.

Data Collection

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

Over a 10-day period in May, the enumerators spent a day in each of the 140 schools to collect the baseline data in ICT. The enumerators received frequent visits and mobile phone calls from the EGRA senior manager, QCOs, coordinator, and field supervisors to check on the status of data collection and to troubleshoot any issues. After collecting the data from the schools, the enumerators submitted their booklets to the supervisors and QCOs for verification and feedback. The supervisors then brought the booklets back to Islamabad for data entry.

Data Entry

Data Entry

In May, the EGRA team developed a customized data entry application so that 1) the exact data from the booklets and questionnaires could be entered into a database, and 2) the computers used for data entry could be networked with a server. In June, the team trained the I-SAPS data coordinator, two supervisors, and 30 data entry operators on the application, with additional hands-on training using actual data (from AJK, GB, and ICT). In June and July, the EGRA and I-SAPS teams did the data entry for over 10,000 student booklets, along with the questionnaires for the students, teachers, and head teachers. This included over 4,000 booklets and questionnaires for ICT.

Data Cleaning

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

Data Analysis

Methodology

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

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 sector or school level are too small for analysis purposes, and any results at those levels would be misleading.

Validity and Reliability

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

For reliability, a generally accepted method is to estimate the internal consistency reliability (Coefficient Alpha) of the test. The minimum reliability threshold is approximately 0.75 to 0.80 for tests of this nature. Reliability was estimated for each province and language. Table 3 shows the reliability estimates for grades 3 and 5 in ICT in English and Urdu.

TABLE 3: RELIABILITY ESTIMATES

Language	Grade Level	Tasks	N-count	Alpha
English	Grade 3	9	1,033	0.79
	Grade 5	9	1,047	0.77
Urdu	Grade 3	9	988	0.83
	Grade 5	9	1,037	0.82

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 ICT students' mastery of the tasks, and third, ICT students' fluency was assessed.

Item P-values and Item-Total Correlations

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

Percent Correct

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

Timed Tasks Scores

The scores on the timed tasks were calculated by taking the number of correct responses times 60 seconds then dividing that number by the number of seconds used to read the stimulus. For instance, if a student read 75 letters correctly in 30 seconds, their letters-correct-per-minute score would be 150 (75 words x 60 seconds/30 seconds). Given another example, if a student read 50 words correctly in 30 seconds, his or her timed task score would be 100 words per minute (50 words x 60 seconds/30 seconds). Table 4 lists the number of stimuli per task. Recall the percent correct scores ranged from zero to 100. The method for calculating phonics and fluency scores yielded much higher maximum values, upwards of 200 at baseline (see the task box plots in Annex 2, Figures 12, 13, 14 and 15).

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 ICT. There are sections on the sample, task and item statistics, score calculation, task and summary scores, timed task scores, and questionnaire findings.

Student Sample

Table 7 shows the number of students in the sample by gender. For grades 3 and 5 English, the actual samples were 98.4 and 99.7 percent of the intended sample, respectively. For grades 3 and 5 Urdu, the actual samples were 94.1 and 98.8 percent, respectively. A small number of students in grade 3 ($n = 6$) and grade 5 ($n = 1$) did not complete the gender item on the questionnaire. The total actual sample in ICT was 97.7 percent of the intended sample.

TABLE 7: ACTUAL STUDENT SAMPLE BY GRADE AND GENDER

Language	Grade Level	Sample	Boys	Girls	Missing	Total
English	Grade 3	Students	503	526	4	1,033
		% of Target	95.8%	100.2%	--	98.4%
	Grade 5	Students	544	503	0	1,047
		% of Target	103.6%	95.8%	--	99.7%
	Total	Students	1,047	1,029	4	2,080
		% of Target	99.7%	98.0%	--	99.0%
Urdu	Grade 3	Students	478	508	2	988
		% of Target	91.0%	96.8%	--	94.1%
	Grade 5	Students	525	511	1	1,037
		% of Target	100.0%	97.3%	--	98.8%
	Total	Students	1,003	1,019	3	2,025
		% of Target	95.5%	97.0%	--	96.4%
English and Urdu	Total	Students	2,050	2,048	7	4,105
		% of Target	97.6%	97.5%	--	97.7%

Task and Item Statistics

Tables 8 and 9 show the statistics for the tasks on the test. On the English test in ICT, the task p-values for grade 3 ranged from 0.02 to 0.75, thus providing a spread on the lower three-quarters of the difficulty spectrum. The p-values for grade 5 ranged from 0.04 to 0.92, which is across almost the entire difficulty spectrum. Such variations in p-values are helpful in terms of measuring pre- to post-test gains in student performance. The variation is also a factor in providing high quality tasks; only one of each of the grade 3 and grade 5 tasks had item-total correlations of less than 0.20, and all values for both grade levels were positive.

On the Urdu test in ICT, the task p-values for grade 3 ranged from 0.08 to 0.50, thus providing a spread on the lower half of the difficulty spectrum. The p-values for grade 5 ranged from 0.32 to 0.77, which is across

the middle part of the spectrum. Only one of the grade 3 tasks (at 0.18), and none of the grade 5 tasks, had item-total correlations of less than 0.20.

TABLE 8: ENGLISH TASK STATISTICS

Task (Subtest)	Grade 3		Grade 5	
	P-Value	Item-Total	P-Value	Item-Total
1. Orientation to print	0.63	0.13	0.71	0.15
2. Letter name recognition	0.75	0.46	0.92	0.31
3. Phonemic awareness	0.42	0.36	0.64	0.33
4. Letter sound knowledge	0.02	0.32	0.04	0.28
5. Familiar word reading	0.43	0.78	0.84	0.67
6. Non-word reading	0.27	0.76	0.59	0.66
7a. Passage reading	0.36	0.80	0.74	0.73
7b. Passage comprehension	0.06	0.53	0.21	0.60
8. Listening comprehension	0.12	0.31	0.23	0.41

TABLE 9: URDU TASK STATISTICS

Task (Subtest)	Grade 3		Grade 5	
	P-Value	Item-Total	P-Value	Item-Total
1. Orientation to print	0.50	0.18	0.54	0.22
2. Letter name recognition	0.37	0.57	0.56	0.45
3. Phonemic awareness	0.36	0.22	0.52	0.35
4. Letter sound knowledge	0.19	0.60	0.32	0.39
5. Familiar word reading	0.25	0.83	0.72	0.79
6. Non-word reading	0.14	0.78	0.45	0.73
7a. Passage reading	0.27	0.82	0.77	0.77
7b. Passage comprehension	0.08	0.70	0.40	0.69
8. Listening comprehension	0.27	0.35	0.49	0.31

Task and Summary Scores

After calculating the test reliability and the item (or task) statistics, the next part of the analysis involves plotting the scores. Histograms of the summary scores (Figures 2 to 5) show that the English and Urdu distributions are moving to the right from grade 3 to grade 5, which is strong evidence that the students are learning basic skills at the primary school level. The Urdu distributions, in fact, are moving slightly more, perhaps due to starting off at a lower point. As with the task and item statistics, it also shows that there is room for growth at each grade level, particularly at grade 3. The goal of the intervention is to see movement of the distributions to the right within the same grade level (i.e., grades 3 and 5) from the baseline to midline to endline.

FIGURE 2: ENGLISH GRADE 3 SUMMARY SCORES

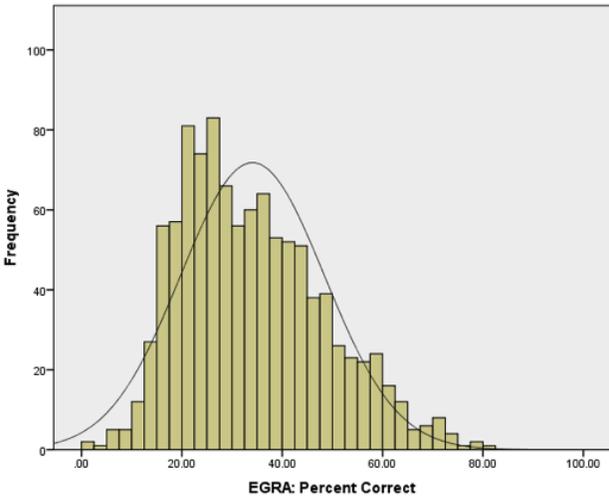


FIGURE 4: ENGLISH GRADE 5 SUMMARY SCORES

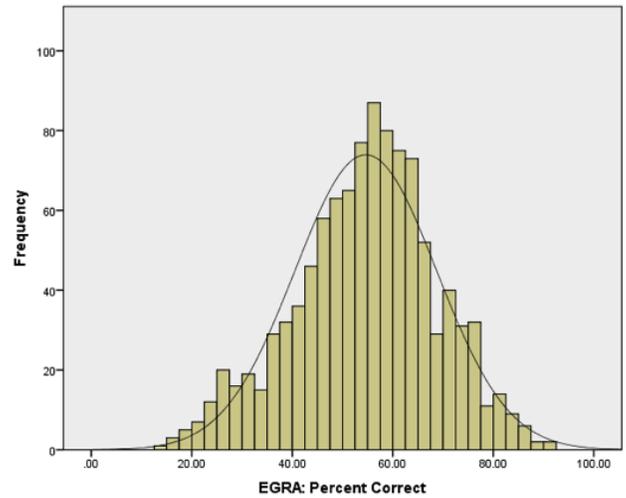


FIGURE 3: URDU GRADE 3 SUMMARY SCORES

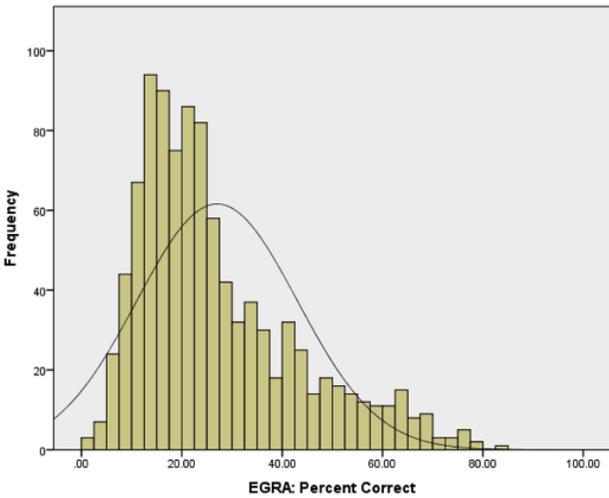


FIGURE 5: URDU GRADE 5 SUMMARY SCORES

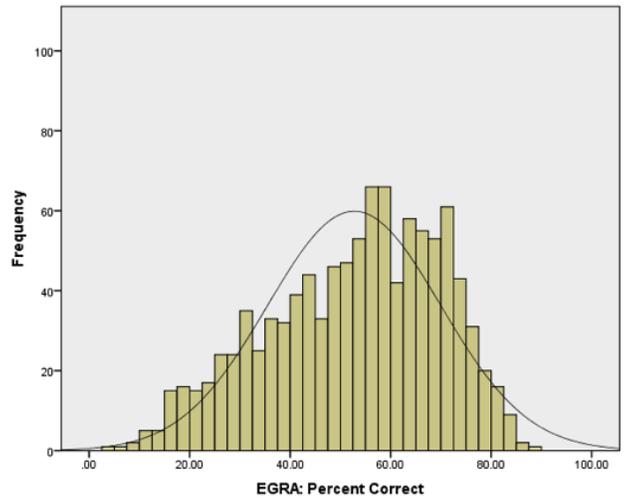


Table 10 and Figure 6 provide the average scores by task using percent correct scores for English and Table 11 and Figure 7 for Urdu. The score for each task was calculated using the total number correct and dividing by the number of items. For instance, a student who scored 3 out of 5 on Task 1 would receive a score of 60 percent. Averages were then calculated for all students on Task 1, which in ICT for English was 63.4 percent for grade 3 and 70.7 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., 70.7 percent minus 63.4 percent equals 7.3 percentage points.

For English, students at Grade 3 demonstrated relatively strong skills in orientation to print and letter name recognition. They had lower skills in areas such as letter sound knowledge and comprehension. Grade 5

students showed strong increases in the reading areas – familiar words, non-words, and passages. They were still doing poorly in letter sound knowledge and comprehension, though they showed some progression in comprehension; these are areas where there is much room for improvement. In areas where there are large differences – i.e., reading and phonemic awareness, and even in letter name recognition and passage comprehension – interventions at grade 3 could have particularly large effects in accelerating children’s learning.

TABLE 10: ENGLISH SCORES BY GRADE AND TASK

Task (Subtest)	Grade 3	Grade 5	Difference
1. Orientation to print	63.4%	70.7%	7.3% points
2. Letter name recognition	75.3%	92.3%	16.8% points
3. Phonemic awareness	42.3%	63.7%	21.4% points
4. Letter sound knowledge	2.2%	3.8%	1.6% points
5. Familiar word reading	43.1%	84.0%	40.9% points
6. Non-word reading	26.6%	58.6%	32.0% points
7a. Passage reading	36.0%	74.3%	38.3% points
7b. Passage comprehension	5.7%	20.8%	15.1% points
8. Listening comprehension	12.1%	23.5%	11.4% points
Reading Summary Score	34.1%	54.6%	20.5% points

FIGURE 6: ENGLISH SCORES BY GRADE AND TASK

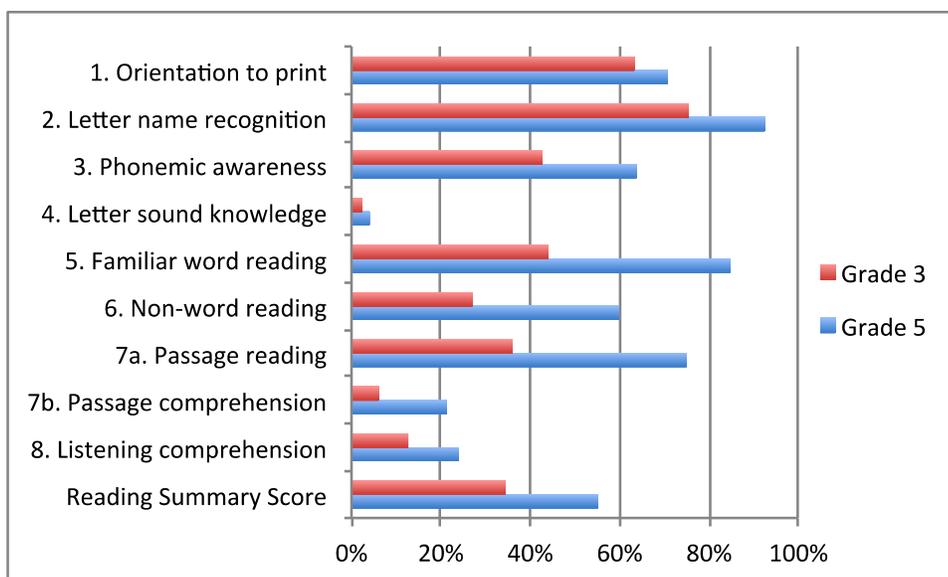


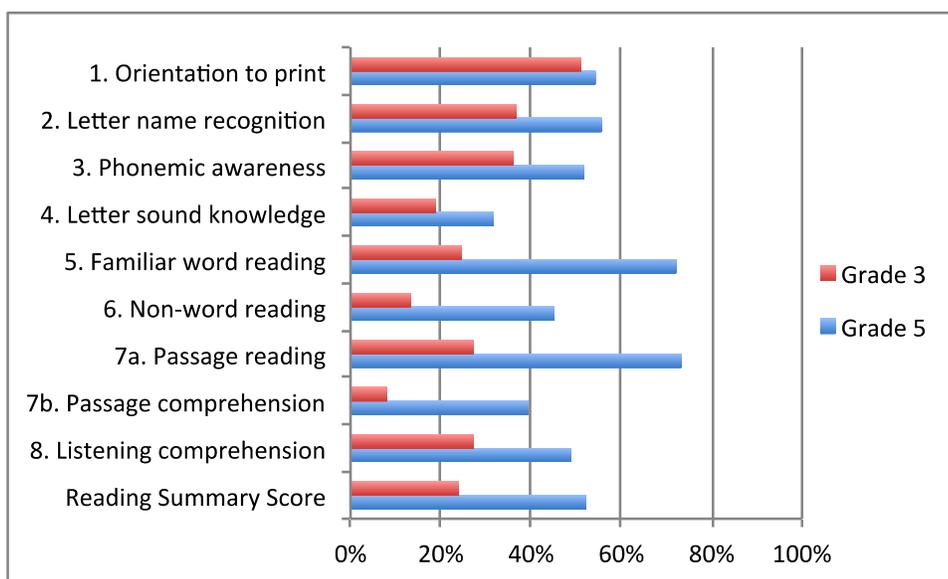
Table 11 and Figure 7 provide the Urdu scores by task, and for the summary (or grand mean). Grade 3 students demonstrated relatively strong skills in orientation to print. However, their scores in areas such as letter sound knowledge, non-word reading, and passage comprehension were low. Grade 5 students showed strong increases in the reading areas – familiar words, non-words, and passages – and in passage comprehension. They were still doing relatively poorly in letter sound knowledge, non-word reading, and

comprehension. In areas where the scores were low and where there were large gains, interventions at grade 3 could lead to substantial improvements in the overall score.

TABLE 11: URDU SCORES BY GRADE AND TASK

Task (Subtest)	Grade 3	Grade 5	Difference
1. Orientation to print	50.4%	54.2%	3.8% points
2. Letter name recognition	36.6%	55.5%	18.9% points
3. Phonemic awareness	36.1%	51.8%	15.7% points
4. Letter sound knowledge	18.9%	31.8%	12.9% points
5. Familiar word reading	24.9%	72.4%	47.5% points
6. Non-word reading	13.6%	45.2%	31.6% points
7a. Passage reading	27.4%	75.1%	47.7% points
7b. Passage comprehension	8.2%	39.7%	32.5% points
8. Listening comprehension	27.1%	48.9%	21.8% points
Reading Summary Score	27.0%	52.7%	25.7% points

FIGURE 7: URDU SCORES BY GRADE AND TASK



When the scores were disaggregated by gender (Tables 12 and 13 and Figures 8 to 11), most of the differences between boys and girls were small, though some were statistically significant in favor of girls. For English, at grades 3 and 5, girls had higher overall scores than boys; the difference was large in the three reading tasks: familiar word reading, non-word reading, and passage reading. The gains from grade 3 to grade 5 were about the same for the boys and girls, with boys increasing by about 21 points as opposed to 20 points for the girls.

For Urdu, the gender differences were similar, with girls having higher overall scores than boys. At grade 3, the girls scored more than 10 points higher than the boys in familiar word reading and passage reading. At grade 5, there was at least a 10-point difference in familiar word reading, non-word reading, passage reading, and passage comprehension. From grades 3 to 5, girls had a 28-point difference and boys a 23-point difference.

TABLE 12: ENGLISH SCORES BY GRADE AND GENDER

Task (Subtest)	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
1. Orientation to print	64.1%	62.3%	71.7%	69.7%
2. Letter name recognition	71.6%	78.5%	91.1%	93.8%
3. Phonemic awareness	38.0%	46.7%	59.4%	67.9%
4. Letter sound knowledge	2.0%	2.5%	2.6%	5.2%
5. Familiar word reading	35.7%	51.5%	79.8%	89.5%
6. Non-word reading	21.4%	32.2%	54.7%	64.5%
7a. Passage reading	28.7%	42.2%	70.3%	79.5%
7b. Passage comprehension	4.7%	7.1%	18.4%	24.0%
8. Listening comprehension	11.7%	13.5%	21.2%	26.6%
Reading Summary Score	30.6%	37.4%*	51.8%	57.6%*

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

FIGURE 8: ENGLISH GRADE 3 SCORES BY TASK AND GENDER

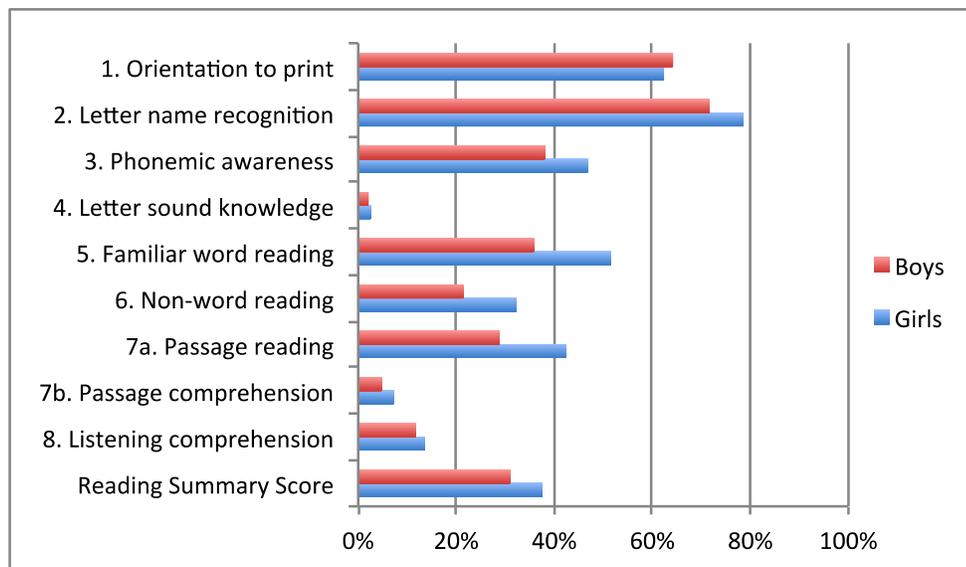


FIGURE 9: ENGLISH GRADE 5 SCORES BY TASK AND GENDER

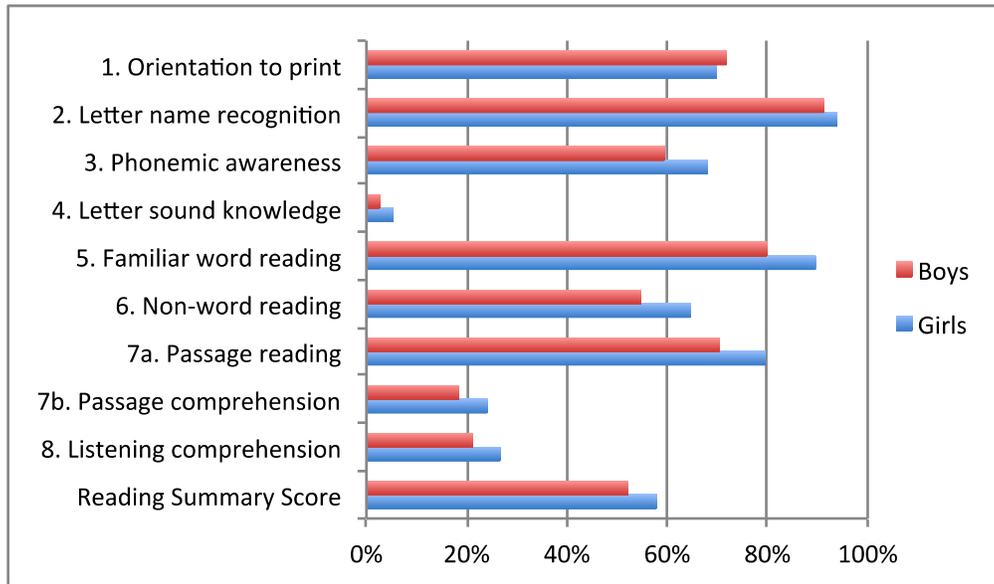


TABLE 13: URDU SCORES BY GRADE AND GENDER

Task (Subtest)	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
1. Orientation to print	53.0%	49.1%	52.5%	55.7%
2. Letter name recognition	32.3%	40.5%	51.0%	60.3%
3. Phonemic awareness	35.2%	36.7%	49.2%	54.2%
4. Letter sound knowledge	15.8%	21.7%	26.8%	36.4%
5. Familiar word reading	17.0%	31.6%	62.8%	81.8%
6. Non-word reading	10.3%	16.2%	37.9%	52.3%
7a. Passage reading	19.6%	34.1%	64.2%	82.9%
7b. Passage comprehension	4.6%	11.2%	28.5%	50.6%
8. Listening comprehension	24.9%	29.4%	47.1%	50.5%
Reading Summary Score	23.6%	30.1%*	46.7%	58.3%*

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

FIGURE 10: URDU GRADE 3 SCORES BY TASK AND GENDER

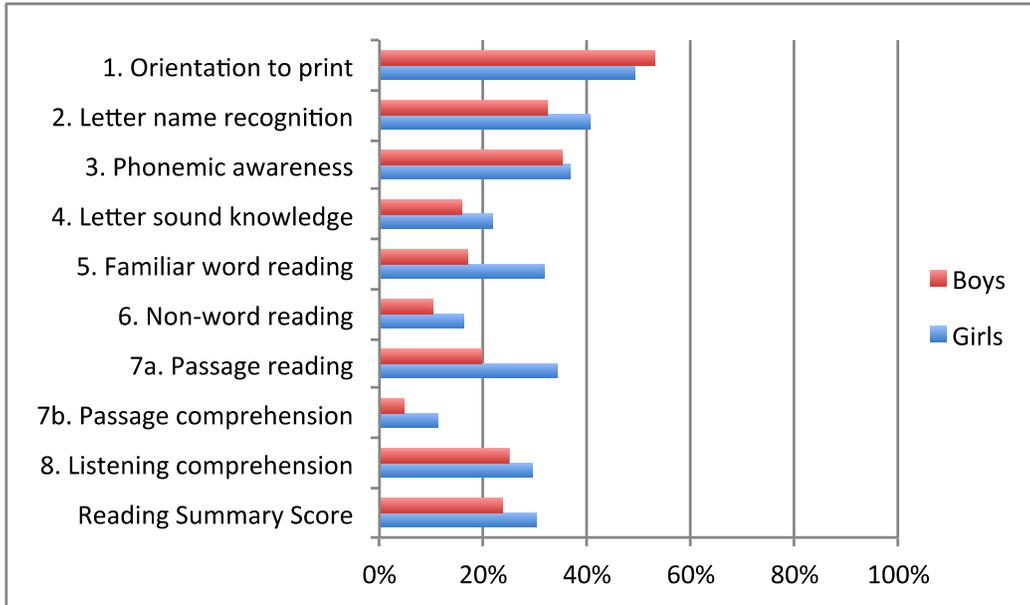
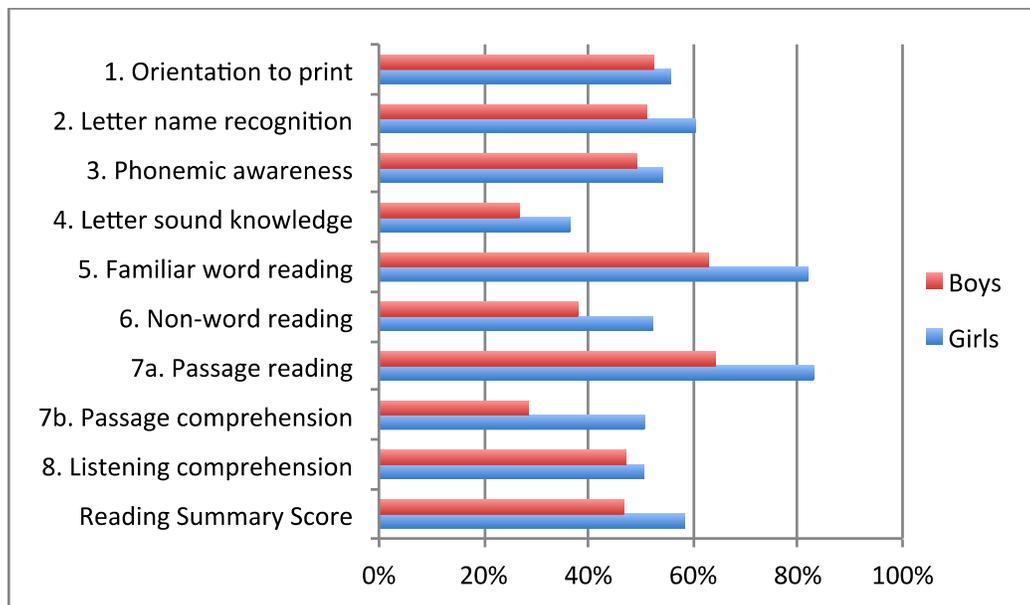


FIGURE 11: URDU GRADE 5 SCORES BY TASK AND GENDER



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 14 to 17 below show scores in terms of raw scores (instead of the percent correct scores on the previous tables). Tables 14 (English) and 15 (Urdu) have the maximum raw scores attained by students on each task at each grade level. Tables 14 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 12, 13, 14, and 15 in Annex 2). Tables 14 and 15 provide the baseline maximum scores at grade 3 and 5 for the five timed tasks.

TABLE 14: ENGLISH BASELINE MAXIMUM SCORES ON FLUENCY (TIMED) TASKS

Phonics Fluency Subtest	Grade 3	Grade 5
2. Letter name recognition	238	198
4. Letter sound knowledge	39	111
6. Non-word reading	93	115
Reading-Rate Fluency Subtest	Grade 3	Grade 5
5. Familiar word reading	143	167
7a. Passage reading	136	193

TABLE 15: URDU BASELINE MAXIMUM SCORES ON FLUENCY (TIMED) TASKS

Phonics Fluency Subtest	Grade 3	Grade 5
2. Letter name recognition	98	150
4. Letter sound knowledge	198	336
6. Non-word reading	66	94
Reading-Rate Fluency Subtest	Grade 3	Grade 5
5. Familiar word reading	100	138
7a. Passage reading	142	217

For English, the lowest scores on the timed tasks were in letter sound knowledge, which also showed the least progression from grade 3 to grade 5 (Table 16). All of the other areas – letter name recognition, familiar word reading, non-word reading, and passage reading – all showed large gains from grade 3 to grade 5. As seen in Table 17, there were differences in favor of girls on all tasks at each grade level, with the largest differences at grade 5 in familiar word reading and passage reading.

For Urdu, the lowest scores on the timed tasks were in non-word reading (Table 18). The highest scores were in letter name recognition at grades 3 and 5, and also in familiar word reading and passage reading at grade 5.

Those latter two areas also showed the largest progression from grade 3 to grade 5. As seen in Table 19, there were differences in favor of girls on all tasks at each grade level, with the largest differences at grade 5 in familiar word reading and passage reading.

TABLE 16: ENGLISH PHONICS AND READING-RATE FLUENCY TASK MEANS BY GRADE

Phonics Fluency Subtest	Grade 3	Grade 5	Difference (G5 – G3)
2. Letter name recognition	77.3	102.6	25.3 points
4. Letter sound knowledge	2.2	3.8	1.6 points
6. Non-word reading	14.0	33.9	19.9 points
Reading-Rate Fluency Subtest	Grade 3	Grade 5	Difference (G5 – G3)
5. Familiar word reading	25.6	63.7	38.1 points
7a. Passage reading	23.5	60.8	37.3 points

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

Phonics Fluency Subtest	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
2. Letter name recognition	73.5	80.7	101.0	104.6
4. Letter sound knowledge	2.0	2.5	2.7	5.2
6. Non-word reading	11.4	16.8	29.8	36.5
Reading-Rate Fluency Subtest	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
5. Familiar word reading	21.0	28.0	59.5	69.0
7a. Passage reading	18.0	26.4	53.8	66.3

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

TABLE 18: URDU PHONICS AND READING-RATE FLUENCY TASK MEANS BY GRADE

Phonics Fluency Subtest	Grade 3	Grade 5	Difference (G5 – G3)
2. Letter name recognition	36.6	55.7	19.1 points
4. Letter sound knowledge	19.0	32.2	13.2 points
6. Non-word reading	7.0	25.2	18.2 points
Reading-Rate Fluency Subtest	Grade 3	Grade 5	Difference (G5 – G3)
5. Familiar word reading	14.1	51.9	37.8 points
7a. Passage reading	19.3	71.0	51.7 points

TABLE 19: URDU PHONICS AND READING-RATE FLUENCY TASK MEANS BY GRADE AND GENDER

Phonics Fluency Subtest	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
2. Letter name recognition	32.3	39.8	49.9	59.0
4. Letter sound knowledge	14.4	21.9	25.6	32.8
6. Non-word reading	2.4	8.3	16.2	28.4
Reading-Rate Fluency Subtest	Grade 3		Grade 5	
	Boys	Girls	Boys	Girls
5. Familiar word reading	9.5	17.9	40.0	60.7
7a. Passage reading	13.4	24.3	51.6	85.9

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

Questionnaire Findings

Selected results are presented below, including results for those characteristics or items that showed significant differences in student scores. 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 overall EGRA averages for the grade 3 and 5 summary scores were used as total averages in the tables below.

Statistical significance was determined based on *t*-tests for indicators with two categories, and analyses of variance for indicators with three or more categories (with post-hoc pairwise comparisons). The significance value was set at $p < 0.05$; a 95 percent confidence level. For many of these analyses, the *n*-counts for the different categories of respondents (students, teachers, or head teachers) was either small, which often made it difficult to find statistically significant differences even though the practical differences may have been relatively large.

Student Questionnaires

Tables 20 and 21 have summary scores by student age and language. 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. The scores were usually the highest for the younger age students and lowest for the older age students. An exception was with Urdu at grade 3, where the normal age students did the best.

TABLE 20: ENGLISH SUMMARY SCORES BY STUDENT AGE

Age Group	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Younger than normal age	187	36.7%	239	60.3%
Normal age	653	35.0%	579	55.6%
Older than normal age	191	28.6%	224	45.6%
Missing	2	--	5	--
Total	1,033	34.3%*	1,047	55.0%*

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

TABLE 21: URDU SUMMARY SCORES BY STUDENT AGE

Age Group	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Younger than normal age	120	25.6%	53	55.8%
Normal age	575	27.5%	515	51.7%
Older than normal age	289	26.7%	465	46.9%
Missing	4	--	4	--
Total	988	27.0%	1,037	49.5%*

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

Tables 22 and 23 show the summary scores according to whether the student reads the Quran at home. There were significant differences in both languages and grades except for Urdu at grade 3 (likely because of the small non-Quran reading sample). Differences were consistently in favor of students who read the Quran at home.

TABLE 22: ENGLISH SUMMARY SCORES BY READING THE QURAN AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	104	25.1%	75	44.5%
Yes	928	35.1%*	966	49.7%*
Missing	1	--	4	--
Total	1,033	34.3%	1,047	49.5%

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

TABLE 23: URDU SUMMARY SCORES BY READING THE QURAN AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	43	23.1%	75	41.9%
Yes	944	27.2%	966	53.1%*
Missing	1	--	4	--
Total	988	27.0%	1,047	52.7%

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

Tables 24 and 25 show the differences in scores based on whether there is a library at the school. While the results were statistically significant in favor of the English-medium students who said that there is a library at their school, there was no difference in Urdu-medium schools results.

TABLE 24: ENGLISH 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	175	30.6%	160	51.2%
Yes	777	35.1%*	860	55.3%*
Missing	81	--	27	--
Total	1,033	34.3%	1,047	54.7%

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

TABLE 25: URDU 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	336	26.6%	271	51.5%
Yes	495	27.7%	712	53.0%
Missing	157	--	54	--
Total	988	27.3%	1,037	52.6%

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

In Tables 26 to 31, the data showed that the existence of newspapers and magazines generally made a difference in reading scores in most cases. The effect of the presence of books at home on scores was mixed. There may be evidence that increasing the presence of reading materials in the home could contribute to raising children's reading levels.

TABLE 26: ENGLISH SUMMARY SCORES BY THE PRESENCE OF NEWSPAPERS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	565	32.7%	530	52.8%
Yes	468	35.7%*	517	56.5%*
Missing	0	--	0	--
Total	1,033	34.1%	1,047	54.6%

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

TABLE 27: URDU SUMMARY SCORES BY THE PRESENCE OF NEWSPAPERS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	669	25.9%	587	51.5%
Yes	318	29.3%*	448	54.3%*
Missing	1	--	2	--
Total	988	27.0%	1,037	52.7%

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

TABLE 28: ENGLISH SUMMARY SCORES BY THE PRESENCE OF MAGAZINES AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	905	33.3%	858	54.0%
Yes	128	39.7%*	189	57.5%*
Missing	0	--	0	--
Total	1,033	34.1%	1,047	54.6%

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

TABLE 29: URDU SUMMARY SCORES BY THE PRESENCE OF MAGAZINES AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	897	26.7%	902	52.5%
Yes	90	30.1%	133	54.6%
Missing	1	--	2	--
Total	988	27.0%	1,037	52.7%

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

TABLE 30: ENGLISH SUMMARY SCORES BY THE PRESENCE OF BOOKS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	635	33.1%	669	54.7%
Yes	398	35.6%*	378	54.5%
Missing	0	--	0	--
Total	1,033	34.1%	1,047	54.6%

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

TABLE 31: URDU SUMMARY SCORES BY THE PRESENCE OF BOOKS AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	583	27.1%	602	54.1%*
Yes	404	27.0%	433	50.9%
Missing	1	--	2	--
Total	988	27.0%	1,037	52.7%

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

This set of student questions (in Tables 33 to 37) pertained to children's reading habits at home. In general, these habits made a difference in student scores at grade 3 but not at grade 5. There seemed to be slightly more benefit for children attending Urdu-medium schools as opposed to those attending English-medium schools.

TABLE 32: ENGLISH 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	379	32.4%	401	54.6%
Yes	652	35.0%*	644	54.7%
Missing	2	--	2	--
Total	1,033	34.1%	1,047	54.6%

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

TABLE 33: URDU 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	420	25.4%	380	52.6%
Yes	558	28.5%*	649	52.8%
Missing	10	--	8	--
Total	988	27.0%	1,037	52.7%

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

TABLE 34: ENGLISH SUMMARY SCORES BY CHILDREN READING TO SOMEONE ELSE AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	431	33.1%	425	54.7%
Yes	600	34.7%	621	54.6%
Missing	2	--	1	--
Total	1,033	34.1%	1,047	54.6%

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

TABLE 35: URDU SUMMARY SCORES BY CHILDREN READING TO SOMEONE ELSE AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	480	25.6%	425	52.3%
Yes	502	28.6%*	621	53.0%
Missing	6	--	1	--
Total	988	27.0%	1,037	52.7%

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

TABLE 36: ENGLISH SUMMARY SCORES BY CHILDREN READING SILENTLY AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	160	32.0%	186	54.4%
Yes	870	34.4%	857	54.7%
Missing	3	--	4	--
Total	1,033	34.1%	1,047	54.6%

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

TABLE 37: URDU SUMMARY SCORES BY CHILDREN READING SILENTLY AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	214	25.3%	277	53.0%
Yes	766	27.6%	757	52.7%
Missing	8	--	3	--
Total	988	27.0%	1,037	52.7%

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

As seen in Tables 38 and 39, having a computer at home seemed to be associated with higher reading scores. The differences were greater in English than in Urdu. Since computers are likely related to socio-economic status, which also tends to lead to higher reading levels, a supplemental study is recommended to find out whether children use computers as reading devices.

TABLE 38: ENGLISH SUMMARY SCORES BY CHILDREN HAVING A COMPUTER AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	526	31.1%	421	50.1%
Yes	503	37.1%*	625	57.7%*
Missing	4	--	1	--
Total	1,033	34.1%	1,047	54.6%

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

TABLE 39: URDU SUMMARY SCORES BY CHILDREN HAVING A COMPUTER AT HOME

Response	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	719	26.4%	694	52.3%
Yes	265	28.9%*	330	53.6%
Missing	4	--	13	--
Total	988	27.0%	1,037	52.7%

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

Teacher Questionnaires

With the smaller sample size, the analysis of the teacher questionnaires was limited to providing descriptive statistics on teacher characteristics and summary scores, i.e., with no group comparisons. Tables 40 to 43 provide information on teacher qualifications. There was little variation in the student scores based on teacher qualifications, either academic or professional for English or Urdu.

TABLE 40: ENGLISH 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.	43	34.4%	41	56.6%
B.A./B.Sc.	15	35.4%	22	52.2%
F.A./F.Sc.	3	29.2%	1	53.1%
Matric	0	--	0	--
Missing	0	--	1	--
Total	61	34.4%	64	55.1%

TABLE 41: URDU 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.	17	25.3%	16	52.5%
B.A./B.Sc.	20	27.2%	23	53.1%
F.A./F.Sc.	12	28.8%	14	50.0%
Matric	11	26.9%	8	54.1%
Missing	0	--	0	--
Total	60	26.9%	61	52.4%

TABLE 42: ENGLISH SUMMARY SCORES BY TEACHER PROFESSIONAL QUALIFICATION

Professional Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.Ed./M.A.	25	35.1%	33	54.4%
B.Ed.	31	34.9%	26	56.6%
C.T.	4	27.3%	3	47.0%
P.T.C.	1	29.7%	2	58.2%
Missing	0	--	0	--
Total	61	34.4%	64	55.1%

TABLE 43: URDU SUMMARY SCORES BY TEACHER PROFESSIONAL QUALIFICATION

Professional Qualification	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
M.Ed./M.A.	3	32.2%	6	54.5%
B.Ed.	24	26.4%	26	51.8%
C.T.	11	26.9%	17	50.2%
P.T.C.	17	28.0%	11	54.2%
Missing	5	--	1	--
Total	60	27.3%	61	52.1%

In an analysis of student scores by teacher age and experience, there were no consistent patterns of younger or older teachers, or those with less or more experience, relating to higher or lower student scores (Tables 44 to 47). Again, small teacher sample sizes made drawing conclusions difficult.

TABLE 44: ENGLISH SUMMARY SCORES BY TEACHER AGE

Age Group in Years	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
40 and less	43	34.7%	43	54.7%
Between 41 and 50	14	34.3%	16	56.1%
51 and more	4	32.0%	5	54.5%
Missing	0	--	0	--
Total	61	34.4%	64	55.1%

TABLE 45: URDU SUMMARY SCORES BY TEACHER AGE

Age Group in Years	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
40 and less	26	25.0%	27	55.0%
Between 41 and 50	27	28.8%	15	49.1%
51 and more	7	26.5%	18	50.0%
Missing	0	--	0	--
Total	60	26.9%	61	52.4%

TABLE 46: ENGLISH SUMMARY SCORES BY TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
10 or less	41	34.0%	36	54.4%
Between 11 and 20	13	36.9%	16	54.0%
Between 21 and 30	6	33.1%	10	58.5%
31 or more	1	27.3%	2	58.2%
Missing	0	--	0	--
Total	61	34.4%	64	55.1%

TABLE 47: URDU SUMMARY SCORES BY TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
10 or less	34	26.7%	30	54.7%
Between 11 and 20	10	28.7%	5	57.1%
Between 21 and 30	15	27.0%	21	48.7%
31 or more	1	16.6%	5	45.8%
Missing	0	--	0	--
Total	60	26.9%	61	52.4%

For frequency of in-service training, there were also no clear patterns (Tables 48 and 49). Again, any differences should be interpreted with caution due to the small sample size.

TABLE 48: ENGLISH SUMMARY SCORES BY TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	27	36.4%	26	55.9%
One time	28	32.9%	27	54.2%
Two times	4	31.9%	10	54.9%
Three times	0	--	1	57.2%
Missing	2	--	0	--
Total	61	34.5%	64	55.1%

TABLE 49: URDU SUMMARY SCORES BY TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	41	26.2%	35	51.7%
One time	15	28.6%	18	51.0%
Two times	2	18.2%	3	55.7%
Three times	2	37.1%	5	57.2%
Missing	0	--	0	--
Total	60	26.9%	61	52.1%

Head Teacher Questionnaires

The sample size for the head teacher questionnaires was small, so data interpretations should be treated with caution. Tables 50 to 53 show head teacher qualifications. The results are inconsistent, with better qualifications associated with higher scores in English but lower scores in Urdu.

TABLE 50: ENGLISH 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.	44	35.2%	43	55.1%
B.A./B.Sc.	23	32.8%	23	53.6%
F.A./F.Sc.	2	27.4%	2	55.4%
Matric	0	--	0	--
Missing	0	--	0	--
Total	69	34.2%	69	54.6%

TABLE 51: URDU 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.	50	26.2%	50	52.2%
B.A./B.Sc.	20	29.1%	20	54.3%
F.A./F.Sc.	0	--	0	--
Matric	0	--	0	--
Missing	0	--	0	--
Total	70	27.1%	70	52.8%

TABLE 52: ENGLISH 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.	32	35.1%	32	55.5%
B.Ed.	35	33.7%	35	53.7%
C.T.	2	27.4%	2	55.4%
P.T.C.	0	--	0	--
Missing	0	--	0	--
Total	69	34.2%	69	54.6%

TABLE 53: URDU 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.	32	25.4%	32	52.0%
B.Ed.	37	28.8%	37	53.6%
C.T.	0	--	0	--
P.T.C.	0	--	0	--
Missing	1	--	1	--
Total	70	27.2%	70	52.9%

Tables 54 to 57 provide information on head teachers' experience and in-service training. In English and Urdu, the relationships between experience, training, and reading scores were inconsistent.

TABLE 54: ENGLISH SUMMARY SCORES BY HEAD TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
2 or less	28	34.1%	28	54.8%
3 to 5	12	30.1%	12	51.0%
6 to 10	16	35.4%	16	54.2%
11 or more	12	37.1%	12	58.1%
Missing	1	--	1	--
Total	69	34.2%	69	54.6%

TABLE 55: URDU SUMMARY SCORES BY HEAD TEACHER EXPERIENCE

Years of Experience	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
2 or less	25	27.0%	25	50.1%
3 to 5	13	28.5%	13	51.0%
6 to 10	23	25.9%	23	55.0%
11 or more	16	31.1%	16	58.8%
Missing	3	--	3	--
Total	70	27.2%	70	52.7%

TABLE 56: ENGLISH SUMMARY SCORES BY HEAD TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	30	33.9%	30	55.7%
1 time	21	36.1%	21	54.4%
2 times	11	31.6%	11	51.7%
More than 2 times	7	33.1%	7	54.7%
Missing	0	--	0	--
Total	69	34.2%	69	54.6%

TABLE 57: URDU SUMMARY SCORES BY HEAD TEACHER IN-SERVICE TRAINING

Frequency of Training	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
None	34	25.8%	34	52.3%
1 time	28	28.0%	28	53.3%
2 times	6	29.2%	6	52.7%
More than 2 times	2	28.3%	2	54.7%
Missing	0	--	0	--
Total	70	27.1%	70	52.8%

Tables 58 to 61 provide data on head teachers' support to teachers in reading and the training that head teachers received in teaching reading. The data were mostly inconsistent, with some possible effects of in-service training for head teachers on student reading scores in the early grades in Urdu.

TABLE 58: ENGLISH SUMMARY SCORES BY HEAD TEACHER SUPPORT TO TEACHERS IN READING

Support to Teachers	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	6	27.8%	6	52.5%
Yes	63	34.8%	63	54.8%
Missing	0	--	0	--
Total	69	34.2%	69	54.6%

TABLE 59: URDU SUMMARY SCORES BY HEAD TEACHER SUPPORT TO TEACHERS IN READING

Support to Teachers	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	1	23.9%	1	47.3%
Yes	69	27.1%	69	52.9%
Missing	0	--	0	--
Total	70	27.1%	70	52.8%

TABLE 60: ENGLISH 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	21	33.2%	21	54.0%
Yes	48	34.6%	48	54.8%
Missing	0	--	0	--
Total	69	34.2%	69	54.6%

TABLE 61: URDU 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	20	23.7%	20	52.9%
Yes	49	28.6%	49	53.0%
Missing	1	--	1	--
Total	70	27.1%	70	52.9%

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 62 to 71). As expected, female schools performed better than male or mixed schools. Urban schools performed better than rural schools, though comparisons were made for the English test by school location, i.e., urban vs. rural, but not for Urdu since only one school in the Urdu-medium sample was classified as urban by NEMIS. Better infrastructure seemed to have a positive relationship with student reading scores.

TABLE 62: ENGLISH SUMMARY SCORES BY SCHOOL GENDER

School Gender	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Male school	10	30.6%	10	50.6%
Female school	9	41.4%	9	61.6%
Mixed school	50	33.6%	50	54.1%
Missing	0	--	0	--
Total	69	34.2%	69	54.6%

TABLE 63: URDU SUMMARY SCORES BY SCHOOL GENDER

School Gender	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Male school	33	23.9%	33	46.9%
Female school	26	31.8%	26	59.8%
Mixed school	11	25.4%	11	54.1%
Missing	0	--	0	--
Total	70	27.1%	70	52.8%

TABLE 64: ENGLISH SUMMARY SCORES BY SCHOOL LOCATION

School Location	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Urban	53	35.9%	53	56.7%
Rural	16	28.2%	16	47.5%
Missing	0	--	0	--
Total	69	34.2%	69	54.6%

TABLE 65: URDU SUMMARY SCORES BY SCHOOL LOCATION

School Location	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
Urban	1	20.4%	1	43.2%
Rural	69	27.2%	69	53.0%
Missing	0	--	0	--
Total	70	27.1%	70	52.9%

TABLE 66: ENGLISH SUMMARY SCORES BY PTA/SMC/PTSMC/PTC

Parent Teacher Committee	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	0	--	0	--
Yes	69	34.2%	69	54.6%
Missing	0	--	0	--
Total	69	34.2%	69	54.6%

TABLE 67: URDU SUMMARY SCORES BY PTA/SMC/PTSMC/PTC

Parent Teacher Committee	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	0	--	0	--
Yes	70	27.1%	70	52.8%
Missing	0	--	0	--
Total	70	27.1%	70	52.8%

TABLE 68: ENGLISH SUMMARY SCORES BY PRESENCE OF A SCHOOL LIBRARY

School Library	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	7	32.2%	7	53.7%
Yes	62	34.4%	62	54.7%
Missing	0	--	0	--
Total	69	34.2%	69	54.6%

TABLE 69: URDU SUMMARY SCORES BY PRESENCE OF A SCHOOL LIBRARY

School Library	Grade 3		Grade 5	
	n-count	Sum. Score	n-count	Sum. Score
No	21	25.4%	21	54.1%
Yes	49	28.0%	49	52.3%
Missing	0	--	0	--
Total	70	27.1%	70	52.8%

TABLE 70: ENGLISH 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
1	0	--	0	--
2	5	25.1%	5	45.7%
3	64	34.9%	64	55.3%
Missing	0	--	0	--
Total	69	34.2%	69	54.6%

TABLE 71: URDU 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
1	1	20.0%	1	45.8%
2	14	24.6%	14	50.8%
3	55	27.8%	55	53.5%
Missing	0	--	0	--
Total	70	27.1%	70	52.8%

CHAPTER 4: CONCLUSIONS AND RECOMMENDATIONS

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

Design and Methodology

1. The design followed USAID evaluation guidelines for a cross-sectional approach. However, due to selecting all of the sectors (and all of the schools) in ICT for full treatment, there is no counterfactual (in either English or Urdu) against which to measure the effects of the full treatment above and beyond the light treatment. With the cross-sectional design, the evaluation will be limited to examining the progress of students in grades 3 and 5 over the course of the PRP project in the two languages.
2. The sampling issues were addressed as well as could have been expected. In a limited number of schools, there was an issue of a lack of the requisite number of students per grade level. However, the actual sample of schools was 100 percent and the actual sample of students reached over 97 percent of the intended sample.
3. The EGRA test was of good quality. The reliability estimates were in the range of previous EGRA administrations in other countries. The task statistics were acceptable, with an appropriate range of p-values and item-total correlations that were at an acceptable level of quality. The characteristics of the test were such that it should be a strong measure of progress over time due to project-led interventions. As with any test, there may be ways to improve on the task and item statistics for the midline and endline.
4. The field implementation was successful. The logistical challenges in ICT were relatively minor. There was a high level of standardization reported by the quality control officers, which they attributed to the effective training process by the EGRA team. The team paid careful attention to detail in the logistics and test administration, which was reflected in the low error rates in the booklets and in the data entry.
5. The timeline was followed for the Round 1 data collection, though the data entry and cleaning were slightly behind schedule due to the need to redesign the data entry software so that it could be networked to a server. The development of this software and system proved to be valuable in entering the data for Round 1, and made for more efficiency in Rounds 2 and 3.

Findings and Results

The ICT evaluation involves two kinds of analyses: 1) a comparison of each group to itself at the baseline, midline, and endline (no control groups), and 2) separate comparisons for each language group (English and Urdu).

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

1. EGRA was administered to a robust sample at each grade level (3 and 5). Test reliabilities were very good, showing that the EGRA tasks and items worked well in measuring reading constructs at both grade levels. The task and item statistics showed that EGRA discriminates well between low- and high-achieving students in both grades. They also showed that there is adequate room for growth by students in each grade level.

2. In English, students were strongest in letter name recognition and orientation to print at grade 3. In grade 5, they had high scores in letter name recognition, familiar word reading, and passage reading. In both grades, they had the most difficulties with letter sound knowledge, followed by passage and listening comprehension. They made substantial progress from grade 3 to grade 5 in several areas, particularly familiar word reading, non-word reading, and passage reading.
3. In Urdu, students were the strongest in orientation to print, letter name recognition, and phonemic awareness at grade 3. At grade 5, they did relatively well in the same areas, plus familiar word reading, and passage reading. The students were weakest in letter sound knowledge, non-word reading, and passage comprehension. They made substantial progress from grade 3 to grade 5 in familiar word reading, non-word reading, and passage reading and comprehension.
4. Female students had higher scores, in general, than did their male counterparts. Areas such as familiar word reading, non-word reading, passage reading, and passage comprehension were areas of particular strength for the females over the males in ICT. The differences were about the same in the two grade levels.
5. 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 and in both languages 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 ICT.
6. Questionnaire findings were mostly inconclusive, due to small sample sizes and the lack of differences in responses within the student, teacher, and head teacher samples. For the students, attending a grade at an appropriate age, or even younger, seemed to have a positive effect on reading outcomes. In terms of the home environment, the presence of reading materials seemed to have a small positive effect on children's reading levels. It was the same with having a person to read with, with more of a positive effect for the younger students. Having a computer at home was associated with better reading scores.
7. Teacher and head teacher qualifications and experience, along with in-service training, were generally not related to student scores. For the English-medium schools, those in urban areas did better than those in rural areas; there were not enough urban Urdu-medium schools to analyze.

Evaluation Recommendations

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

1. The instrument development and trans-adaptation process was comprehensive and resulted in high quality EGRA tools. This should be repeated as soon as possible with the tasks that need to be changed for the midline and endline tools (to minimize test-retest effects and security breaches), so that reading progress can be accurately measured over time.

2. The EGRA items and tasks had good reliability values and covered the low-to-middle difficulty range. At baseline, the reading scores were relatively low for both grades, and show room for growth. In addition, histograms and box plots provided evidence that the tool is expected to measure higher levels of reading that are anticipated due to project-led interventions. Therefore, the baseline data indicates that the EGRA is appropriate for measuring increases in reading ability at midline and endline.
3. The sampling was reasonable in terms of finding a balance between the resources available, the required sample size, and the geographic coverage. It should be maintained in the midline and endline, i.e., keep the same sectors and schools, along with the methods at the school level.
4. The systems developed for field data collection should be repeated. The different layers of management, coordination, supervision, and quality control contributed to successful planning, implementation, and problem solving. The quality control officers were particularly important in maintaining standards and providing support for the local subcontractors.
5. The data entry process took time to develop but it eventually proved to be advantageous in terms of having the data entry operators connect to a central server. This facilitated the two rounds of data entry and the reconciliation process. This system should also be repeated in subsequent data entry activities.
6. The methods for analysis also took some time to develop, but it was important to create templates and agree on a methodology due to the volume of analysis and reporting that needs to be done for eight provinces. Again, the investment of time and effort in this process paid dividends for Rounds 2 and 3 of the baseline and will do so for the midline and endline.
7. Reading proficiency levels should be created to provide educators and other stakeholders with meaningful results. Most parents and educators better understand reading achievement in useful terms or levels, such as emerging, proficient, or advanced, rather than interpreting a percent-correct test score that may differ by test or reading passage difficulty. Education officials are encouraged to select specific EGRA scores to serve as levels of reading proficiency for both grades. Percent correct for each task, summary score, as well as fluency rates are recommended for this purpose. The baseline EGRA data can be used for establishing these reading proficiency levels.
8. Finally, it may be advisable to add items to the student, teacher, and head teacher questionnaires to collect data on PRP- and SRP-supported interventions so that student scores can be correlated with these indicators.

In general, the ICT 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, in both English and Urdu.

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

Tables A1 (English) and A2 (Urdu) present 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). The timed tasks are lists of letters, sounds, and words, 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. In English, the difficulty values ranged from 0.00 to 0.89 for grade 3 and 0.03 to 0.93 for grade 5, indicating an acceptable range. A total of 19 and 21 items for grades 3 and 5 out of the 23 items per grade had item-total correlations of at least 0.20, indicating high quality items.

TABLE AI: ENGLISH 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	0.89	0.27	0.87	0.30
	Q2	0.68	0.23	0.80	0.22
	Q3	0.58	0.03	0.39	0.06
	Q4	0.20	-0.03	0.55	0.05
	Q5	0.81	0.14	0.93	0.20
2. Letter name recognition (timed)	--				
3. Phonemic awareness (untimed)	Q1	0.62	0.47	0.82	0.54
	Q2	0.36	0.10	0.38	0.26
	Q3	0.43	0.49	0.71	0.54
	Q4	0.35	0.35	0.55	0.44
	Q5	0.35	0.41	0.60	0.44
	Q6	0.58	0.37	0.78	0.49
	Q7	0.37	0.45	0.56	0.54
	Q8	0.31	0.49	0.59	0.51
	Q9	0.50	0.45	0.75	0.50
	Q10	0.36	0.50	0.62	0.58
4. Letter sound knowledge (timed)	--				
5. Familiar word reading (timed)	--				
6. Non-word reading (timed)	--				
7a. Passage reading (timed)	--				
7b. Passage comprehension (untimed)	Q1	0.10	0.54	0.37	0.48
	Q2	0.14	0.45	0.37	0.45
	Q3	0.00	0.24	0.03	0.38
	Q4	0.04	0.49	0.21	0.55
	Q5	0.01	0.31	0.06	0.47
8. Listening comprehension (untimed)	Q1	0.23	0.31	0.36	0.42
	Q2	0.06	0.26	0.15	0.42
	Q3	0.07	0.32	0.20	0.47

In Urdu, the difficulty values ranged from 0.00 to 0.74 for grade 3 and 0.01 to 0.83 for grade 5, also indicating a strong range of items. A total of 20 and 18 items for grades 3 and 5 out of the 23 items per grade had item-total correlations of at least 0.20, also indicating high quality items.

TABLE A2: URDU ITEM STATISTICS BY GRADE

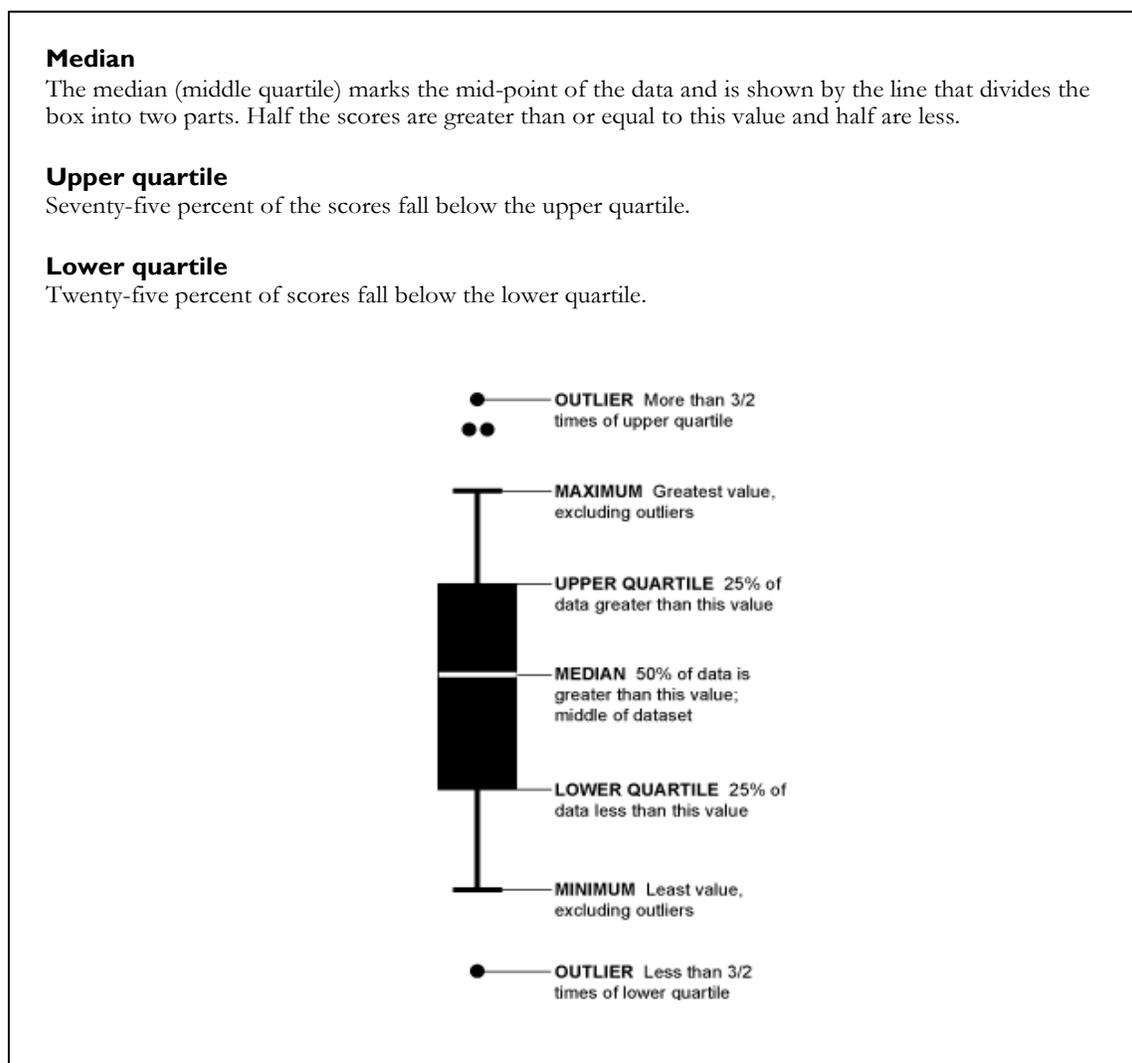
Task (Subtest)	Item	Grade 3		Grade 5	
		P-Value	Item-Total	P-Value	Item-Total
1. Orientation to print (untimed)	Q1	0.74	0.27	0.64	0.19
	Q2	0.58	0.29	0.63	0.27
	Q3	0.45	0.24	0.34	0.17
	Q4	0.12	0.02	0.30	0.07
	Q5	0.63	0.22	0.80	0.23
2. Letter name recognition (timed)	--				
3. Phonemic awareness (untimed)	Q1	0.59	0.36	0.78	0.42
	Q2	0.38	0.41	0.59	0.54
	Q3	0.33	0.31	0.45	0.38
	Q4	0.27	0.30	0.46	0.50
	Q5	0.34	0.32	0.48	0.45
	Q6	0.47	0.35	0.64	0.41
	Q7	0.22	0.28	0.36	0.41
	Q8	0.30	0.31	0.42	0.39
	Q9	0.20	0.28	0.35	0.42
	Q10	0.51	0.35	0.65	0.41
4. Letter sound knowledge (timed)	--				
5. Familiar word reading (timed)	--				
6. Non-word reading (timed)	--				
7a. Passage reading (timed)	--				
7b. Passage comprehension (untimed)	Q1	0.09	0.48	0.38	0.41
	Q2	0.08	0.48	0.28	0.30
	Q3	0.05	0.47	0.20	0.29
	Q4	0.09	0.49	0.54	0.35
	Q5	0.00	0.07	0.01	0.09
8. Listening comprehension (untimed)	Q1	0.28	0.23	0.50	0.23
	Q2	0.06	0.19	0.14	0.16
	Q3	0.47	0.21	0.83	0.22

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



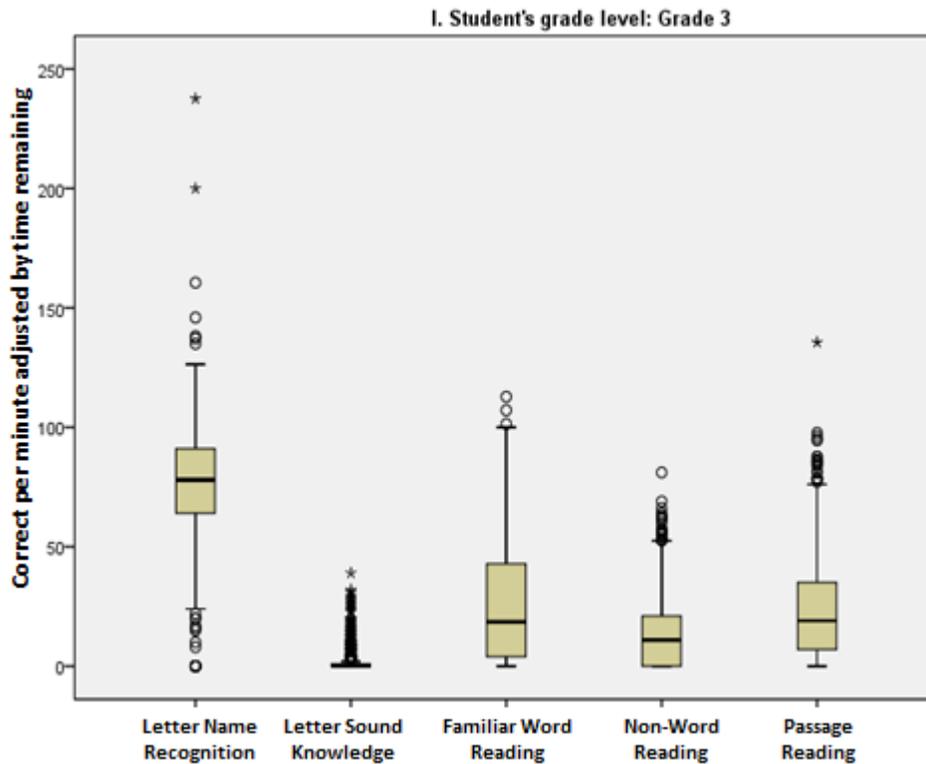
Box plots are presented below (Figures A2 to A5) for the results by language and grade level on the five fluency (timed) tasks: letter name recognition (task 2), letter sound knowledge (task 4), familiar word reading (task 5), non-word reading (task 6), and passage reading (task 7a).

Grade 3, English

For English 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 (letter-sound knowledge) to about 60 (letter name recognition) items per minute. It shows that the students had much better knowledge of letter names than phonics.

The variation (i.e., the range of scores, without outliers) for each of the tasks varied from about 0 (letter sound knowledge) to about 100 (familiar word reading). It shows that the scores were more spread out when reading words than providing the sound of letters distributed in random order.

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



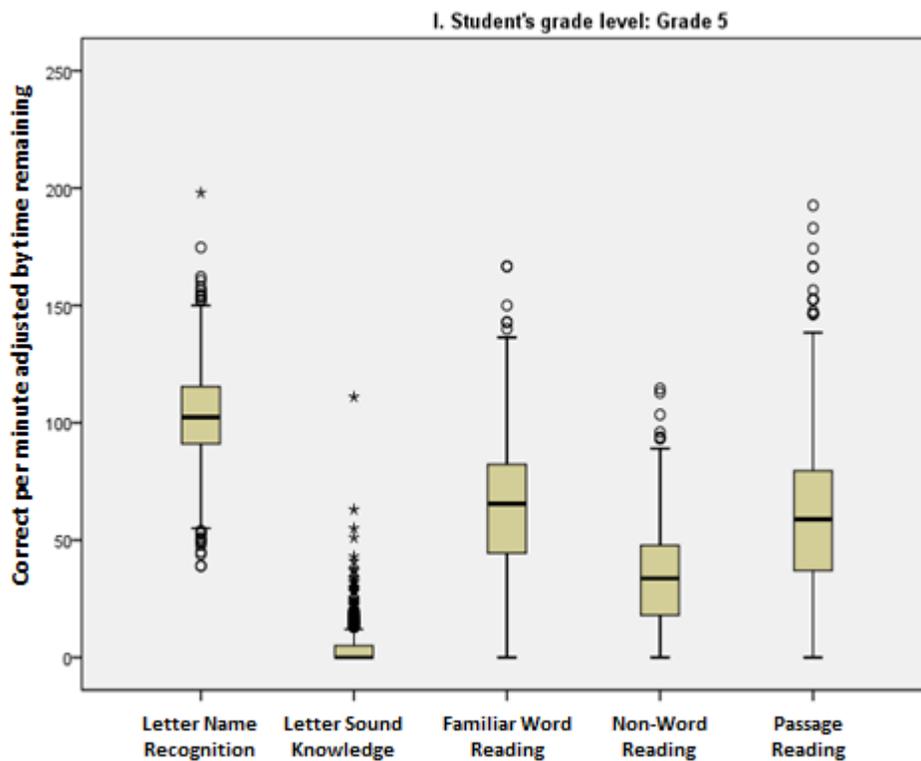
Grade 5, English

For English grade 5, the central tendency (the median speed) for each of the tasks ranged from about 0 (letter sound knowledge) to about 100 (letter name recognition) items per minute. . It shows that the students had much better knowledge of letter names than phonics.

The variation (i.e., the range of scores, without outliers) for each of the tasks varied from about 10 (letter sound knowledge) to about 140 (familiar words). It shows that the scores were more spread out when reading words than producing the sound of letters distributed in random order.

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

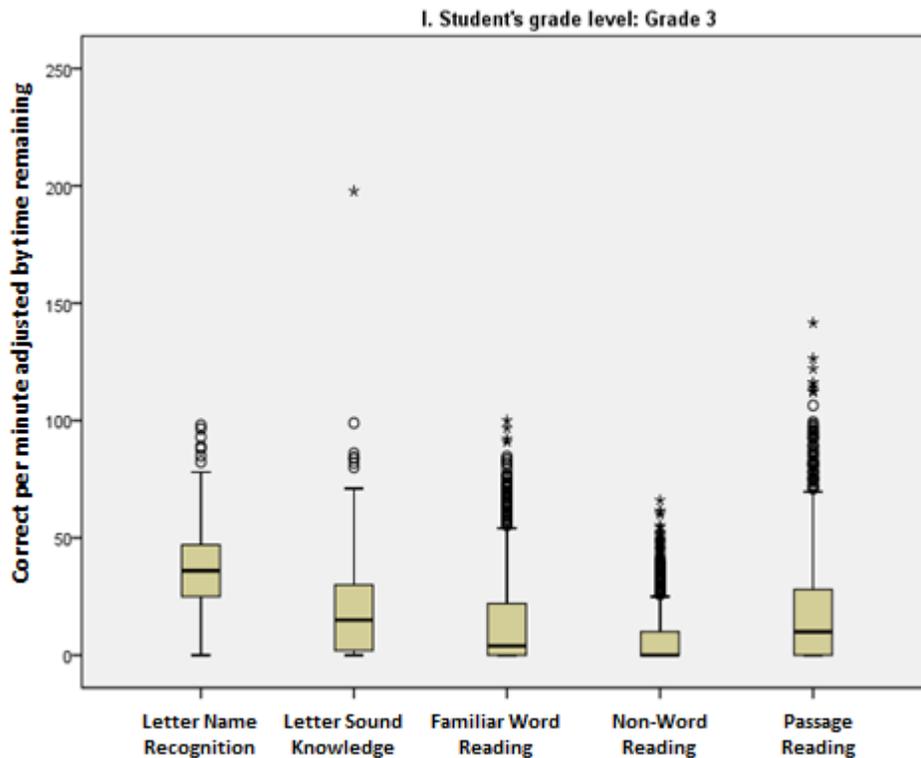


Grade 3, Urdu

For Urdu grade 3, the central tendency (i.e., the median speed, or the line in the middle) for each of the tasks ranged from about 0 (non-word reading) to about 40 (letter name recognition) items per minute. It shows that the students had much better knowledge of letter names than grapheme-morpheme correspondence.

The variation (i.e., the range of scores, without outliers) for each of the tasks varied from about 30 (non-word reading) to about 70 (letter name recognition). It shows that the scores were more spread out when recognizing letters than sounding out pseudo-words.

FIGURE A4: PHONICS AND READING-RATE FLUENCY BOX PLOTS FOR GRADE 3, URDU



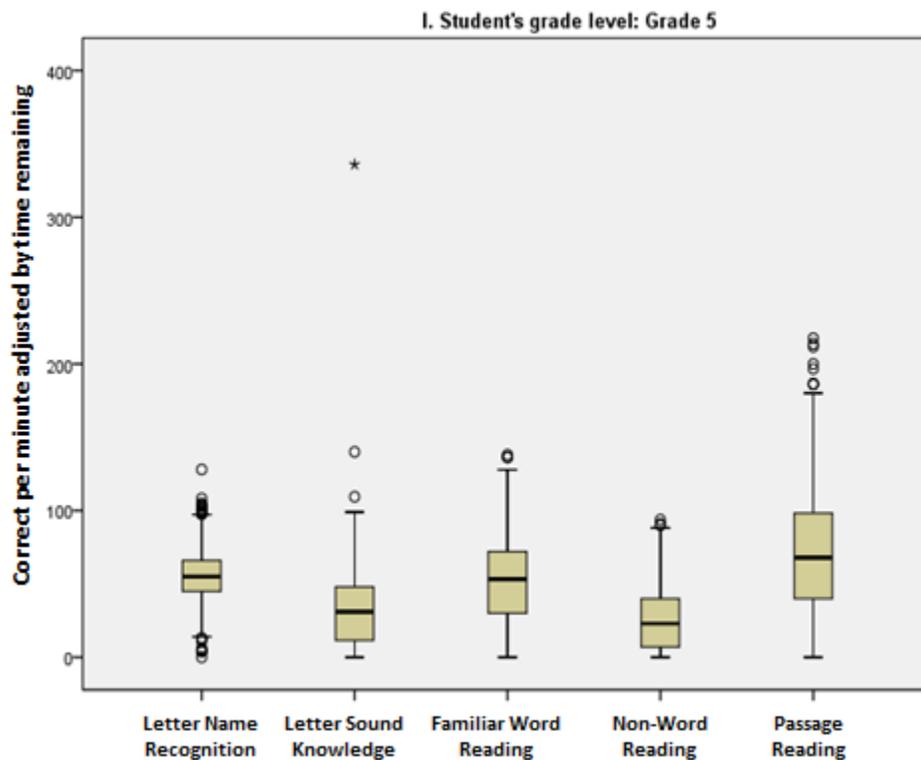
Grade 5, Urdu

For Urdu grade 5, the central tendency (the median speed) for each of the tasks ranged from about 30 (letter sound knowledge) to about 60 (passage reading) items per minute. It shows that the students had more fluency reading connected words than 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 A5: PHONICS AND READING-RATE FLUENCY BOX PLOTS FOR GRADE 5, URDU



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).

English

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 75.8 (rounded to 76). With this method, 76 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 101.1 (rounded to 101). Then 101 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 A3 below.

TABLE A3: ENGLISH THRESHOLDS FOR CWPM WITH 80 PERCENT COMPREHENSION

Category (Performance Level)	Grade 3		Grade 5	
	CWPM	% of Students	CWPM	% of Students
Non-Reader	0	15.7%	0	2.2%
Non-Fluent Reader	1 to 75	82.2%	1 to 100	86.5%
Fluent Reader	76 and above	2.1%	101 and above	11.3%
Total	--	100.0%	--	100.0%

Urdu

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 79.7 (rounded to 80). With this method, 80 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 98.2 (rounded to 98). Then 98 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 A4 below.

TABLE A4: URDU THRESHOLDS FOR WCPM WITH 80 PERCENT COMPREHENSION

Category (Performance Level)	Grade 3		Grade 5	
	CWPM	% of Students	CWPM	% of Students
Non-Reader	0	37.7%	0	3.4%
Non-Fluent Reader	1 to 79	57.7%	1 to 97	71.4%
Fluent Reader	80 and above	4.6%	98 and above	25.2%
Total	--	100.0%	--	100.0%

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

Fluency using fixed interval thresholds

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

TABLE A5: ENGLISH THRESHOLDS FOR CWPM WITH FIXED INTERVALS

Category (Performance Level)	CWPM	% of Students	
		Grade 3	Grade 5
Non-Reader	0	15.7%	2.2%
Early Reader	1 to 40	64.2%	25.8%
Intermediate Reader	41 to 80	18.7%	48.4%
Fluent Reader	81 to 120	1.3%	19.3%
Advanced Reader	121 and above	0.1%	4.2%
Total	--	100.0%	100.0%

TABLE A6: URDU THRESHOLDS FOR CWPM WITH FIXED INTERVALS

Category (Performance Level)	CWPM	% of Students	
		Grade 3	Grade 5
Non-Reader	0	37.7%	3.4%
Early Reader	1 to 40	45.3%	22.2%
Intermediate Reader	41 to 80	12.8%	35.2%
Fluent Reader	81 to 120	4%	27.9%
Advanced Reader	121 and above	0.3%	11.4%
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 for each language; 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 A7-A10 and Figures A6-A9) 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 in English, 100 percent of the non-readers have 0 percent comprehension and 23 percent of the fluent readers have 80 percent comprehension.

English

TABLE A7: GRADE 3 READING FLUENCY AND COMPREHENSION, ENGLISH

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	87%	10%	3%	0%	0%	0%	100%
Intermediate Reader	41 to 80	53%	26%	11%	8%	2%	1%	100%
Fluent Reader	81 to 120	23%	15%	15%	23%	23%	0%	100%
Advanced Reader	121 and above	0%	0%	0%	100%	0%	0%	100%

FIGURE A6: GRADE 3 READING FLUENCY AND COMPREHENSION, ENGLISH

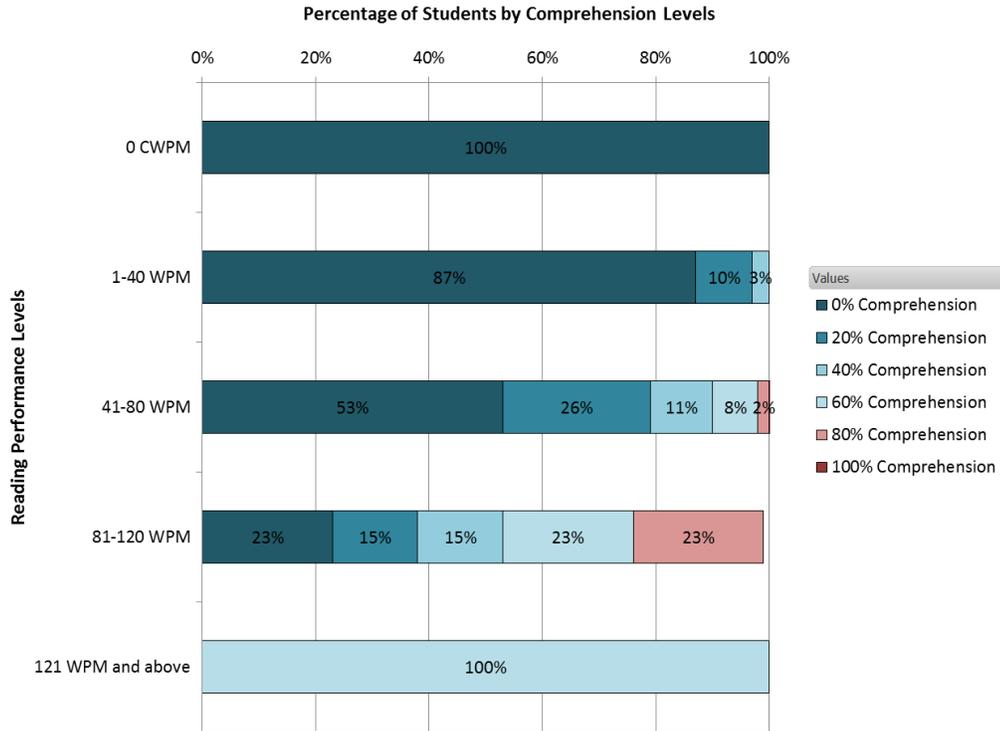
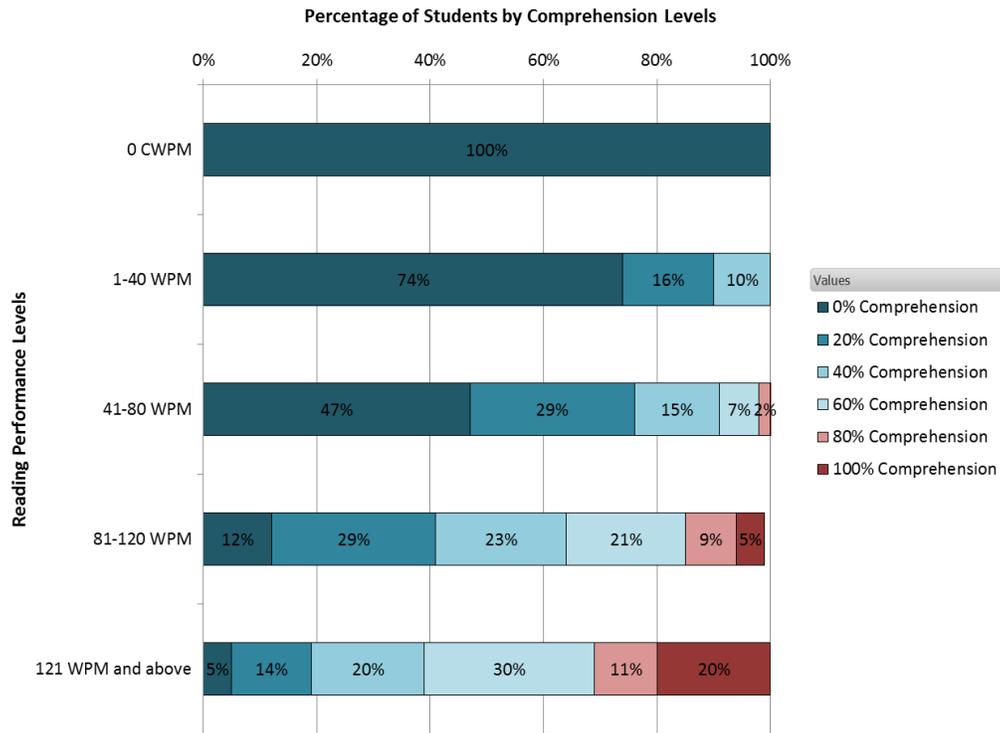


TABLE A8: GRADE 5 READING FLUENCY AND COMPREHENSION, ENGLISH

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	74%	16%	10%	0%	0%	0%	100%
Intermediate Reader	41 to 80	47%	29%	15%	7%	2%	1%	100%
Fluent Reader	81 to 120	12%	29%	23%	21%	9%	5%	100%
Advanced Reader	121 and above	5%	14%	20%	30%	11%	20%	100%

FIGURE A7: GRADE 5 READING FLUENCY AND COMPREHENSION, ENGLISH



Urdu

TABLE A9: GRADE 3 READING FLUENCY AND COMPREHENSION, URDU

Category (Performance Level)	CWPM	% of Students by Comprehension Level						Total
		0%	20%	40%	60%	80%	100%	
Non-Reader	0	100%	0%	0%	0%	0%	0%	100%
Early Reader	1 to 40	86%	11%	2%	1%	0%	0%	100%
Intermediate Reader	41 to 80	24%	26%	24%	17%	7%	2%	100%
Fluent Reader	81 to 120	8%	13%	26%	18%	18%	18%	100%
Advanced Reader	121 and above	0%	0%	0%	67%	33%	0%	100%

FIGURE A8: GRADE 3 READING FLUENCY AND COMPREHENSION, URDU

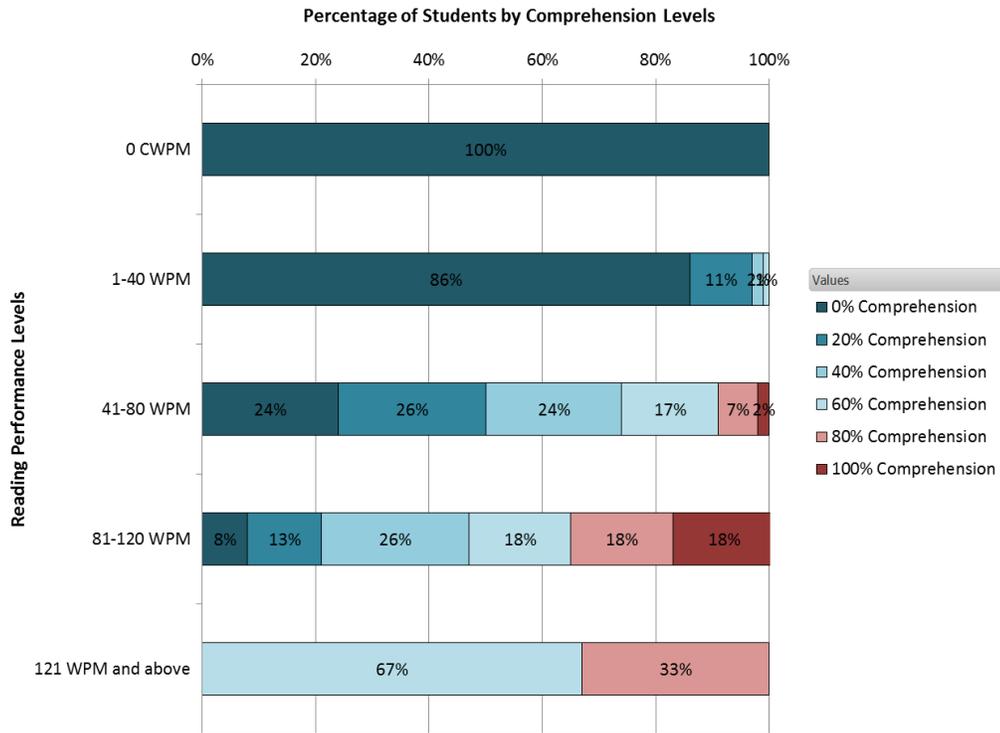
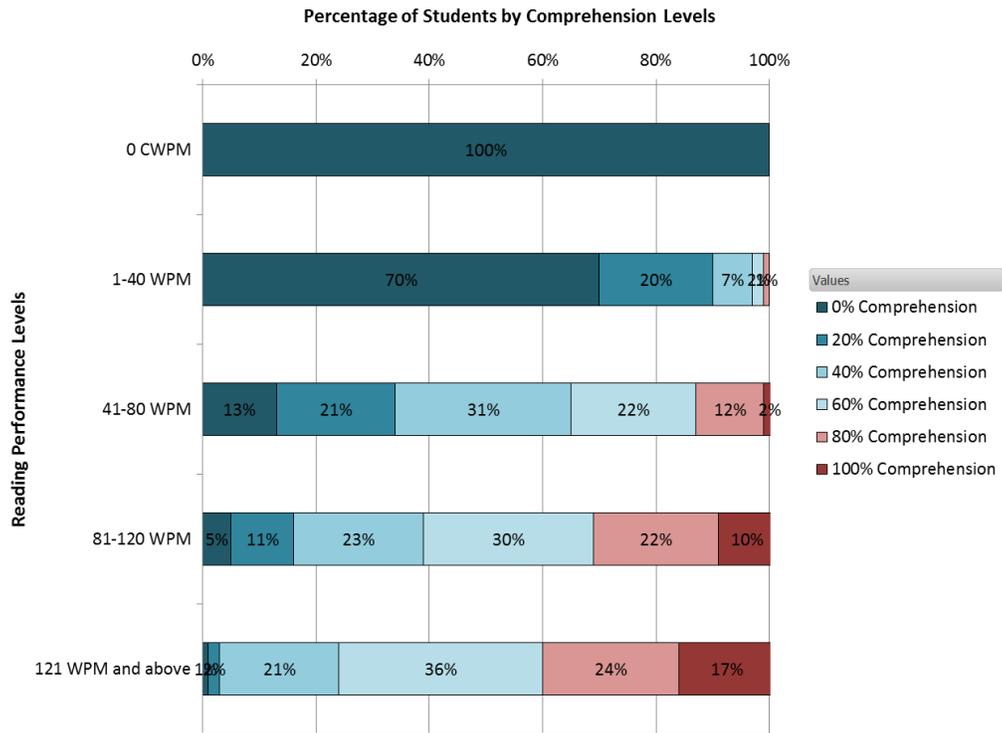


TABLE A10: GRADE 5 READING FLUENCY AND COMPREHENSION, URDU

Category (Performance Level)	CWPM	% of Students by Comprehension Level						Total
		0%	20%	40%	60%	80%	100%	
Non-Reader	0	100%	0%	0%	0%	0%	0%	100%
Early Reader	1 to 40	70%	20%	7%	2%	1%	0%	100%
Intermediate Reader	41 to 80	13%	21%	31%	22%	12%	2%	100%
Fluent Reader	81 to 120	5%	11%	23%	30%	22%	10%	100%
Advanced Reader	121 and above	1%	2%	21%	36%	24%	17%	100%

FIGURE A9: GRADE 5 READING FLUENCY AND COMPREHENSION, URDU



Main Results

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 (in grades 3 and 5 and English and Urdu) had 0 percent comprehension.
- Early Readers (1-40 CWPM) – Most of the early readers (87 percent at grade 3 and 74 percent at grade 5 in English and 86 percent at grade 3 and 70 percent at grade 5 in Urdu) had 0 percent comprehension. Almost none of them achieved 80 percent comprehension.
- Intermediate Readers (41-80 CWPM) – In English, about half of the intermediate readers had 0 percent comprehension (53 percent in grade 3 and 47 percent in grade 5), while in Urdu this percentage was about half as much (24 percent for grade 3 and 13 percent for grade 5). A small minority of them (3 percent at grades 3 and 5 in English and 9 percent at grade 3 and 14 at grade 5 in Urdu) achieved at least 80 percent comprehension.
- Fluent Readers (81-120 CWPM) – A minority of the fluent readers had 0 percent comprehension (in English, 23 percent at grade 3 and 12 percent at grade 5 and in Urdu 8 percent in grade 3 and 5 percent in grade 5).
- Advanced Readers (121 CWPM and above) – A small percentage of the advanced readers had 0 percent comprehension. More than a third achieved at least 80 percent comprehension. (in English, 23 percent at grade 3 and 31 percent at grade 5 and in Urdu 36 percent in grade 3 and 41 percent in grade 5).

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.