USAID FROM THE AMERICAN PEOPLE

Maktabgacha va maktab ta'limi vazirligi

RTI INTERNATIONAL

FY2023–YEAR 4

# Early Grade Reading and Mathematics Endline Impact Evaluation Report, 2023

UZBEKISTAN EDUCATION for EXCELLENCE PROGRAM

FLORIDA STATE UNIVERSITY 1851

MISSISSIPPI STATE UNIVERSITY

# Uzbekistan Education for Excellence Program

Early Grade Reading and Mathematics Endline Impact Evaluation Report, 2023
Cooperative Agreement No. 72011519CA00004

**Submitted to:**

Abdugani Bazarov
Agreement Officer's Representative
USAID/Central Asia/Uzbekistan
3 Moyqorghon Street
Tashkent, Uzbekistan 70093
Tel.: (99871) 140 2486
Fax: (99871) 120 6309
abazarov@usaid.gov

**Submitted by:**

RTI International
3040 East Cornwallis Rd
Research Triangle Park, NC 27709

Dr. Carmen Strigel
Project Manager
cstrigel@rti.org

September 25, 2023

# ACKNOWLEDGEMENTS

On behalf of the United States Agency for International Development (USAID) Uzbekistan Education for Excellence Program, the authors of this report wish to acknowledge the contributions of all collaborators who made it possible to complete the Program's Early Grade Reading and Mathematics Assessment (EGRA/EGMA) Endline report.

We acknowledge and appreciate the close collaboration of colleagues from the Ministry of Preschool and School Education (MoPSE) of the Republic of Uzbekistan, who supported us throughout the process.

We wish to acknowledge the school directors and teachers who supported the assessment in the 176 schools in the EGRA/EGMA Endline. Without your effort and dedication, we would not have been able to achieve our goal of testing 4,141 students.

We acknowledge Middle Asia Management Consulting for managing the data collection process.

We acknowledge the Uzbekistan Education for Excellence Program staff who contributed to the assessment process by providing technical and logistical support for training and data collection. Specifically, we thank Temurbek Rakhmatov, Furkat Sharipov, Retno Utaira, and Bakhtiyor Mamatkulov.

In addition, we acknowledge the RTI International home office staff who contributed, including Laiba Bahrahwar for data cleaning, analysis, and supporting report writing, and Joseph DeStefano, Jennifer Ryan, Margaret (Peggy) Dubeck, Peter Muyingo, Yasmin Sitabkhan, and Geri Burkholder for writing the report.

Finally, we express our gratitude to Abdugani Bazarov, the Program's Agreement Officer's Representative at USAID/Uzbekistan, for his support and encouragement throughout the work.

# EXHIBITS

# ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| COVID-19 | coronavirus disease 2019 |
| cwpm | correct words per minute |
| EFL | English as a Foreign Language |
| EGMA | Early Grade Mathematics Assessment |
| EGRA | Early Grade Reading Assessment |
| ICT | Information and Communication Technology |
| MoPSE | Ministry of Preschool and School Education |
| PPS | probability proportional to size |
| RTI | RTI International |
| SES | socioeconomic status |
| STB | student textbook |
| TG | teacher guide |
| TIMSS | Trends in International Mathematics and Science Study |
| TLM | teaching and learning material |
| TPD | teacher professional development |
| ULA | Uzbek Language Arts |
| USAID | United States Agency for International Development |
| wpm | words per minute |

# EXECUTIVE SUMMARY

The Government of Uzbekistan Ministry of Preschool and School Education (MoPSE) is committed to an ambitious program of systematic and comprehensive reforms. The country aims to create an education system that can produce graduates with critical thinking, problem solving, and practical skills that will enable them to succeed.

To support the MoPSE in achieving its reform agenda, the United States Agency for International Development (USAID) initiated the 4-year Uzbekistan Education for Excellence Program (the Program) on December 9, 2019, which will end on December 8, 2023.

This Program aims to provide the expertise and experience needed to help the MoPSE achieve and sustain three overarching results:

(1) Improved Uzbek Language Arts (ULA) and Mathematics outcomes in grades 1–4.

(2) Enhanced Information and Communication Technology instruction for grades 1–11, and

(3) Improved English as a Foreign Language instruction in grades 1–11.

Cross-cutting themes include capacity building, gender equality and social inclusion, transparency, local ownership, and sustainability.

## PURPOSE OF THE EVALUATION

To evaluate the impact of the Program's reading and mathematics components on student learning, this report will compare results from the endline assessment against the values from the baseline assessment. The baseline Early Grade Reading and Mathematics Assessments (EGRA and EGMA, respectively) were conducted in November and December 2021. The endline was administered in May 2023.

The EGRA/EGMA baseline was originally planned to assess students completing grades 2 and 4 at the end of the 2019–2020 school year, in May 2020. However, the assessment was postponed because of coronavirus disease 2019 (COVID-19). The Program decided to assess grade 3 and 5 students close to the start of the school year in November–December 2021, as proxies for students completing grades 2 and 4. Therefore, throughout this report, the baseline results are presented as grade 2 and 4 results to facilitate comparison with the results of the endline, which was conducted at the end of grades 2 and 4. Baseline results were also adjusted to account for the time difference, relative to the start and end of the school year, between baseline (Nov/Dec 2021) and endline (May 2023; see **Annex A**).

At baseline, 2,334 grade 3 and 2,324 grade 5 students from 140 Program and 59 comparison schools participated in the EGRA/EGMA evaluation study. The endline included 2,065 grade 2 and 2,076 grade 4 students in 126 Program and 50 comparison schools.

The EGRA/EGMA endline sought to answer the following research question: What is the overall impact of the Uzbekistan Education for Excellence Program in grades 2 and 4 on Uzbek language reading and mathematics skills?

## SUMMARY OF FINDINGS

The Program used EGRA to measure changes in basic reading skills. In both grade 2 and grade 4, students were assessed on decoding and the higher order skills of fluency and comprehension as shown in **Exhibit ES-1.** Due to the difference between when the Program

conducted the baseline and endline assessments, the results show adjusted baseline averages and adjusted difference values, besides the original values. Details of our adjustment methodology are provided in **Annex A**. A comparison of the adjusted baseline to the endline average scores for grade 2 EGRA showed that the greatest improvement was in reading comprehension, followed by nonword decoding. Performance on the oral reading fluency remained almost the same at endline. Results for grade 4 showed significant improvement ($p$<0.001) in student achievement on all subtasks, with the greatest change (~12 correct words per minute [cwpm]) in oral reading fluency.

**Exhibit ES-1. Uzbek Language Literacy Achievement, by Grade, Task**

| Grade1 | Task | Baseline Average | Adjusted Baseline Average | Endline Average | Difference [Endline – Baseline] | Adjusted Difference [Endline – Baseline] |
|---|---|---|---|---|---|---|
| Grade 2 | Nonwords (cwpm) | 38.9 | 31.6 | 35.4 | -3.5 | +3.8** |
| | Oral reading fluency (cwpm) | 47.9 | 39.7 | 40.5 | -7.4 | +0.8 |
| | Reading comprehension (percent score) | 79.1 | 61.6 | 69 | -10.1 | +7.4*** |
| Grade 4 | Nonwords (cwpm) | 47.2 | 41.1 | 47.8 | 0.6 | +6.7*** |
| | Oral reading fluency (cwpm) | 68.3 | 58.9 | 70.6 | 2.3 | +11.7*** |
| | Silent reading comprehension (percent score) | 64.8 | 56.8 | 64.9 | 0.1 | +8.1*** |

*** p<0.001, ** p<0.01, * p<0.05

**Exhibit ES-2** presents Program impact on grade 2 and 4 student mathematics skills. Students were assessed on different mathematics skills (described in Section 3.2) appropriate for their grade. A comparison of the adjusted baseline average scores to the endline average scores for grade 2 EGMA showed a statistically significant improvement +8.1% (p<0.001) in the missing number subtask and a slight increase of 1.5% in the three-dimensional spatial thinking. There was a decline in all other subtasks (e.g., word problems, addition, subtraction, and relational reasoning) with the most significant decrease (-7.4%) in the word problems subtask. In grade 4, there was a large improvement of around 6 percentage points in the overall average score at endline (baseline adjusted score of 53% vs 59% correct at baseline). The greatest improvement (+8.5%) was in the numbers and operations subtask (p<0.001).

1 During baseline the Program assessed grades 3 and 5 closer to the beginning of school year as proxy to 2 and 4 years of schooling, respectively. However, during endline the Program assessed grade 2 and 4 at the end of school year.

**Exhibit ES-2. Mathematics Achievement, by Grade, Subtask**

| Grade | Task | Baseline Average | Adjusted Baseline Average | Endline Average | Difference | Adjusted Difference |
|---|---|---|---|---|---|---|
| Grade 2 | Missing number *(percent score)* | 67.9 [±2.0] | 66.5 | 74.6 | +6.7 | +8.1*** |
| | Word problems *(percent score)* | 75.4 [±2.1] | 72.2 | 64.8 | -10.6 | -7.4*** |
| | Addition *(percent score)* | 83.2 [±1.8] | 79.9 | 76.6 | -6.6 | -3.3 |
| | Subtraction *(percent score)* | 74.7 [±2.3] | 72.1 | 66 | -8.7 | -6.1* |
| | Relational reasoning *(percent score)* | 62.2 [±3.0] | 59.0 | 53.9 | -8.3 | -5.1 |
| | Three-dimensional (3D) spatial thinking *(percent score)* | 62.8 [±2.2] | 62.5 | 64 | +1.2 | +1.5 |
| Grade 4 | Overall mathematics *(percent score)* | 57.4 [±2.5] | 52.6 | 59 | +1.6 | +6.4*** |
| | Numbers and operations *(percent score)* | 61.0 [±2.5] | 54.8 | 63.3 | +2.3 | +8.5*** |
| | Geometry *(percent score)* | 41.7 [±2.4] | 38.5 | 43.1 | +1.4 | +4.6** |
| | Measurement *(percent score)* | 53.5 [±3.3] | 50.0 | 51.3 | -2.3 | +1.3 |
| | Statistics *(percent score)* | 60.7 [±3.6] | 59.3 | 63.4 | +2.7 | +4.1 |

*** $p<0.001$, ** $p<0.01$, * $p<0.05$

Presenting results by gender is an important way to investigate equity issues. **Exhibit ES-3** presents achievement on selected tasks by gender. Overall, we see almost similar gains or losses for girls and boys from the adjusted baseline to endline. In grade 2, girls outperformed boys in all EGRA tasks. The difference was particularly noticeable for oral reading fluency with grade 2 girls reading on average 6.5 more cwpm than boys. Girls also performed significantly better in reading comprehension. Results are similar for grade 4 with girls reading on average 11 more cwpm than boys. For mathematics, boys generally outperformed girls in both grades. The differences between genders in reading for grade 2, however, were not statistically significant. For grade 4, apart from oral reading comprehension, most of the differences in grade 2 gains or losses between genders were not statistically significant; however, girls' grade 4 gains in oral reading fluency were statistically significant. Differences in student achievement by gender are further explored in Section 3.3.

**Exhibit ES-2. Mathematics and Reading Achievement, by Grade and Gender for Intervention Schools**

| Grade | Subject | Task | Gender | Adjusted Baseline Average | Endline Average | Gain |
|---|---|---|---|---|---|---|
| Grade 2 | Reading | Oral reading fluency *(cwpm)* | boys | 34.8 | 37.2 | +2.4 |
| | | | girls | 44.7 | 43.7 | -1.0 |
| | | Reading comprehension *(percent score)* | boys | 60.3 | 67.3 | +6.9** |
| | | | girls | 63.0 | 70.7 | +7.7*** |
| | Mathematics | Relational reasoning *(percent score)* | boys | 60.1 | 56.5 | -3.6 |
| | | | girls | 58.0 | 51.4 | -6.6 |
| | | 3D spatial thinking *(percent score)* | boys | 66.1 | 67.2 | +1.1 |
| | | | girls | 58.8 | 60.7 | +1.9 |
| Grade 4 | Reading | Oral reading fluency *(cwpm)* | boys | 53.0 | 65 | +12*** |
| | | | girls | 64.8 | 76 | +11.2*** |
| | | Silent reading comprehension *(percent score)* | boys | 58.0 | 64.4 | +6.4** |
| | | | girls | 55.6 | 65.4 | +9.8*** |
| | Mathematics | Overall mathematics *(percent score)* | boys | 52.9 | 60.3 | +7.4*** |
| | | | girls | 52.2 | 57.8 | +5.6** |

*** $p<0.001$, ** $p<0.01$, * $p<0.05$

A comparison of results across students' socioeconomic status (SES) indicates that overall, students across the three SES tertiles scored similarly, on average, in both grades across all reading subtasks and mathematics domains. More detail on how the Program defined and measured SES is in **Annex B**.

Overall, the endline shows that the Program had significant gains on students' oral reading comprehension in grade 2, and nonword decoding, oral reading fluency, and silent reading comprehension in grade 4. Despite the impressive progress, there is still room for improvement especially in grade 2 oral reading fluency that remained almost the same at endline. Regarding Mathematics, the Program had a large and positive effect on student performance in only the missing numbers domain in grade 2. In grade 4, the gain in the overall mathematics score was statistically significant. These results imply that the Program was more successful in improving grade 4 mathematics abilities than in grade 2. In both grades, girls outperformed boys in EGRA while boys outperformed girls in Mathematics.

# SECTION 1: BACKGROUND

## 1.1 PROGRAM OVERVIEW

The Government of Uzbekistan Ministry of Preschool and School Education (MoPSE) is committed to an ambitious program of systematic and comprehensive reforms. The country aims to create an education system that can produce graduates with critical thinking, problem solving, and practical skills that will enable them to succeed.

To support the MoPSE in achieving its reform agenda, the United States Agency for International Development (USAID) initiated the 4-year Uzbekistan Education for Excellence Program (the Program) on December 9, 2019. The Program is implemented by a consortium of implementing partners including RTI International (RTI) as the consortium lead and Florida State University and Mississippi State University as partners. The RTI consortium provides the expertise and experience needed to help the MoPSE achieve and sustain three overarching results:

1. Improved Uzbek Language Arts (ULA) and Mathematics outcomes in grades 1–4.
2. Enhanced Information and Communication Technology (ICT) instruction for grades 1–11; and
3. Improved English as a Foreign Language (EFL) instruction in grades 1–11.

## 1.2 LIFE OF THE PROGRAM ANTICIPATED ACHIEVEMENTS

Over the life of the Program, in close collaboration with the MoPSE, the Program:

- Developed relevant and appropriate student learning standards for ULA, Mathematics, ICT, and EFL.
- Customized or developed and piloted revised student textbooks (STBs) and teacher guides (TGs).
- Designed and implemented an in-service teacher professional development (TPD) approach.
- Conducted Program monitoring, evaluation, and learning activities, including impact evaluation research.

The Program developed new approaches to curriculum products development and support for TPD, including a digital platform for standards and instructional materials. These materials and approaches were used as the centerpieces to help enhance teachers' capacity to understand, apply, reflect on, and improve classroom practices. It was expected that the

improvements in curriculum products and in teacher capacity would translate into improvements in student achievement over time.

Between baseline and endline assessments, the Program strategically enhanced student learning outcomes through various initiatives including curriculum development, digital platform creation, TPD, monitoring and evaluation, capacity building workshops, and close collaboration with stakeholders.

Prior to the start of the 2022–2023 school year the Program developed and distributed ULA and Mathematics STBs and TGs. The new teaching and learning materials (TLMs) were aligned with modern standards that emphasized student-centered learning. Furthermore, the Program established a central digital platform housing instructional materials and resources, enhancing content accessibility and interactive learning, and aiding lesson planning for teachers in key subjects of the Program.

In addition, the Program implemented an evidence based continuous TPD approach that was comprised of an effective quality assurance feedback loop. The Program's TPD approach provided a combination of three 2-day group-based training workshops and short, regular, and frequent learning opportunities during weekly Methodological Days[2] that are part of Uzbekistan's education system. This purposeful design enabled teachers to practice and master specific techniques before the next training. Each training session built upon the previous training. The Program utilized a cascade model: Beginning with the training of 80 master trainers (Tier 1), who, in turn, trained 800 teacher trainers (Tier 2), who then cascaded the training to approximately 9,000 primary grade teachers from the 919 schools in Sirdaryo and Namangan that participated in the pilot. In addition, the Program introduced online self-paced learning modules and Zoom question and answer sessions, allowing teachers to seek clarifications for their questions, exchange ideas, and maintain a sense of community outside of the structured learning events. The TPD program took place from August 2022 to April 2023, during which time the Program offered this combination of training workshops and Methodological Days. Over the 9 months, the Program developed content for 7 main topics for the ULA and Mathematics primary teachers.

The Program's robust monitoring and evaluation activities helped examine the Program's progress along its entire theory of change and consisted of monitoring the distribution of TLM, monitoring gains in teachers' knowledge, skills, and attitudes emanating from the TPD program (add link to TPD Effectiveness Study report) and the use of the new TLMs in the classroom (add link to the TLM Uptake Study report). The Program also collected feedback through Telegram to inform the revision of the TLM (see the Desk Review of ULA and Mathematics report) and conducted cognitive interviews with 60 grade 2 students half of which from comparison schools. The purpose of the cognitive assessment study was to better understand the development of higher-order skills in reading and mathematics. These activities complemented and informed the EGRA/EGMA baseline and endline assessments and the discussion of results in this report.

2 Methodological Days occur once a week and were established by MoPSE to provide primary teachers with dedicated time for class preparation and professional development.

# SECTION 2: STUDY DESIGN

## 2.1 PURPOSE OF THE STUDY

This EGRA/EGMA study seeks to evaluate the impact on learning outcomes of the USAID Uzbekistan Education for Excellence Program in mathematics and Uzbek language reading. By analyzing the baseline and endline results, the assessment seeks to provide valuable insights into the effectiveness of the Program's interventions and their contribution to enhancing students' proficiency and skills in these subjects.

## 2.2 RESEARCH QUESTIONS

The main goal of the EGRA/EGMA assessment is to measure the impact the Program had on student learning outcomes. To achieve this goal, the EGRA/EGMA endline addressed the following research question:

> *What is the overall impact of the Uzbekistan Education for Excellence Program in grades 2 and 4 on Uzbek language reading and mathematics skills?*

To establish a baseline against which to measure its impact, the Program conducted the EGRA/EGMA baseline assessment in November–December 2021. The baseline was originally scheduled for May 2020, to assess grade 2 and 4 students at the end of the 2019–2020 school year. However, because of the global COVID-19 pandemic, the assessment had to be postponed to November–December 2021. As these months represented the beginning of the school year, instead of the end, the Program administered the assessments to students in grades 3 and 5 as proxies for students who had completed 2 and 4 years of schooling. The endline was conducted in May 2023, including students who were completing grades 2 and 4 in the same schools assessed at baseline. It was not possible to conduct the endline at the same time in the school year as the baseline given the Program's December 8 end date.

## 2.3 MEASURING IMPACT

The baseline and endline assessments originally included students in both Program schools and comparison schools. As the Program was implemented in Namangan and Sirdayo Regions, Program schools included in the impact evaluation were from those regions. Comparison schools came from Jizzakh Region. At baseline, sample performance was balanced, with small and acceptable differences between the comparison school and Program school averages. We applied a difference-in-differences analysis at endline to measure impact. This analysis is a calculation of the difference between the comparison and Program schools' average gains in learning outcomes. The analysis revealed that comparison schools outperformed Program schools on all grade 2 and grade 4 reading subtasks and mathematics domains, and improvement in grade 4 mathematics scores was exceptionally high in comparison schools. These results were surprising. To understand why comparison schools performed better than Program schools, we did the following:

- Conducted further analysis of the data, and the patterns in the comparison school data indicate that a non-representative selection of students at endline in enough schools skewed the data considerably (i.e., the best students, rather than random students, were selected in some schools).

- Reviewed the assessment results for each school and aligned them to the assessor teams that were administering the tests in each school. This review revealed that a certain assessor administered the EGRA in a significant percentage of the higher performing comparison schools. For example, the same assessor was present in 44% of schools with average scores of over 80% correct on grade 2 word problems. That same assessor also was present in 36% of schools with average grade 4 reading comprehension scores over 80% correct.
- Conducted a rapid survey of some of the comparison schools to find out how the tests were administered during the endline. A small number of initial interviews revealed some potential issues; for example, most of the comparison assessors were teachers from the sampled comparison schools (which was not the case for the Program sample, for which we used contracted assessors). In some cases, a teacher who was trained as an assessor was present to help at her school on the day of the assessment. Data indicate that when a school's own teacher, who was trained as an EGRA assessor, was present at the school on the day of assessment, the scores recorded were higher than those recorded when a school's own teacher, trained on EGRA, was not present at the school on the day of assessment.

Based on the above issues related to results in the comparison schools at endline, the comparison sample was dropped, and Program impact was evaluated by looking only at the baseline and endline changes within Program schools. The baseline scores were also adjusted to accommodate for the difference between when data were collected for the baseline and endline—that is, beginning of the school year for baseline, and end of the school year for endline. We calculated this adjustment by multiplying the baseline scores by a fraction (total days in school at endline divided by total days in school at baseline). See **Annex A** for more details.

## 2.4 SAMPLE METHODLOGY

The population of interest were grade 2 and grade 4 students in all schools in the UEEP program, which included government primary schools in the Namangan and Sirdaryo regions. The 2019-2020 census list of schools was used as the sampling frame. This list included a total of 199 schools (122 in Namangan; 77 in Sirdaryo), 12,465 grade 2 students (9,021 in Namangan; 3,444 in Sirdaryo) and 14,592 grade 4 students (10,525 in Namangan; 4,067 in Sirdaryo).

A two-stage sample of schools and students was conducted for both baseline and endline. The Program conducted the endline assessment in the same schools assessed at baseline. Schools were selected at baseline using a probability proportional to size (PPS) methodology, meaning students in large schools had the same probability of being selected as students attending smaller schools. Within the selected schools, grade 2 students were stratified by gender, and 6 girls and 6 boys were randomly selected with equal probability. The same was done for the grade 4 students.

Since the same schools were visited at endline, the endline school weights were inherited from the baseline. Weights for each stage were calculated as the inverse of the probability of selection. The school weights were scaled to the region's population of schools. The students' weights were scaled to the population of grade 2 and grade 4 students in the given region.

## 2.5 SCHOOL AND STUDENT CHARACTERISTICS

**Exhibit 1** presents school and student sample characteristics for baseline and endline.

**Exhibit 1. School Sample Characteristics by Grade**

| Timepoint | Region | Namangan Region | | | | Sirdaryo Region | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Grade 2 | | Grade 4 | | Grade 2 | | Grade 4 | |
| | | Girls | Boys | Girls | Boys | Girls | Boys | Girls | Boys |
| Baseline[3] | Number of Schools | 77 | | | | 63 | | | |
| | Number of Students | 443 | 443 | 455 | 452 | 368 | 369 | 365 | 357 |
| Endline | Number of Schools | 67 | | | | 59 | | | |
| | Number of Students | 405 | 405 | 401 | 398 | 342 | 348 | 343 | 360 |

As **Exhibit 1** shows, the endline included 126 Program sampled schools from the two targeted regions and a total of 3,002 students (1,491 girl and 1,511 boys). Although the Program planned to conduct an endline assessment in all schools that were selected for the baseline assessment, we had to drop some schools that were undergoing reconstruction or changing their structure from ordinary schools to specialized schools during the time of the endline. The magnitude of the difference in the number of schools in the sample at baseline and endline, however, is considered permissible and does not affect the precision of the results.

## 2.6 ASSESSOR TRAINING AND DATA COLLECTION

Training of assessors started with the master trainer training on April 19–21, 2023. The Program trained a total of 8 master trainers in this initial training. The master trainer training was followed by the assessor training April 24–28, 2023. A total of 72 assessors were trained, and out of these, 70 assessors scored at least 90% or above on the Assessor Accuracy Measure. We selected only these to collect the endline data. Data collection took place between May 1–18, 2023. During this time, assessor teams, consisting of four members each, visited the sampled schools, with the trainers acting as field coordinators and supporting the assessor teams. Program staff conducted spot checks and supported assessors during data collection.

## 2.7 SURVEY INSTRUMENTS

**Exhibit 2** provides an overview of the EGRA and EGMA tasks administered by grade.

**Exhibit 2. Overview of EGRA/EGMA Task by Grade**

| Language | Grade 2 | Grade 4 |
|---|---|---|
| **EGRA** | | |
| Assessed in Uzbek | ▪ Nonword decoding<br>▪ Oral reading fluency (grade 2-level text)<br>▪ Oral reading comprehension (grade 2-level text) | ▪ Nonword decoding<br>▪ Oral reading fluency (grade 4-level text)<br>▪ Silent reading comprehension (grade 4-level text) |

3 Baseline student numbers are for grades 3 and 5 instead of grades 2 and 4.

| Exhibit 2. Overview of EGRA/EGMA Task by Grade | | |
|---|---|---|
| **Language** | **Grade 2** | **Grade 4** |
| **EGMA** | | |
| Instructions given in the language of instruction | ▪ Missing number<br>▪ Addition/subtraction<br>▪ Word problems<br>▪ Relational reasoning<br>▪ Three-dimensional (3D) spatial thinking | ▪ Numbers and operations<br>▪ Geometry<br>▪ Measurement<br>▪ Statistics |

## 2.7 LIMITATIONS

The Program baseline data were collected under a separate project (All Children Reading Asia) which resulted in the data being collected before the curriculum content was defined and piloted. This limited the Program's ability to change the assessment before the full baseline was complete.

In the baseline, many students already achieved grade level foundational skills; notably in phonemic awareness and letter sound knowledge as well as addition and subtraction. For this reason, in the endline, the Program dropped entire subtasks but was not able to replace these subtasks with higher-level assessment items.

Study findings must also be considered in context. Due to delays, specifically owing to the coronavirus disease 2019 (COVID-19) pandemic, the completion of the TLMs and thereby the pilot, were delayed by one school year. This shortened the period students and teachers used the Program materials to a single school year, i.e., to just 9 months of instruction without the chance to pilot complete TLMs before.

The pilot was supported by a TPD approach that focused on specific techniques that built teachers' skills over time which meant that students were only exposed to the complete range of techniques and content by the end of the school year. Therefore, teachers did not have a comprehensive and practical understanding of the content that would be covered in a given grade until the end of the pilot year.

# SECTION 3: MAIN RESULTS

## 3.1 GRADE 2 AND GRADE 4 EGRA FINDINGS

### 3.1.1 Grade 2 EGRA Findings

There were three EGRA subtasks administered in the grade 2 EGRA: nonwords, oral reading, and reading comprehension. As shown in **Exhibit 3**, on average at endline, students scored 35.4 correct words per minute (cwpm) on the nonwords subtask, 40.5 cwpm on the oral reading subtask, and 69% correct on the reading comprehension subtask. As expected, students were less adept at reading nonwords than at reading connected text of real words. The increase of 3.8 cwpm in nonwords suggests that students have had more phonics instruction, and they are using that knowledge to read unfamiliar words (as all nonwords would be unfamiliar). That increase in nonword reading a likely contributor to the increases in oral reading fluency. The 7.4% increase in reading comprehension scores were likely influenced by the improved word recognition ability (3.8 cwpm) and reading rate (+0.8 cwpm).

**Exhibit 3. Average Grade 2 Reading Achievement by Task, Baseline and Endline**

| Grade 2 | Adjusted Baseline | Endline Average | Change from Adjusted Baseline |
|---|---|---|---|
| Nonword reading (cwpm) | 31.6 | 35.4 | +3.8** |
| Oral reading fluency (cwpm) | 39.7 | 40.5 | +0.8 |
| Reading comprehension (% correct) | 61.6 | 69.0 | +7.4*** |

*** p<0.001, ** p<0.01, p<0.05

To further understand the shifts in student performance from baseline to endline, we analyzed the grade 2 EGRA results by reading proficiency levels developed earlier in Program that measure student performance based on the oral reading fluency and reading comprehension subtasks. **Exhibit 4** defines the various categories of readers.

**Exhibit 4. Grade 2 Reading Proficiency Levels**

| Category | Definition* |
|---|---|
| Low Grade 2 Reader | Fewer than 15 correct words per minute (cwpm) |
| Emergent Grade 2 Reader | 15–45 cwpm and reading comprehension 60% or above |
| Proficient Grade 2 Reader | 45–61 cwpm and reading comprehension 80% or above |
| Fluent Grade 2 Reader | 61 or more cwpm and reading comprehension 80% or above |

*Definitions derived from performance on a grade 2 passage

**Exhibit 5** presents shifts in grade 2 reading proficiency levels between baseline and endline. The analysis shows a very slight decrease in the percentage of students in the low grade 2 reader category, but a substantial increase in the percentage of students in the fluent grade 2 reader category.

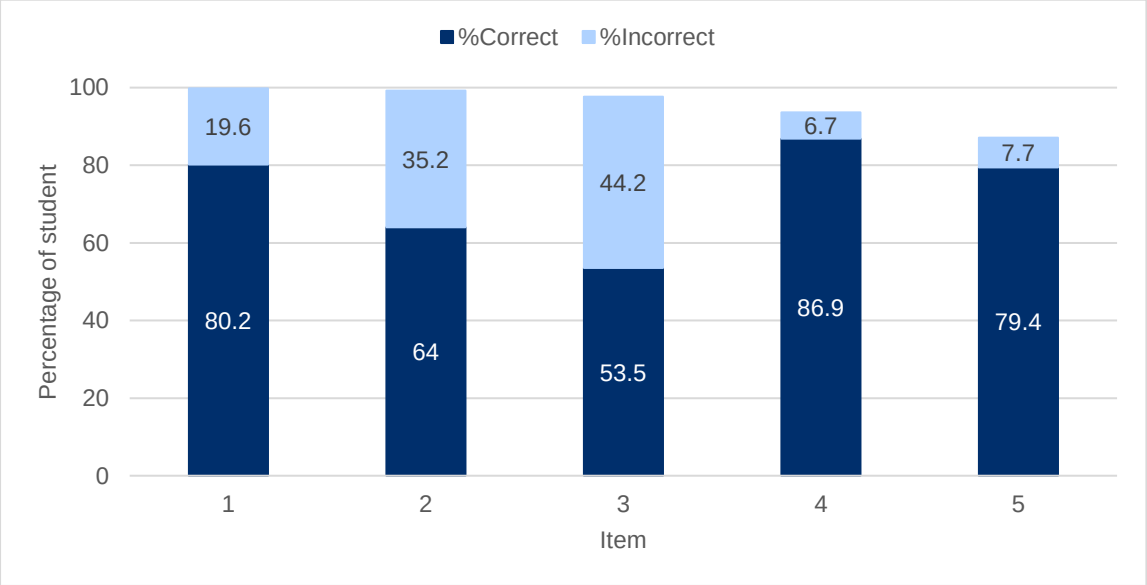
**Exhibit 5. Shifts in Grade 2 Reading Proficiency Levels, Baseline and Endline (Percentages)**

| Category | Adjusted Baseline | Endline | Change from Adjusted Baseline |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Low Grade 2 Reader | 10.6% | 9.7% | -0.9 |
| Emergent Grade 2 Reader | 53.1% | 58.6% | +5.5 |
| Proficient Grade 2 Reader | 30.0% | 19.5% | -10.5 |
| Fluent Grade 2 Reader | 6.3% | 12.2% | +5.9 |

A breakdown of grade 2 student performance in the form of score distributions at endline is presented in **Annex C**. Of note is that the relationship between nonwords and reading connected words is as expected. Most students scored between 20 to 60 cwpm on both nonwords and oral reading subtasks. Skill in reading unfamiliar words (i.e., nonwords) contributes to the ability to read connected text accurately and automatically. The reading comprehension subtask consisted of four lower-level questions that were text-based (i.e., explicit) and one higher-order question that required making an inference. On the reading comprehension subtask, most students (>50%) scored either a 4 or a 5 out of the 5 comprehension questions, which is 80%–100%. Item level analysis for the reading comprehension subtask at endline is shown in **Exhibit 6** indicates that 80% of grade 2 students got question 1 correct, 64% got question 2 correct, and 54% got question 3 correct, whereas 87% got question 4 correct and 79% got question 5 correct.

**Exhibit 6. Reading Comprehension Scores by Item**




The reading passage questions are arranged such that the comprehension questions are revealed as the student reads more of the passage. So, question 5 would only be asked if the student read to the end of the passage. Yet, students at endline demonstrated greater levels of accuracy for Questions 4 and 5 than questions 1, 2, and 3. Results also showed that the students who read further into the passage and qualified to be asked questions 4 and 5, read words more accurately and faster than their peers who were asked fewer questions. Generally, and in this case, these two skills (accuracy and speed) contribute to higher reading comprehension. Students who were less accurate and slower had limited resources to support their reading comprehension. Comprehension was also an area of learning on which the Program specifically focused its intervention. With specially designed TLMs aimed at improving comprehension, students could grasp higher level comprehension questions with greater accuracy.

Overall, for grade 2, endline results suggest some positive impact, especially on reading comprehension. Students were able to answer questions with greater levels of accuracy, and the difference between adjusted baseline scores and endline scores was also found to be statistically significant, implying that the program was successful in improving this critical reading skill.

### 3.1.2 Grade 4 EGRA Findings

There were three EGRA subtasks administered in the grade 4 EGRA: nonwords, oral reading, and silent reading comprehension. On average at endline, students scored 47.8 cwpm on the nonwords subtask, 70.6 cwpm on the oral reading subtask, and 64.9% correct on the silent reading comprehension subtask, as shown in **Exhibit 7**. As in grade 2, skill at reading nonwords was lower in grade 4 than skill at reading connected text of real words. The increase of 6.7 cwpm in nonword reading suggests that students had had instruction in word recognition (i.e., phonics and word study), and they were using that knowledge to read unfamiliar words (as all nonwords would be unfamiliar). That increase in nonword reading likely contributed to the nearly 12 words increase in oral reading fluency. The 8.1 cwpm increase in silent reading comprehension is likely influenced by the improved word recognition ability (+6.7) and rate (+11.7).

**Exhibit 7. Grade 4 Reading Achievement by Task, Baseline and Endline**

| Grade 4 | Adjusted Baseline | Endline Average | Change from Adjusted Baseline |
|---|---|---|---|
| Nonword reading (cwpm) | 41.1 | 47.8 | +6.7*** |
| Oral reading fluency (cwpm) | 58.9 | 70.6 | +11.7*** |
| Silent reading comprehension (% correct) | 56.8 | 64.9 | +8.1*** |

*** $p<0.001$, ** $p<0.01$, * $p<0.05$

We also analyzed grade 4 EGRA results by reading proficiency levels to further understand changes in student performance from baseline to endline. **Exhibit 8** below defines the various categories of readers for the grade 4 EGRA.

**Exhibit 8. Grade 4 Reading Proficiency Levels**

| Category | Definition* |
|---|---|
| Low Grade 4 Reader | Fewer than 40 cwpm |
| Emergent Grade 4 Reader | 40–55 cwpm and reading comprehension 60% or above. |
| Proficient Grade 4 Reader | 55–70 cwpm and reading comprehension 70% or above |
| Fluent Grade 4 Reader | 70 or more cwpm and reading comprehension 70% or above |

*Definitions derived from performance on a grade 4 passage

The analysis of grade 4 student proficiency levels, as shown in **Exhibit 9**, indicates a large decrease (-10.0%) in the percentage of students in the low reader category, and remarkable increase of nearly18% in the percentage of students in the fluent reader category.

**Exhibit 9. Shifts in Grade 4 Reading Proficiency Levels, Baseline and Endline (Percentages)**

| Category | Adjusted Baseline | Endline | Change from Adjusted Baseline |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Low Grade 4 Reader | 19.3% | 9.3% | -10.0 |
| Emergent Grade 4 Reader | 25.9% | 29.7% | +3.8 |
| Proficient Grade 4 Reader | 39.8% | 28.2% | -11.6 |
| Fluent Grade 4 Reader | 15.0% | 32.9% | +17.9 |

A further breakdown of grade 4 student performance in the form of score distributions at endline is presented in **Annex F**. A little over 50% of students scored 40–80 cwpm on both nonwords and oral reading subtasks. As with grade 2, the relationship between nonwords and reading connected words was as expected. Skill in reading unfamiliar words (i.e., nonwords) contributed to the ability to read connected text accurately and automatically. On the silent reading comprehension subtask, the distribution was more spread out with 78% students scoring 50%–100%.

As we did for grade 2, we only generated the item level analysis for the silent reading comprehension subtask at endline in grade 4 (**Annex G**) because of the many items in other subtasks. Findings show that 50% or more of students were able to answer 90% of the questions accurately. Only one question (question 9) on the silent reading comprehension subtask had below 50% accuracy. Most of the questions were explicit questions. They could be answered directly from the text or by close word matching. Question 9 was challenging for most of the students as it required connecting pieces of information from across the story and evaluating this information. This is a higher-order comprehension skill that many students were lacking.

Overall, for grade 4, we saw improvement in students' reading ability across the three subtasks at endline compared to adjusted baseline scores. The change in scores from baseline to endline was also significant, hence confirming the Program's success in improving grade 4 reading.

## 3.2 GRADE 2 AND GRADE 4 EGMA FINDINGS

### 3.2.1 Grade 2 EGMA Findings

The Program administered six EGMA subtasks in the grade 2 EGMA: missing number, word problems, addition, subtraction, relational reasoning, and spatial thinking. On average at endline, students scored 74.6% correct on the missing number subtask, 64.8% correct on the word problems subtask, 76.6% correct on the addition subtask, 66% correct on the subtraction subtask, 53.9% correct on relational reasoning, and 64% correct on spatial thinking. **Exhibit 10** below shows the change between endline average and adjusted baseline average scores. The most significant improvement was in the missing number subtask (+8.1%), while the greatest decline was in word problems subtask (-7.4%), followed by the subtraction subtask (-6.1%). The change from baseline to endline was not statistically significant in three of the six subtasks—addition, relational reasoning, and spatial thinking.

**Exhibit 10. Average Grade 2 Mathematics Achievement by Subtask, Baseline and Endline**

| Grade 2 | Adjusted Baseline | Endline Average | Change from Adjusted Baseline |
|---|---|---|---|
| Missing number (% correct) | 66.5 | 74.6 | +8.1*** |
| Word problems (% correct) | 72.2 | 64.8 | -7.4*** |
| Addition (% correct) | 79.9 | 76.6 | -3.3 |
| Subtraction (% average score) | 72.1 | 66 | -6.1* |
| Relational reasoning (% correct) | 59.0 | 53.9 | -5.1 |

| | | | |
|---|---|---|---|
| Spatial thinking (% correct) | 62.5 | 64 | +1.5 |

*** $p<0.001$, ** $p<0.01$, * $p<0.05$

A further breakdown of grade 2 student performance in mathematics in the form of score distributions at endline is presented in **Annex D**. One interesting finding from the score distributions that is worth highlighting is that over 50% of students scored above 80% correct on the missing numbers subtask, whereas for higher-order subtasks, most students did not score as high. In fact, on the relational reasoning subtask, the score distribution was even: out of a total of 5 points, students were relatively evenly spread across 1–5, with 15% scoring 1 out of 5 and 16% scoring 5 out of 5. On the word problems subtask, 62% of students scored a 4, 5, or 6 out of a total of 6 questions.

The item analyses by subtask at endline are presented in **Annex E.** On missing numbers, students did well with obvious patterns, but struggled when the pattern was not as readily apparent (i.e., 3, 8, __, 18). Similarly, for the subtraction subtask, more than twice as many students incorrectly solved a two-digit item that required borrowing compared to one that did not require it. For word problems, one item that required applying division to equally distribute sweets among a group of children was incorrectly solved by 60% of students. For the relational reasoning, students scored high on items for which they had to calculate the value of an expression presented in a familiar way and the items for which they had to complete the decomposition of a number. However, students scored low on equivalence expressions in which relationships were presented and the students had to determine the missing number needed to satisfy the relationship (i.e., ☐+26=27+27). Regarding spatial thinking, students did well on items that required simple visualization, but scored low on items that required intricate visualization (e.g., where cubes were hidden from direct view meaning students had to count the cubes they could see and mentally manipulate the pictures to know how many cubes were hidden from their view).

The decline in student performance from baseline to endline on word problems and subtraction merits reflection on the factors that may have contributed to that result. These are summarized in four key reflections below.

**Key reflection #1: In the Program's Mathematics course, the curriculum was overloaded, especially in grades 1 and 2, because of new content added without changing any of the previous content**.

The Program was developed to respond to the government's call for education that emphasized the 4Cs: collaboration, creativity, communication, and critical thinking. For Mathematics, this also meant creating a curriculum that would allow students to successfully take the Third International Mathematics and Science Study (TIMSS) assessment in grade 4. The development of the Program's Mathematics curriculum used the TIMSS assessment framework to determine the scope and pacing of content.

In grades 1–4, Program materials continued Uzbekistan's prior focus on instruction that was fast paced, with students reaching fluency in addition and subtraction basic facts by the end of grade 1, and multiplication basic facts memorized by the end of grade 2. Much of the prior content was focused on getting students to become fluent in these facts, and most of the time in grades 1–2 was spent on students performing complex operations accurately. However, because of the alignment with TIMSS, the Program materials also added in new concepts for students to learn (e.g., fractions, statistics and data analysis, algebraic concepts such as functions, and geometry). These concepts were either absent or only lightly addressed in Uzbekistan's prior materials.

The new additions to the curriculum as well as the prior high-level of standards led to teachers feeling like the Program materials did not offer enough time to build mathematics expertise and understanding in grades 1–2 during the pilot. To address this, the Program conducted two distinct actions:

1. We created workbooks for grades 1–4, which gave students extra practice on key skills such as operations and memorization of basic facts. This was meant to supplement the Program STB textbook and give teachers opportunities to help students continue to practice core skills that they did not have enough time for during regular lessons.
2. We held a consultation meeting in January 2023 with representatives of the Math Institute, government officials, and teachers and curriculum writers. We explained the conundrum of adding in new content while keeping current expectations, and the overloading of the curriculum, and we asked for advice. Together, this group decided that future Program materials would change some grade-level standards (for example, instead of being fluent with multiplication facts with multipliers 1-10 by the end of grade 2, it was changed to be: fluency with multiplication with multipliers 1-5 by the end of grade 2, and fluency with multipliers 1-10 by the end of grade 3). These types of changes were documented and would be made in future iterations of the Program STBs.

**Key reflection #2: Building the 4Cs is a time intensive endeavor, and contrary to existing classroom practices. It is time intensive in two ways:**

1. Most of the allocated 45 minutes for Mathematics lessons was dedicated to classroom discussions and building critical thinking skills. Necessarily, this took time away from memorization and practice of rote skills.
2. A shift to critical thinking and the other 4Cs takes time. Teachers need practice, and in the short term, it may be that a program needs more than just 1 year of piloting to better see any shifts in teachers' ability to promote, and students' ability to acquire, higher-order skills.

**Key Reflection #3: The assessment for grade 2 was not designed to measure the type of instruction that the Program was emphasizing.**

The EGMA for grade 2 is designed to measure the foundational learning skills that most children should know by the end of grade 3 and is targeted to students in low- and middle-income countries. In Uzbekistan, because of high levels of existing learning, only the more advanced subtasks were used (such as addition and subtraction Level 2, not Level 1). In fact, most of the skills measured by the EGMA used in Uzbekistan for grade 2 are above minimum proficiency levels on the Global Proficiency Framework. This assessment was adapted to Uzbekistan but not designed to measure Program improvement, as the Program was aimed not at developing foundational skills but instead moving instruction to be aligned with the 4Cs. For example, although fractions were introduced in grade 1 to Program students, there was no assessment of fractions on the grade 2 EGMA.

In contrast, the grade 4 assessment was aligned to the TIMSS and the 4Cs and designed specifically for the context of Uzbekistan.

**Key Reflection #4: The Program should revisit its approach to word problems.**

In Uzbekistan, students are very familiar with creating a model of word problems every time they solve a problem. For example, given the problem "Feruza has 4 apples, and then her

friend gives her 2 more. How many apples does she have altogether?", students would be asked to recreate the problem. For the Program, curriculum writers and designers of the Mathematics program felt that this process was cumbersome and time consuming to do every single time, and a leftover from earlier eras of instruction. The Program materials also expanded into different types of problems and featured specific lessons on word problems, instead of every lesson including word problems (as prior Uzbekistan country materials did). Likely because of this shift, though, word problem scores at endline were lower than at baseline. It could be that teachers and students just need more time to learn the new methodology for word problems, or it could be that the old methodology, while time intensive, was crucial to building understanding. Before creating future iterations of the textbooks, more evidence is needed into how best to support students in Uzbekistan with word problems and how teachers are treating word problems in the classroom.

In sum, we did not see any statistically significant changes in grade 2 EGMA from baseline to endline, except in the missing numbers, word problems, and subtraction subtasks. While student performance declined in word problems and subtraction, the results can be explained by reflecting on gaps in the Program approach towards mathematics as outlined above. In contrast, endline findings suggest significant improvements in grade 2 students' performance on the missing number subtask.

### 3.2.2. Grade 4 Mathematics Findings

There were four domains assessed in the grade 4 written Mathematics assessment: number and operations, geometry, measurement, and statistics. On average at endline, students' overall score was 59% correct, whereas students' overall adjusted baseline score was 52.6% correct, showing some improvement (~6%), after Program implementation. **Exhibit 11** below shows the change between endline average and adjusted baseline average scores.

**Exhibit 11. Average Grade 4 Mathematics Achievement by Domain and Treatment, Baseline and Endline**

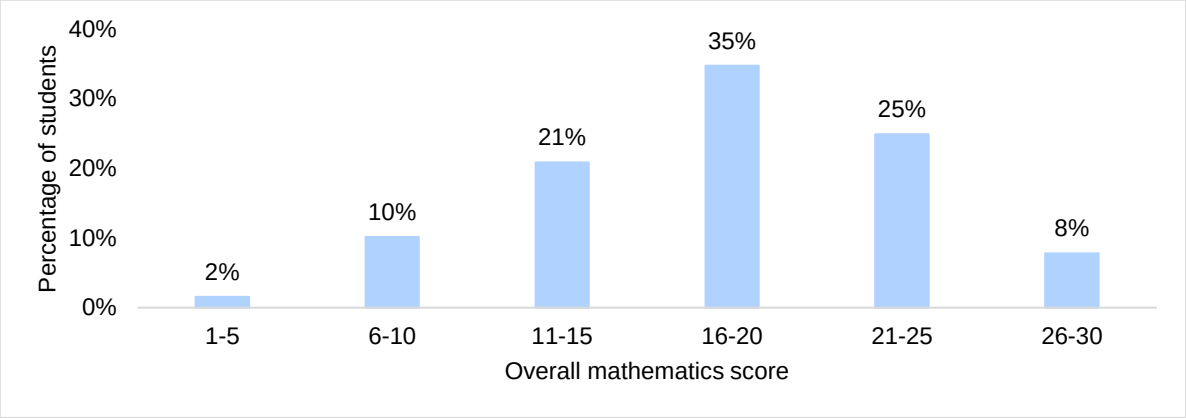
| Grade 4 | Adjusted Baseline | Endline Average | Change from Adjusted Baseline |
|---|---|---|---|
| Overall score (% correct) | 52.6 | 59 | +6.4*** |
| Number and operations (% correct) | 54.8 | 63.3 | +8.5*** |
| Geometry (% correct) | 38.5 | 43.1 | +4.6** |
| Measurement (% correct) | 50.0 | 51.3 | +1.3 |
| Statistics (% correct) | 59.3 | 63.4 | +4.1 |

*** $p<0.001$, ** $p<0.01$, * $p<0.05$

The Program targeted all domains for grade 4 Mathematics, adding content for fractions, geometry, and statistics that was not present before; hence it is very likely that the improvement in those areas is indeed due to the Program's implementation. The highest improvements were on the number and operations domain, whereas the average scores on the measurement domain remained about the same.

The overall score distribution for grade 4 mathematics is shown in **Exhibit 12** below. A further breakdown of grade 4 student performance on each domain in the form of score distributions is shown in **Annex H.**

**Exhibit 12. Overall Distribution of Mathematics Scores, Grade 4**




Over 50% of students scored below 20 out of a total of 30 questions (total questions across all the domains). As for the individual domains, students got more questions right on the number and operations domain and the statistics domain (with 50% or more students correctly answering 50% or more questions on each domain). This may be because the Program materials provided a systematic way for teachers to explain key concepts in mathematics, building proficiency through deep dives into the content.

Overall, for the grade 4 written Mathematics assessment, we generally saw an improvement in student performance from baseline to endline. The change in the overall Mathematics score was statistically significant, so there is evidence to suggest that the Program was successful in improving grade 4 mathematics abilities. The Program had more impact in grade 4 than in grade 2 because, unlike in grade 2, the grade 4 assessment tool was designed to be aligned with the TIMSS assessment framework and developed specifically for the context of Uzbekistan. The Program curriculum was also designed to align with the TIMSS framework. In the case of grade 4, therefore, the assessment measured what the Program curriculum emphasized. For example, the Program curriculum introduced fractions as a content area, and put more emphasis on geometry. Number and operations (which includes fractions) and geometry are two domains in which students improved significantly from baseline to endline. In addition, as discussed above, most of the reports about the curriculum being overloaded that the Program team received were for grade 2. When the curriculum was reviewed at the consultation meeting in January 2023, the team of experts decided that adjustments to grades 1 and 2 were needed, but that grades 3 and 4 were not overloaded and that the time allocated to the different topics in the curriculum was adequate. For this reason, the Program materials in grades 3 and 4 were a better fit for the school day hours than those for grades 1 and 2.

## 3.3 FINDINGS BY STUDENT GENDER

**Exhibit 13** presents the EGRA and EGMA results for grade 2 and 4 students by subtask and gender. In grade 2, girls outperformed boys in all EGRA tasks. The difference in performance between girls and boys in grade 2 was particularly substantial for oral reading fluency, with girls reading on average 6.5 more cwpm than boys. The gain in scores on the reading comprehension subtask from adjusted baseline to endline was higher for girls than boys. Grade 4 girls also performed significantly better than grade 4 boys on oral reading fluency, with girls reading 11 more cwpm than boys. Girls increased their reading comprehension scores in both grades and their oral reading fluency rate on the grade 4 EGRA at endline compared to the adjusted baseline; however, their oral reading fluency rate in grade 2 remained about the same. The same was true for boys, that is, their scores on

reading comprehension in both grades and grade 4 oral reading fluency improved, however grade 2 oral reading fluency remained about the same.

Grade 2 and 4 students' Mathematics performance by task and gender is also highlighted in Exhibit 12. Boys generally outperformed girls in Mathematics subtasks in both grades. For grade 2, the gains or losses in scores from adjusted baseline to endline on all subtasks were not statistically significant for boys and girls. Grade 4 estimates show that overall, boys outperformed girls, with the grade 4 boys achieving an average score of 60.3% and the girls achieving an average score of 57.8%. These scores were, respectively, 7.4% and 5.6% higher than boys' and girls' overall adjusted baseline Mathematics scores. Girls outperforming boys on reading subtasks and boys outperforming girls in Mathematics was also an apparent phenomenon at baseline.

**Annex I** displays gender disaggregated student performance across all the subtasks. The relative difference in gains from adjusted baseline to endline for each gender were negligible, with the highest relative difference being found in the grade 2 oral reading fluency subtask, grade 2 relational reasoning subtask, and grade 4 silent reading comprehension subtask.

**Exhibit 13. Mathematics and Reading Achievement, by Grade and Gender for Intervention Schools**

| Grade | Subject | Task | Gender | Adjusted Baseline Average | Endline Average | Gain |
|---|---|---|---|---|---|---|
| Grade 2 | Reading | Oral reading fluency (cwpm) | boys | 34.8 | 37.2 | +2.4 |
| | | | girls | 44.7 | 43.7 | -1.0 |
| | | Reading comprehension (percent score) | boys | 60.3 | 67.3 | +6.9** |
| | | | girls | 63.0 | 70.7 | +7.7*** |
| | Mathematics | Relational reasoning (percent score) | boys | 60.1 | 56.5 | -3.6 |
| | | | girls | 58.0 | 51.4 | -6.6 |
| | | 3D spatial thinking (percent score) | boys | 66.1 | 67.2 | +1.1 |
| | | | girls | 58.8 | 60.7 | +1.9 |
| Grade 4 | Reading | Oral reading fluency (cwpm) | boys | 53.0 | 65 | +12*** |
| | | | girls | 64.8 | 76 | +11.2*** |
| | | Silent reading comprehension (percent score) | boys | 58.0 | 64.4 | +6.4** |
| | | | girls | 55.6 | 65.4 | +9.8*** |
| | Mathematics | Overall mathematics (percent score) | boys | 52.9 | 60.3 | +7.4*** |
| | | | girls | 52.2 | 57.8 | +5.6** |

*** $p<0.001$, ** $p<0.01$, * $p<0.05$

## 3.4 FINDINGS BY STUDENT SOCIOECONOMIC STATUS

Although students' socioeconomic status (SES) was not a focus of Program interventions, we asked students questions at the end of the EGRA/EGMA that we used to construct the SES index to investigate the relationship between students' SES and learning achievement. The questions that were asked and how we created the SES index are presented in **Annex B. Exhibit 14** below displays averages by SES tertile for each grade 2 and grade 4 subtask. Overall, students across the three SES tertiles scored more or less similarly on average in both grades across all reading and mathematics subtasks. This implies that SES did not

have significant association with reading and mathematics achievement in grades 2 and 4 in the Uzbekistan context.

**Exhibit 14. Average Scores by SES Tertile, Grade, and Subtask**

| | SES 1 | SES 2 | SES 3 |
|---|---|---|---|
| **Grade 2** | | | |
| **Reading** | | | |
| Oral reading fluency (cwpm) | 38.4 | 40.8 | 42.5 |
| Nonwords (cwpm) | 33.6 | 35.6 | 37.3 |
| Reading comprehension (average score out of 5) | 3.3 | 3.5 | 3.6 |
| **Mathematics** | | | |
| Missing number (average score out of 10) | 7.2 | 7.6 | 7.5 |
| Word problems (average score out of 6) | 3.8 | 4.0 | 4.0 |
| Addition (average score out of 5) | 3.7 | 4.0 | 4.0 |
| Subtraction (average score out of 5) | 3.1 | 3.5 | 3.3 |
| Relational reasoning (average score out of 5) | 2.5 | 2.8 | 2.8 |
| Spatial thinking (average score out of 4) | 2.5 | 2.6 | 2.7 |
| **Grade 4** | | | |
| **Reading** | | | |
| Oral reading fluency (cwpm) | 68.4 | 72.2 | 70.6 |
| Nonwords (cwpm) | 45.6 | 48.1 | 49.6 |
| Silent reading comprehension (average score out of 10) | 6.4 | 6.6 | 6.5 |
| **Mathematics** | | | |
| Numbers and operations (average score) | 11.0 | 11.4 | 11.9 |
| Geometry (average score) | 1.6 | 1.7 | 1.8 |
| Measurement (average score) | 1.8 | 2.1 | 2.3 |
| Statistics (average score) | 2.4 | 2.6 | 2.7 |
| Overall score (average) | 16.8 | 17.8 | 18.6 |

# SECTION 4: CONCLUSIONS AND RECOMMENDATIONS

This section presents conclusions and recommendations based on the EGRA and EGMA findings.

## 4.1 EGRA

The results of the Program endline EGRA show improvements in students' reading achievement across all grade 2 and 4 subtasks. A comparison of the adjusted baseline to the endline average scores showed that the gain in grade 2 oral reading comprehension scores was statistically significant, suggesting that the Program successfully improved this component of reading. This substantial increase in reading comprehension may be a result of improvement in students' word recognition ability and reading rate. In grade 4, the gains were significant across all subtasks, thus confirming Program success in improving grade 4 reading.

Reading achievement by gender indicates that in grade 2 and 4, girls outperformed boys on all EGRA tasks, and the difference in performance was greater for oral reading fluency. Girls outperforming boys on reading subtasks was also an apparent phenomenon at baseline. In terms of changes, results show almost similar gains or losses on all grade 2 and 4 EGRA subtasks for boys and girls from the adjusted baseline to endline. The gains for both girls and boys in grade 2 were not statistically significant, except for those in oral reading comprehension. However, grade 4 gains were statistically significant for boys and girls on all subtasks.

Results by student SES show that overall, students across three SES tertiles scored similarly on average in both grades on all reading subtasks. These results suggest that SES did not have significant association with reading achievement in grades 2 and 4 in the Uzbekistan context.

### 4.1.1 Recommendations based on EGRA results

The integration of Uzbek grammar and literature into a single subject, Uzbek Language Arts, for grades 1 to 4 has provided an opportunity to expand the curriculum content toward a more diverse range of texts with a varied vocabulary, and instructional methodologies that focus on enhancing communication and creativity. These shifts in approach and content take time for teachers to perfect.

- Teachers should continue to implement student-centered strategies. These evidence-based instructional approaches support student acquisition of reading and higher-order skills (e.g., fluency, comprehension) as students advance to higher grades.
- Reading with fluency means that students are reading with speed, accuracy, and understanding. Teachers should continue to build on the speed-reading tradition with greater, continued attention to students' accuracy and understanding. Teachers should integrate techniques to improve reading comprehension (e.g., questioning, visualization, predicting, reciprocal teaching). Teachers should ask more reading comprehension questions and teach students strategies for working with texts, help them better understand the difference between open and inferential questions, and develop strategies for working with both types of questions.

- Teachers should continue to develop nonword reading activities (e.g., jigsaw word reading,[4] jumbled words,[5] crosswords with nonwords[6]) and include nonword decoding activities in their lessons.
- Vocabulary and specifically academic vocabulary, and skills and strategies for complex vocabulary comprehension, are critical for success in upper grades. Future efforts may include adapting digital early grade reading games like Feed the Monster, Antura and the Letters, onebillion, or GraphoGame for Uzbek language phonics aligned with the government's Digital Nation policy.
- Effective instruction should be complemented with appropriate supplementary reading materials.

## 4.2 EGMA

Student achievement on the grade 2 EGMA increased for only two subtasks: missing number and spatial thinking. A comparison of the adjusted baseline to the endline average scores showed that the gain in missing numbers was statistically significant, while the gain in spatial thinking was not. Student performance declined on all other subtasks (i.e., word problems, addition, subtraction, and relational reasoning) with the greatest decline on the word problems subtask followed by the subtraction subtask. The low performance on word problems and subtraction may be a result of a number of factors, including (1) an overloading of the curriculum, especially in grades 1 and 2, due to addition of new content without removing previous content; (2) a shift to critical thinking and the other 4Cs that requires time to develop in class, and is contrary to existing classroom practices; (3) the design of the grade 2 EGMA, which did not measure the type of instruction that the Program was emphasizing; and (4) a shift away from traditional methods of solving word problems.

Results of the grade 4 written mathematics assessment show increases in student achievement from adjusted baseline to endline on all subtasks. The change in the overall mathematics score was statistically significant, implying that the Program was successful in improving grade 4 mathematics abilities.

Like the baseline, endline mathematics results by gender indicate that boys generally outperformed girls on grade 2 and 4 mathematics subtasks. In terms of gains, results show that gains or losses on all grade 2 subtasks were not statistically significant for boys and girls. However, for grade 4, the overall gains for both girls and boys were significant.

Results by student SES show that overall, students across three SES tertiles scored similarly on average in both grades on all mathematics subtasks. These results suggest that SES did not have significant association with mathematics achievement in grades 2 and 4 in the Uzbekistan context.

### 4.2.1 Recommendations based on EGMA results

Shifting instruction from more traditional, rote ways of teaching toward instruction that is aligned to the 4Cs (creativity, communication, critical thinking, and collaboration) is difficult and may take time. Teachers may not be used to teaching in this manner.

4 Jigsaw Word Reading: a set of words broken into syllables that students need to reassemble to figure out all possible words.
5 Jumbled words: a mixed set of letters that students need to use to restore the original word.
6 Crosswords with nonwords: creating crosswords puzzles using nonsensical or pseudowords.

In addition to shifts in instruction, preparing students to take international assessments such as the TIMSS assessment is challenging and requires sustained effort over years to include new skills and domains. To do this, teachers need support in:

- Using mathematics manipulatives (e.g., sticks, number line, multiplication charts), which make numbers less abstract, in all grades.
- Connecting mathematics concepts to students' daily lives to help students see the mathematics that is all around them.
- Using explanation and justification to ensure that students are understanding math concepts.
- Discussing incorrect answers with students.
- Encouraging problem solving through complex, multi-step problems.

TPD activities should emphasize training teachers on strategies to support student development of the 4Cs across various domains.

Future iterations of the STBs and TGs should carefully review the scope of the curricula in grades 1 and 2 and ensure that only the key competencies are included. Efforts should be made to reduce the breadth of topics in each grade's curriculum and focus more on depth in each topic. In addition, students need time to develop new skills over the course of primary school, and interventions should attempt to better understand how students develop skills from grades 1–4. Future research should provide more evidence to inform decision-making on these topics, such as, what sequence of content is most appropriate for students in grades 1-4 in Uzbekistan? Which topics should be emphasized? What types of methodologies support student understanding of word problems? This evidence can provide the Government of Uzbekistan with a clear way forward to support student learning in math in primary school.

# ANNEXES

## ANNEX A: Methodology for adjusting time difference between baseline and endline assessments

In order to account for the time difference between when during the school year data were collected at baseline and when they were collected at endline, we decided to adjust the baseline data by multiplying each individual student's scores at baseline by a ratio of days that a student from the respective grade was in school at endline to days a student from the respective grade was in school at baseline. The ratio was calculated to be 83% for grade 2 (0.827) and 91% for grade 4 (0.906).

**To calculate adjusted score:**

- Assume 165 days of school in years 2017–2018, 2018–2019, 2019–2020, 2020–2021
- 170 days of school in year 2021–2022 (for baseline we take only 63 days for year 2021–2022)
- 165 days of school in year 2022–2023 (for endline we count 155 days to account for when data collection ended)

**Step 1.** Sum the total days of school for baseline (grade 1 and 2 for grade 2 students, and grades 1, 2, 3, and 4 for grade 4 students).

- Total for grade 2 would be 165+165+63 = 393
- Total for grade 4 would be 165+165+165+165+63 = 723

**Step 2.** Sum the total days of school for endline.

- Total for grade 2 would be 170+155 = 325 days
- Total for grade 4 would be 165+165+170+155 = 655 days

**Step 3.** Divide endline by baseline to get fraction.

- grade 2 = 325 days /393 days = 0.827
- grade 4 = 655 days /723 days = 0.906

**Step 4.** Apply that fraction to baseline student level data to get adjusted student scores.

- For example: reading fluency x 0.827 = adjusted reading fluency.

While this adjustment methodology attempts to mitigate the inflation in scores at baseline due to the time difference, there are still some limitations to this approach. One limitation was that using this methodology we could not accurately account for zero scores. This is because a student scoring above zero at the original baseline would always show up in the data as scoring above zero even after the adjustment; for example, a student reading as low as 5 cwpm in the original baseline data would still show up as reading as 5 x 0.827 cwpm (in case of grade 2) which is above zero and hence the student would not be captured in the zero score category. Though in reality, it is very likely that such a student may have picked up some basic reading in the 63 additional days that they were in school, so if data collection were to have taken place 63 days (about 2 months) prior at baseline, it is possible that they would have scored a zero.

## ANNEX B: Creation of Student Socioeconomic Status index

The socioeconomic status (SES) index was created using a set of questions below identified in previous successful indices as well as new questions more specific to the context of Uzbekistan. The index was used to create tertiles, which allow us to compare student performance based on which tertile their family falls into.

**Exhibit B-1. Socioeconomic Status Index**

| Question | |
|---|---|
| Q1 | Did you go to kindergarten before school? |
| Q2 | How many people live in your house (how many live in a family)? |
| Q3 | What is the main source of drinking water in your home? (Where do you get drinking water?) |
| *Q4 | Do you have simple or automatic washing machines in your house? |
| *Q5 | Do you have bikes, cars, trucks, microbuses, motorcycles, scooters, mopeds, motor bicycles, or cars in your family? |
| Q6 | Is there any personal property in your family (at home)? |
| Q7 | Have there been sheep, lambs, goats, cows, calves, bulls, bulls, horses, or donkeys in your family since last spring (irrespective of age)? |
| *Q8 | How will your family warm the house during the winter? |
| *Q9 | Does anyone in your house use smartphones (phones without buttons)? |
| *Q10 | How many people in your home use smartphones? |
| *Q11 | Is there an Internet (WiFi) connection (network) in your home? |
| *Q12 | Have you been to Tashkent since last spring (in the last year)? |
| *Q13 | If you have gone, how many times have you traveled to Tashkent since last spring (in the last year)? |
| *Q14 | Have you been to the central city of the province in the last 6 months (since the beginning of the year)? |
| *Q15 | If you went, how many times have you traveled to the central city of the province (since the beginning of the year) in the last 6 months? |

*represents the questions that were included in the SES index based on factor analysis

We created an SES index from the questions above. Not all these questions were successful in qualifying to be part of the index. We ran a factor analysis on all the SES questions and identified the questions that had a factor of around or greater than 0.3. These questions are identified in the above table with an asterisk.

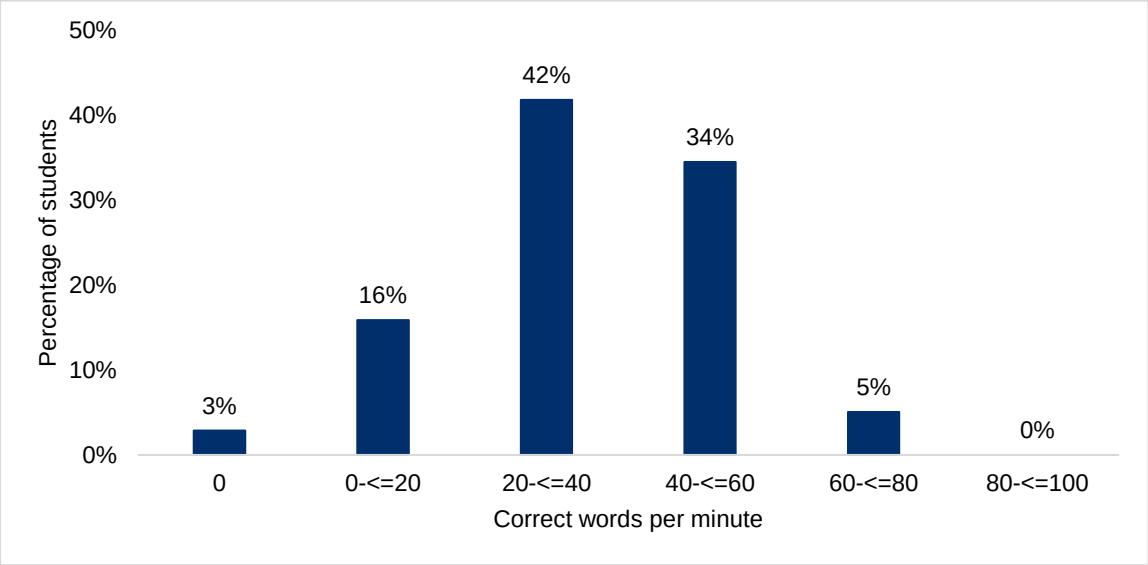
Once these questions were identified, we then generated an overall index for each student by multiplying each question's factor with each question's value in the student level dataset, and then adding the products.

- index = factor1*variable1 + factor2*variable2…

And finally, we created SES tertiles from this overall index.

## ANNEX C: Grade 2 Score Distributions by EGRA Subtask

### Exhibit C-1. Nonword Score Distribution, Grade 2




### Exhibit C-2. Oral Reading Fluency Score Distribution, Grade 2

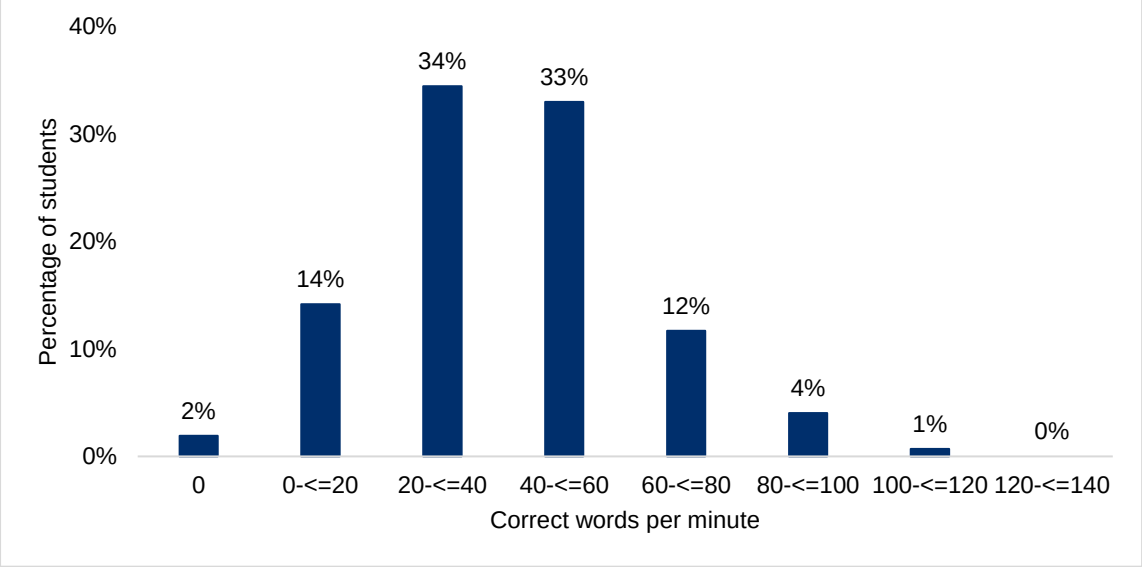
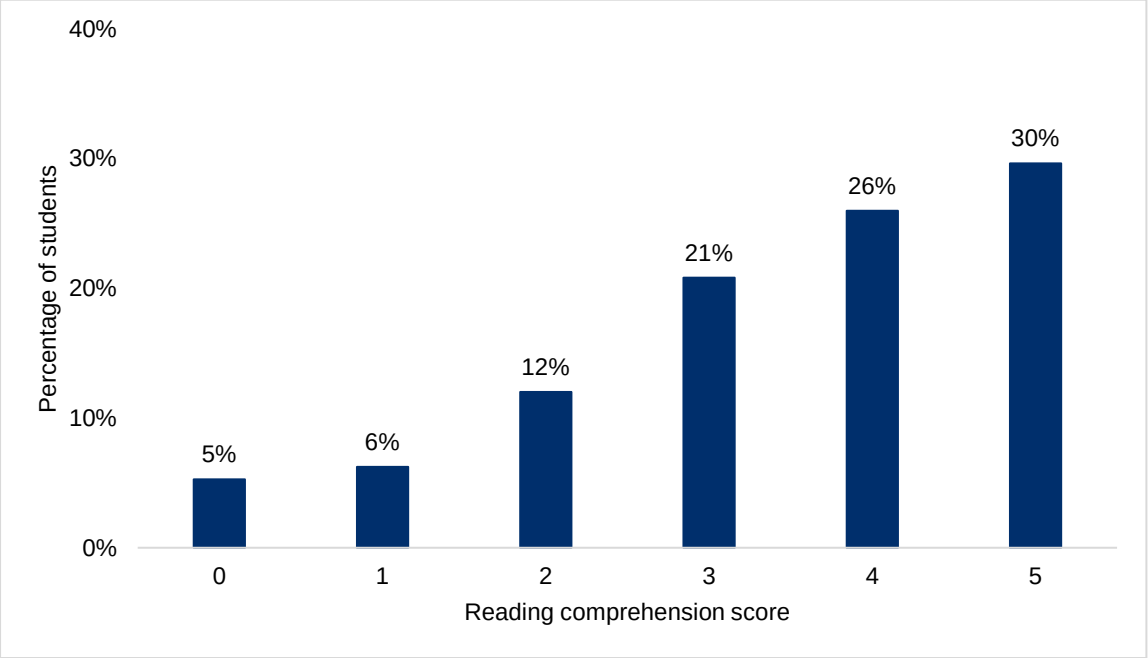
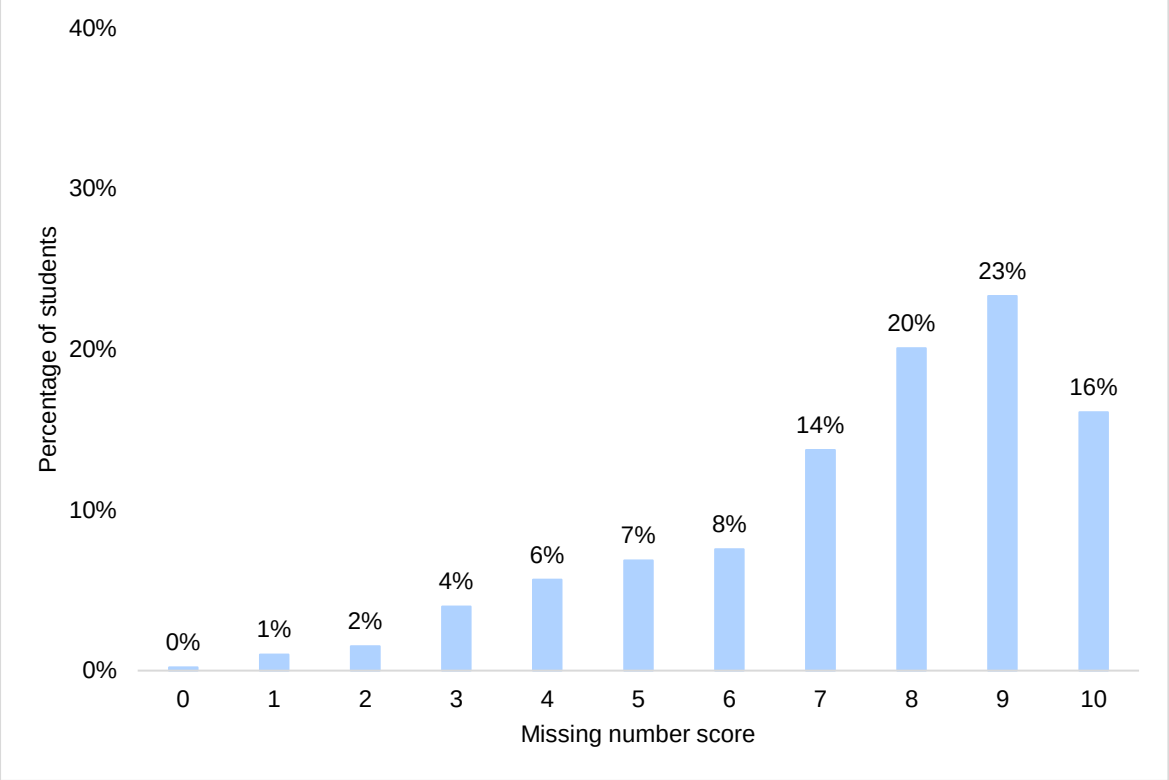


**Exhibit C-3. Reading Comprehension Score Distribution, Grade 2**

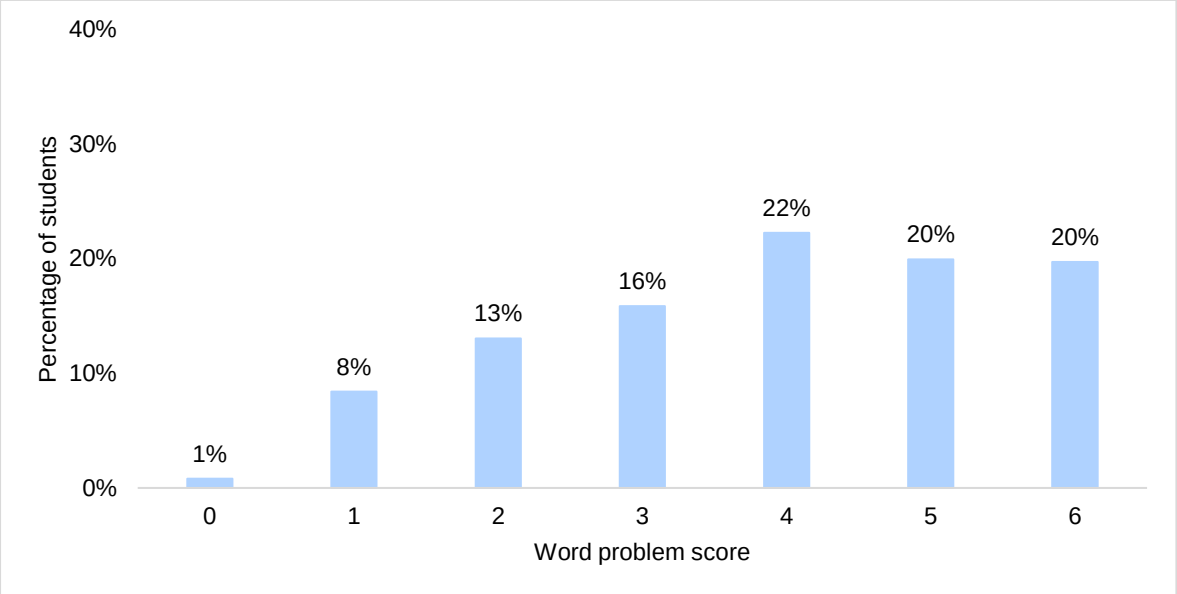


## ANNEX D: Grade 2 Score Distributions by EGMA Subtask

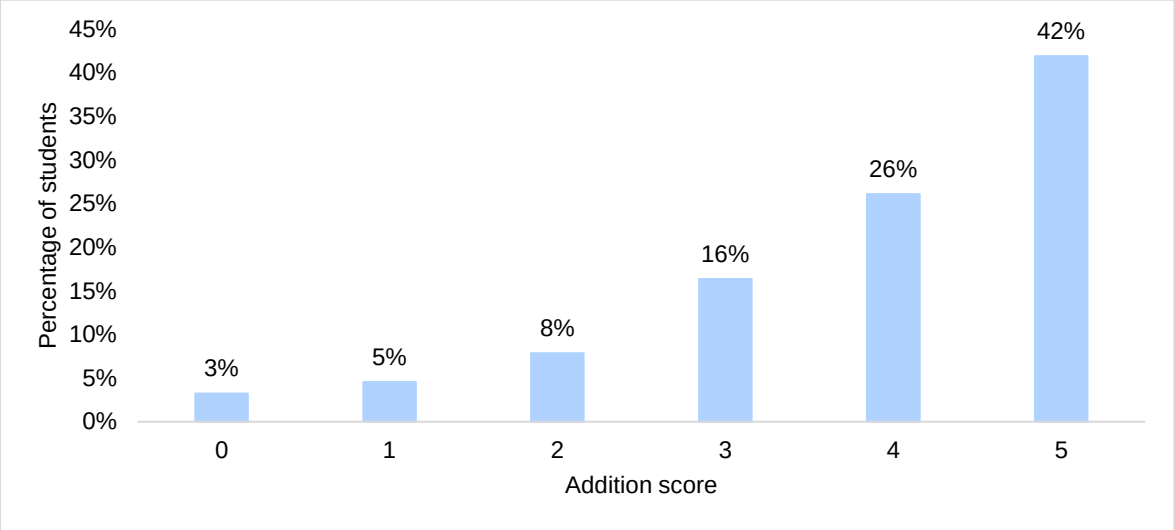
### Exhibit D-1. Missing Numbers Score Distribution, Grade 2



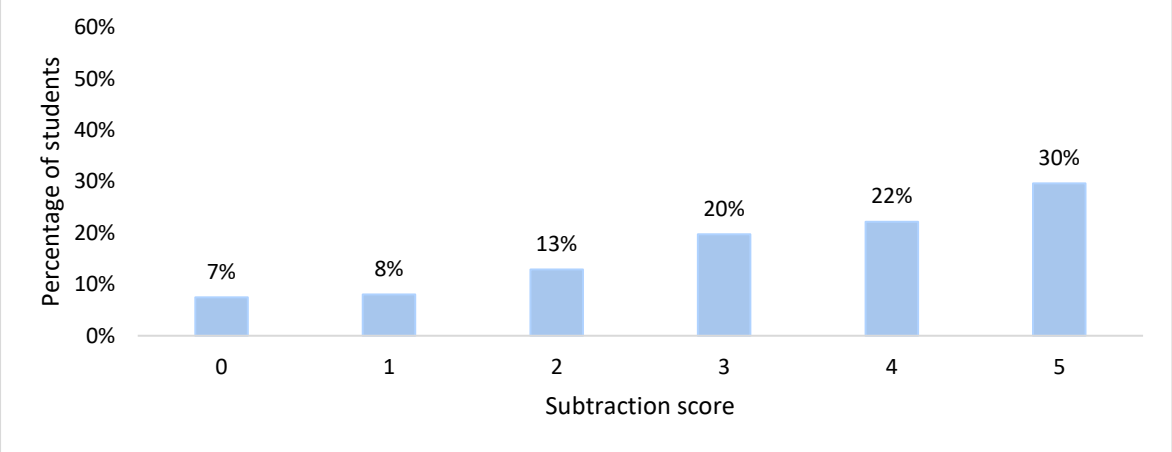

### Exhibit D-2. Word Problems Score Distribution, Grade 2



## Exhibit D-3. Addition Score Distribution, Grade 2




## Exhibit D-4. Subtraction Score Distribution, Grade 2




## Exhibit D-5. Relational Reasoning Score Distribution, Grade 2

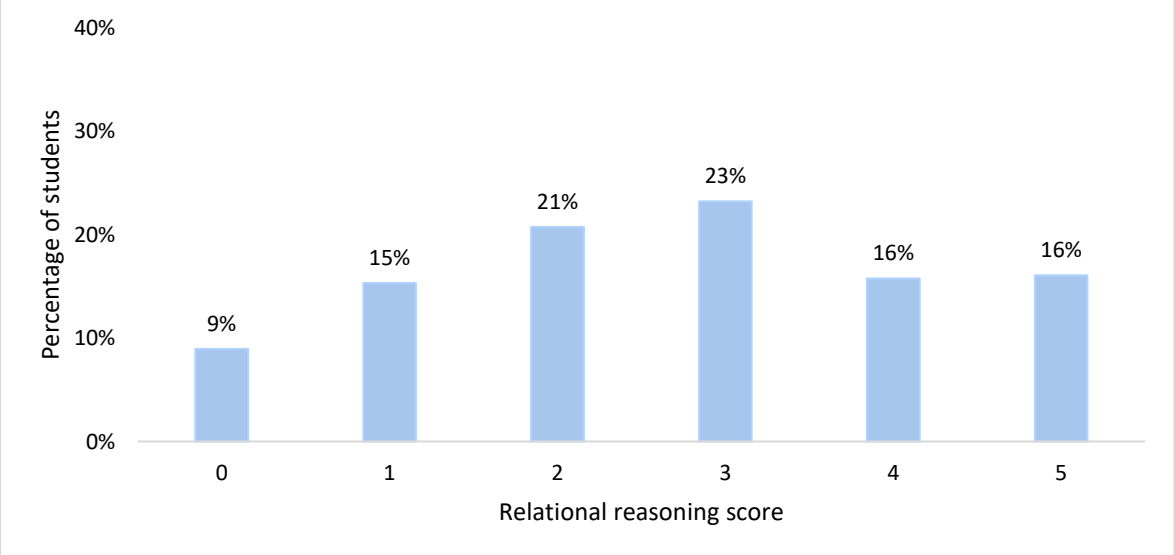
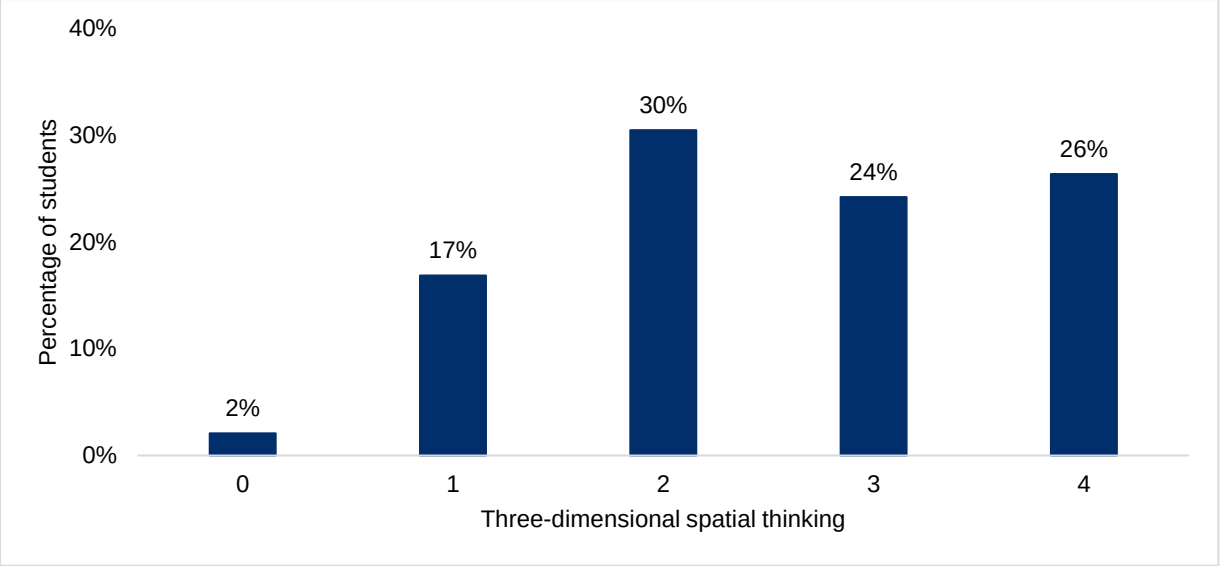
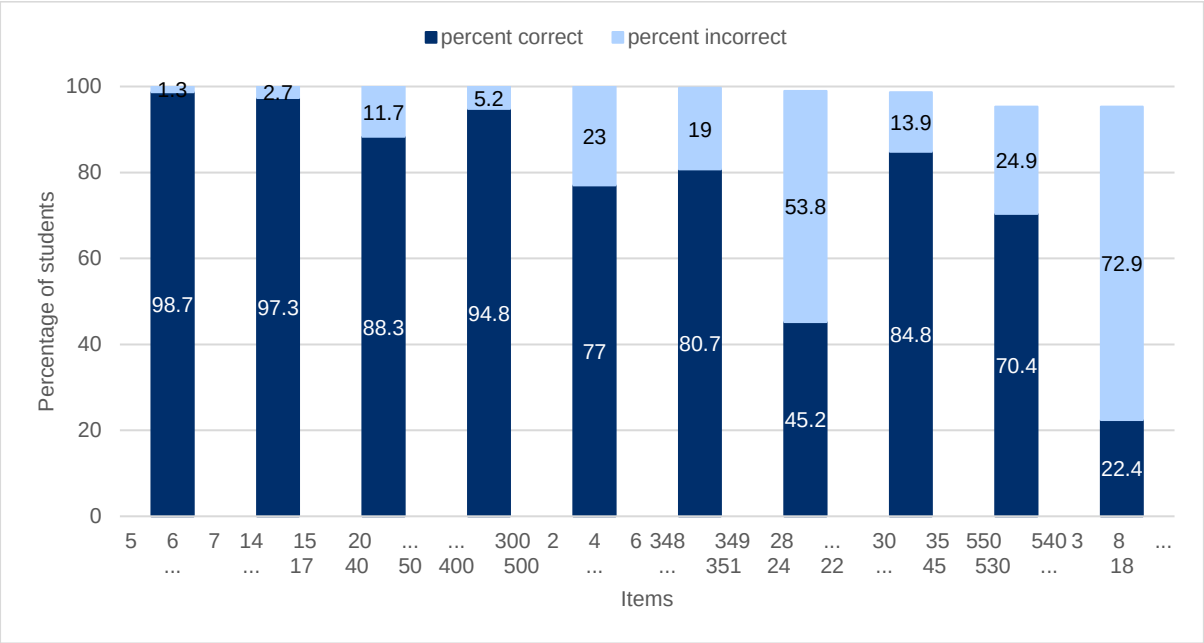


**Exhibit D-6. Spatial Thinking Score Distribution, Grade 2**

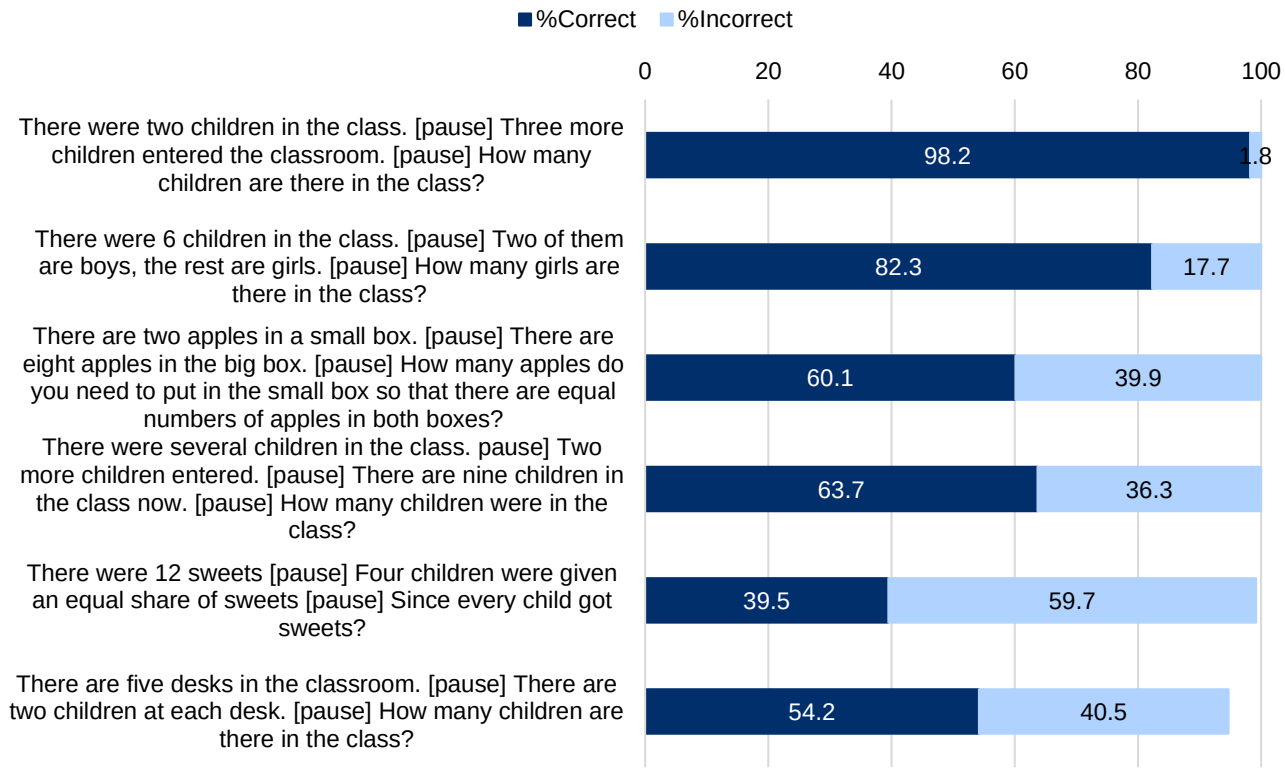


## ANNEX E: Grade 2 Item Analysis By EGMA Subtask

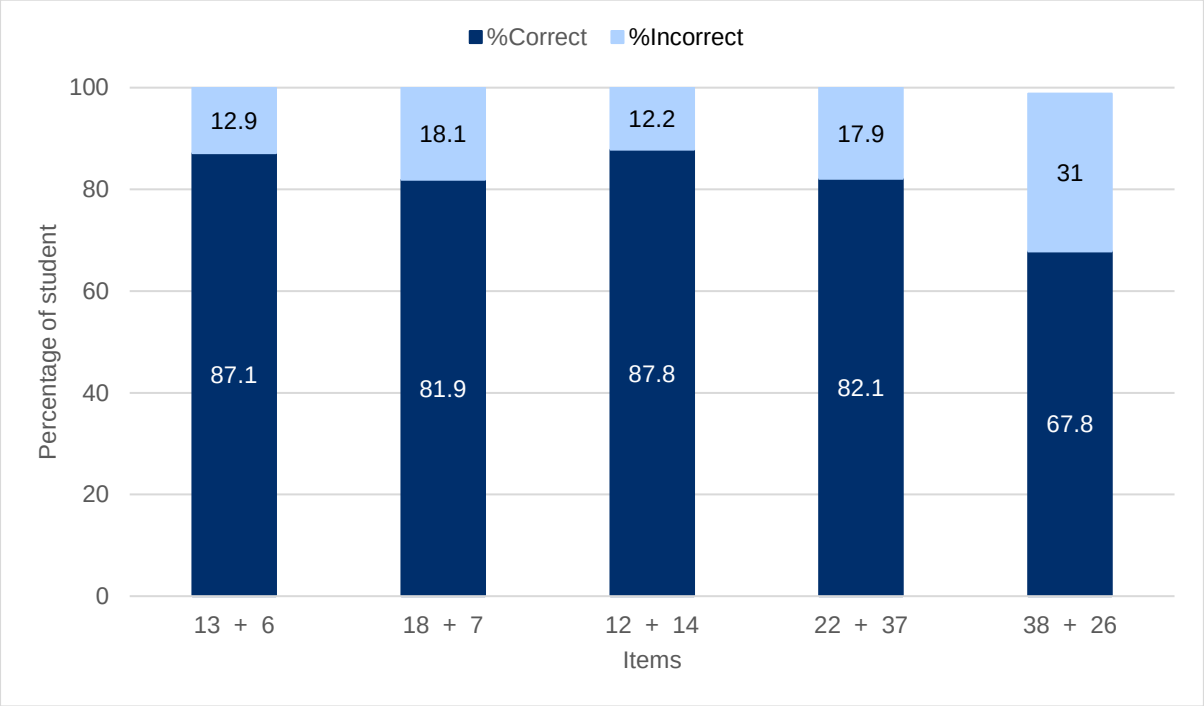
### Exhibit E-1. Missing Number Scores by Item, Grade 2



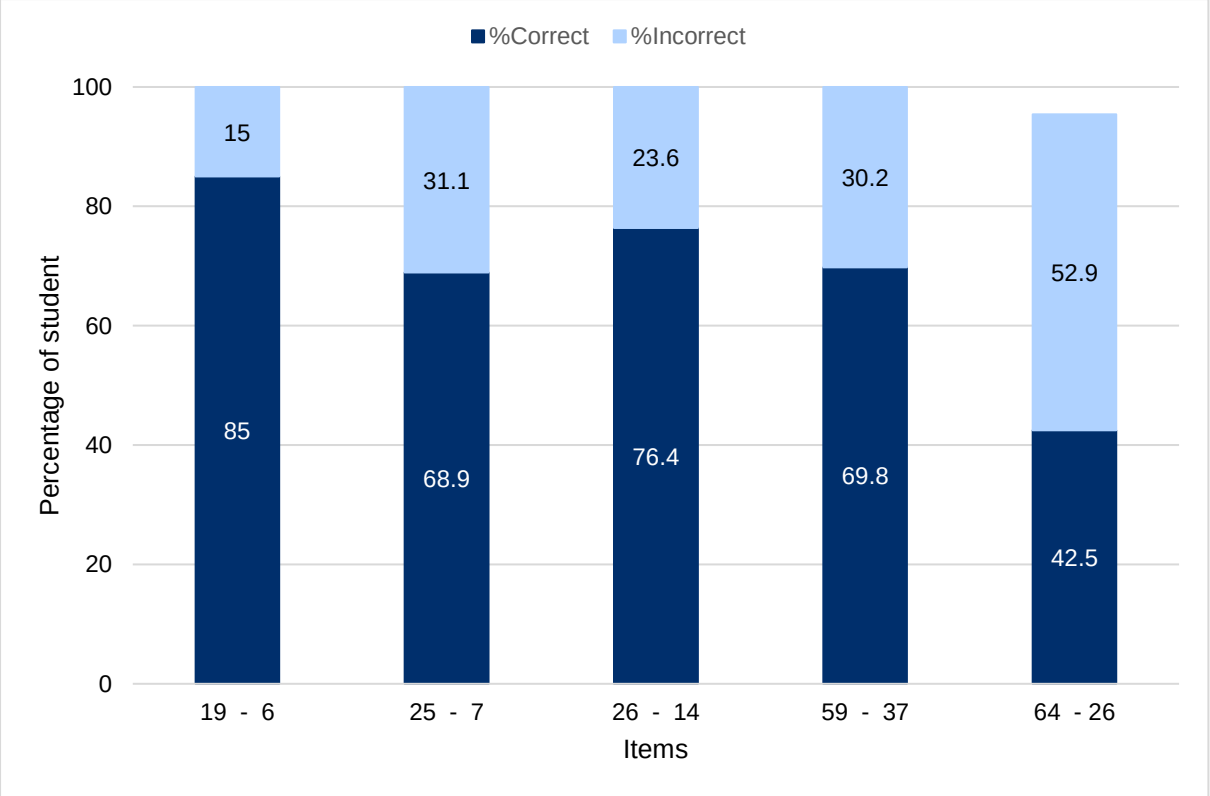

### Exhibit E-2. Word Problems Scores by Item, Grade 2



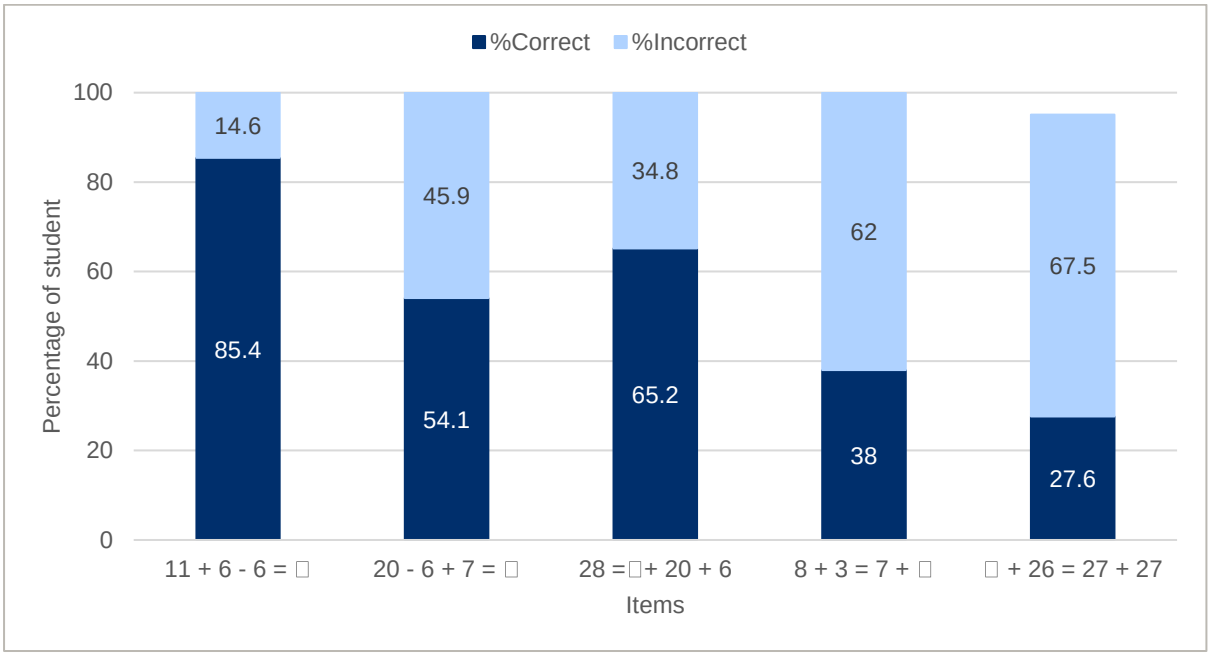
## Exhibit E-3. Addition Scores by Item, Grade 2



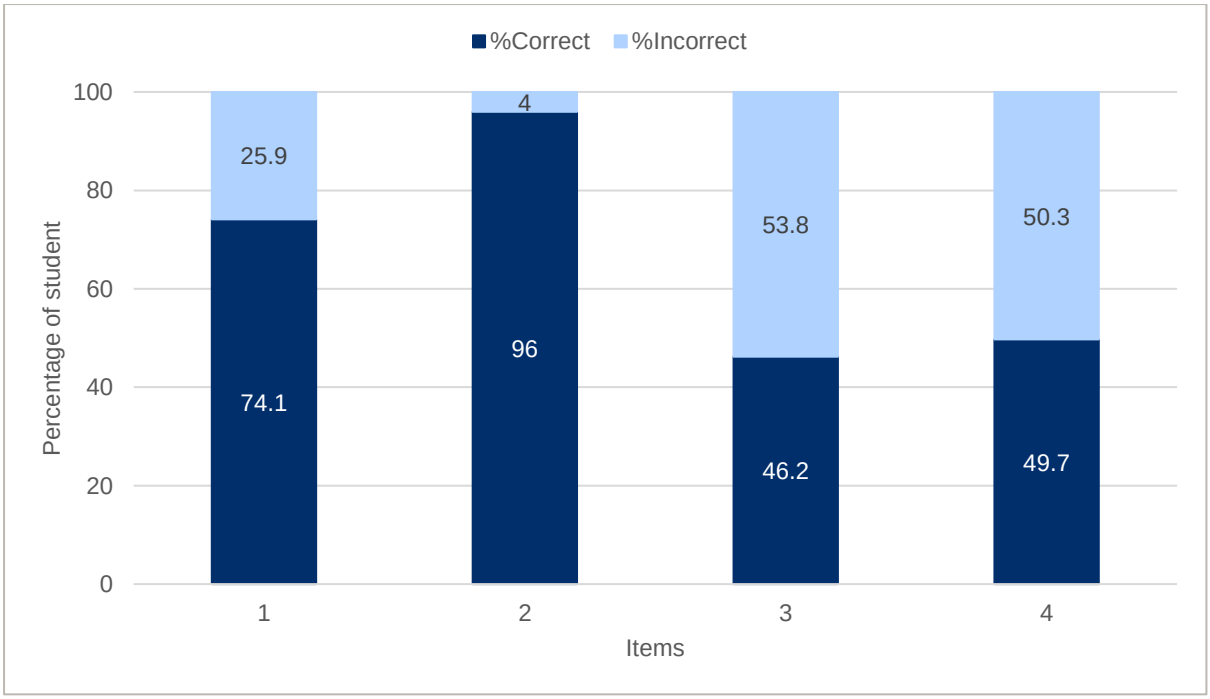

## Exhibit E-4. Subtraction Scores by Item, Grade 2



## Exhibit E-5. Relational Reasoning Scores by Item, Grade 2

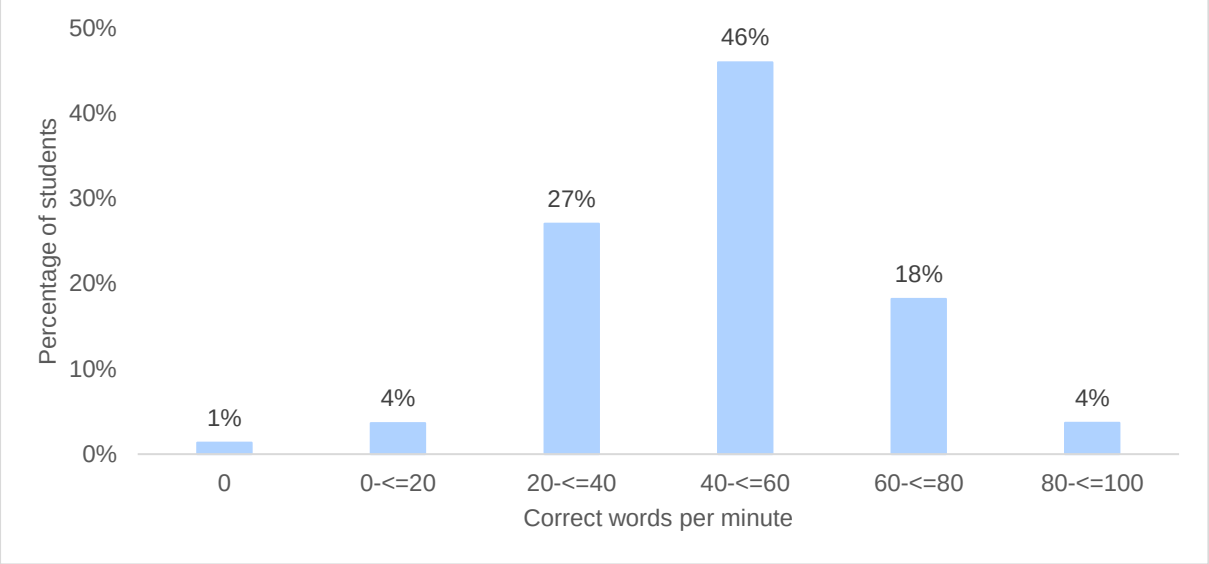



## Exhibit E-6. Spatial Thinking Scores by Item, Grade 2



## ANNEX F: Grade 4 Score Distributions By EGRA Subtask

### Exhibit F-1. Nonword Score Distribution, Grade 4




### Exhibit F-2. Oral Reading Fluency Score Distribution, Grade 4

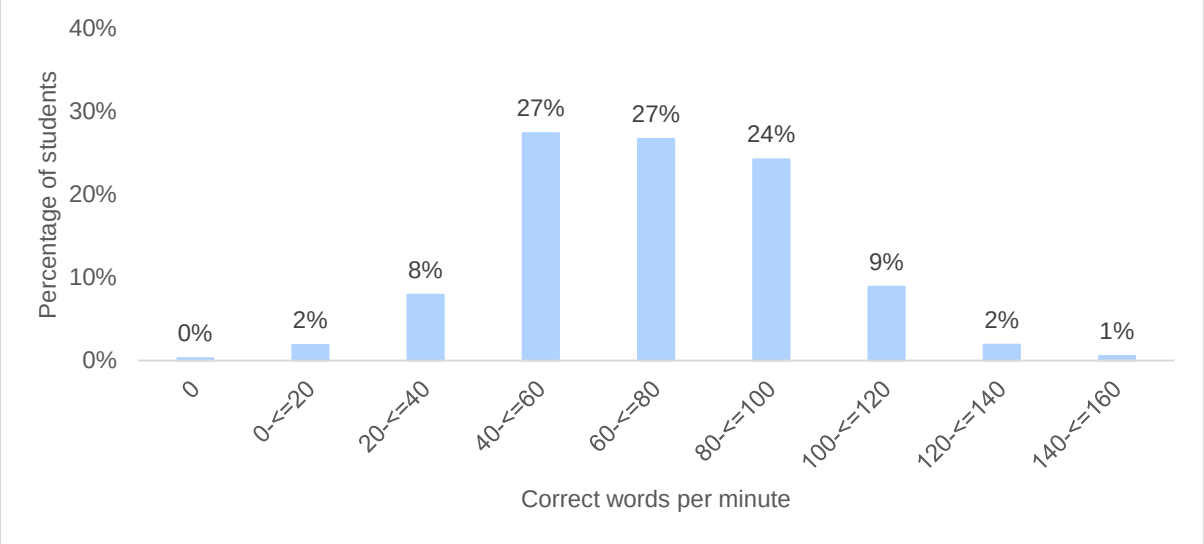
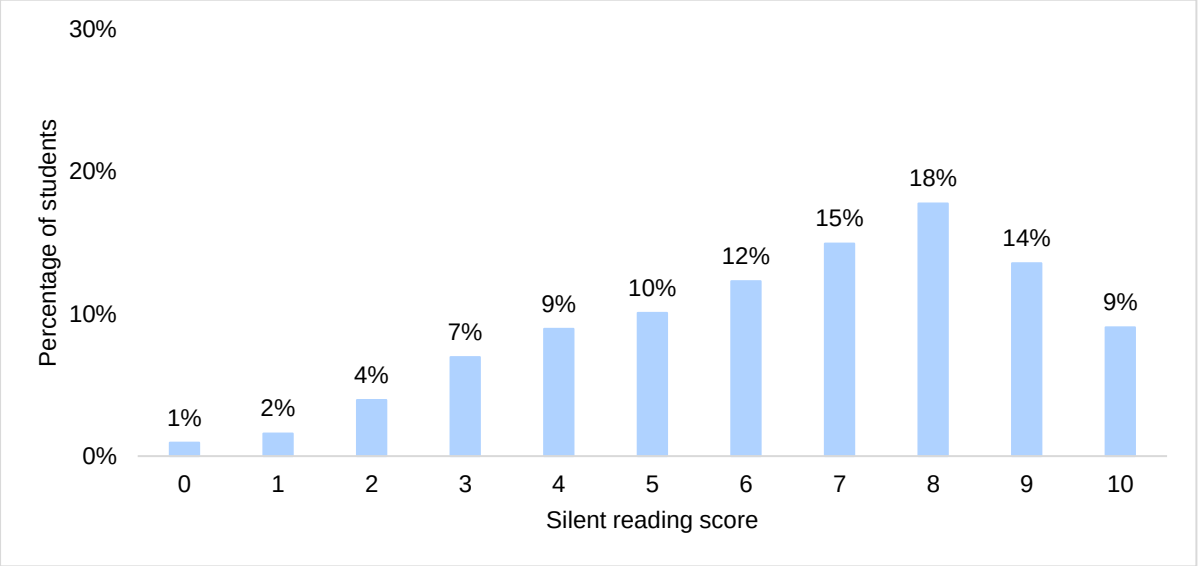
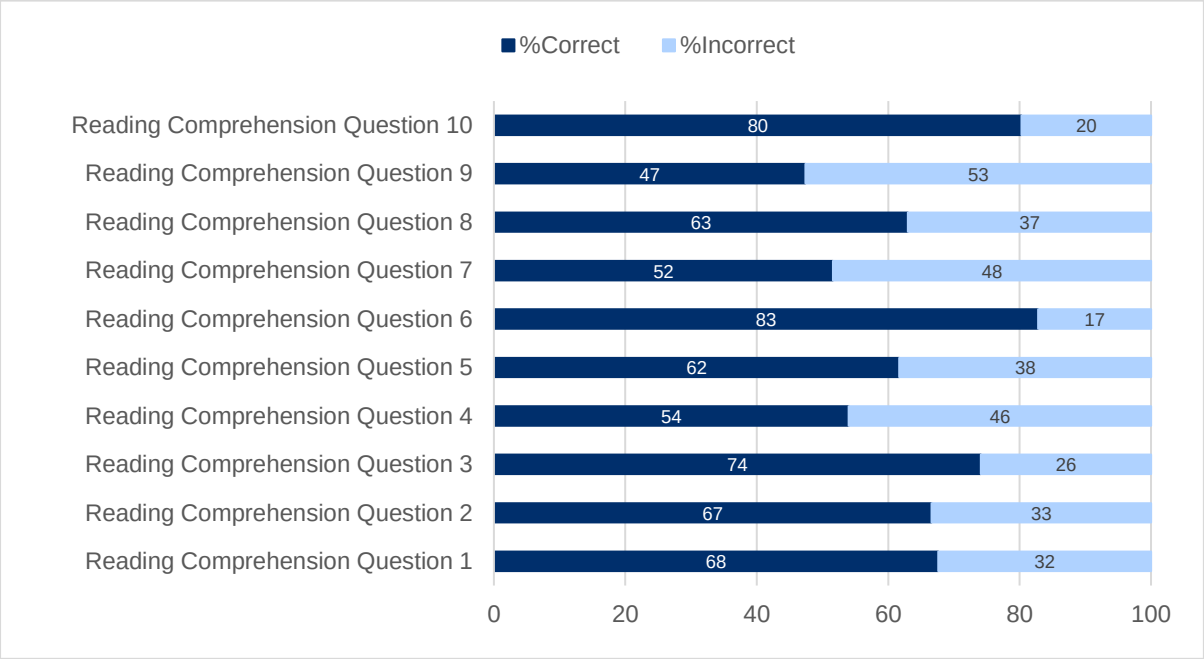


**Exhibit F-3. Silent Reading Comprehension Score Distribution, Grade 4**



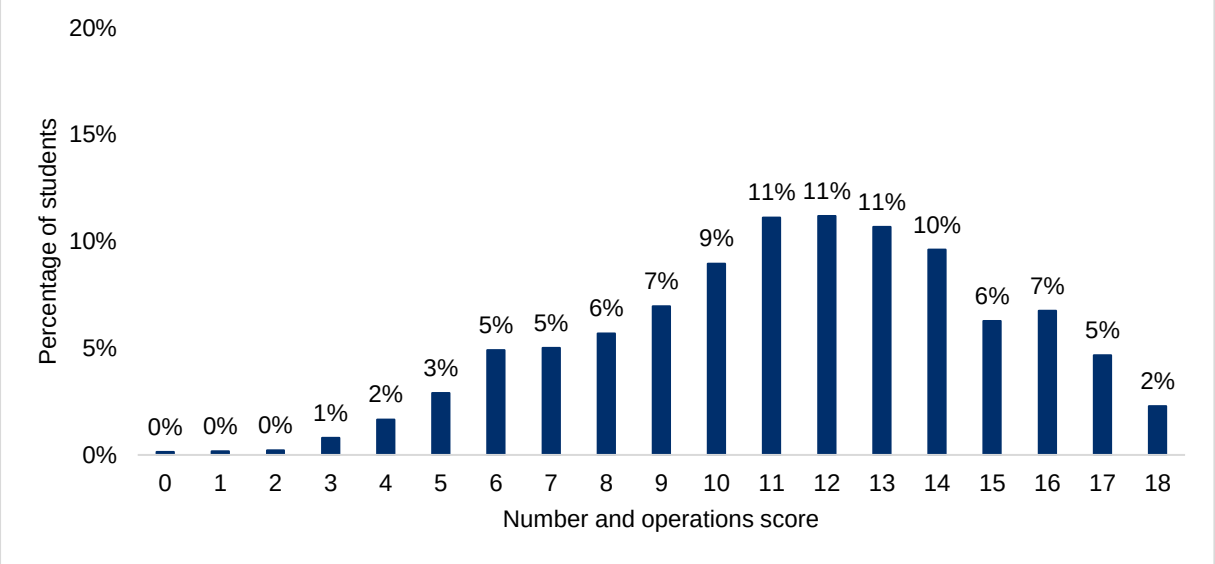
## ANNEX G: Grade 4 Silent Reading Comprehension Scores By Item

### Exhibit G-1. Silent Reading Comprehension Scores by Item

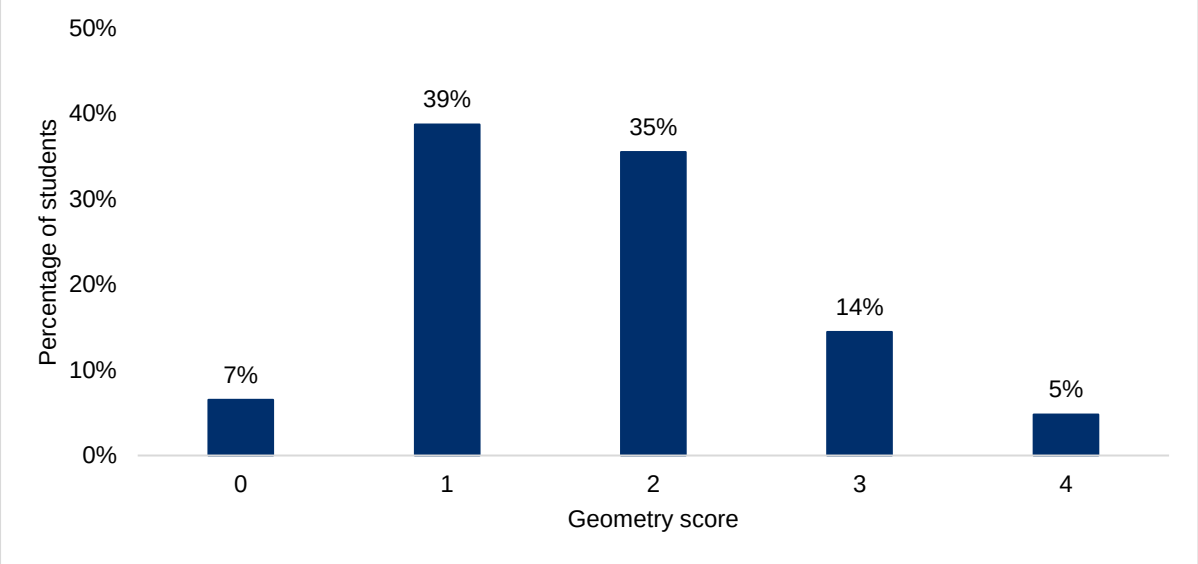


## ANNEX H: Grade 4 Score Distributions By Mathematics Domains

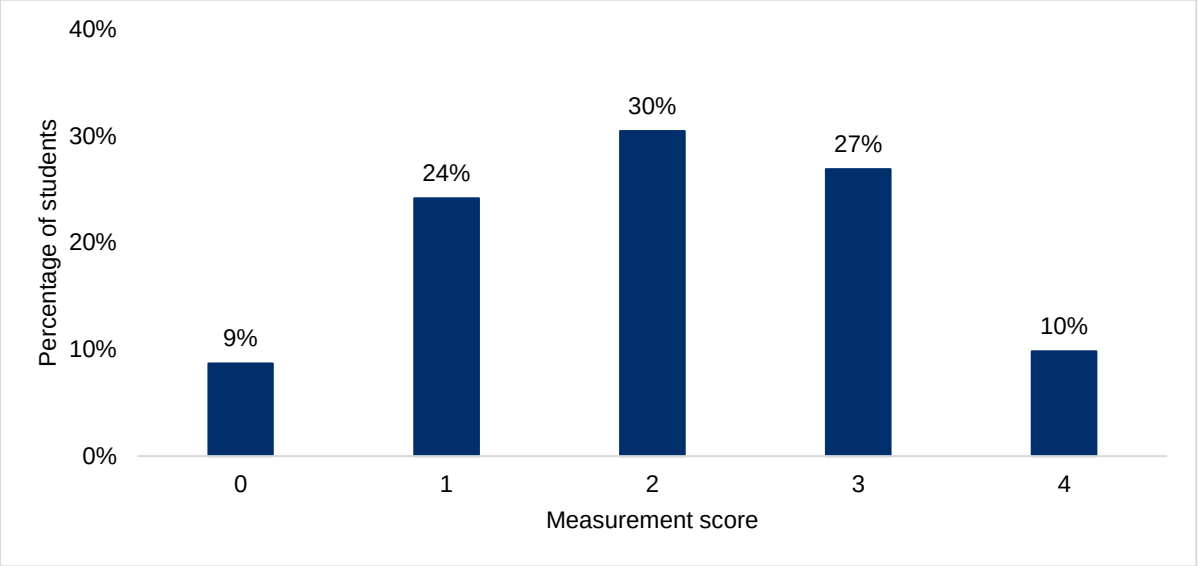
### Exhibit H-1. Numbers and Operations Score Distribution, Grade 4



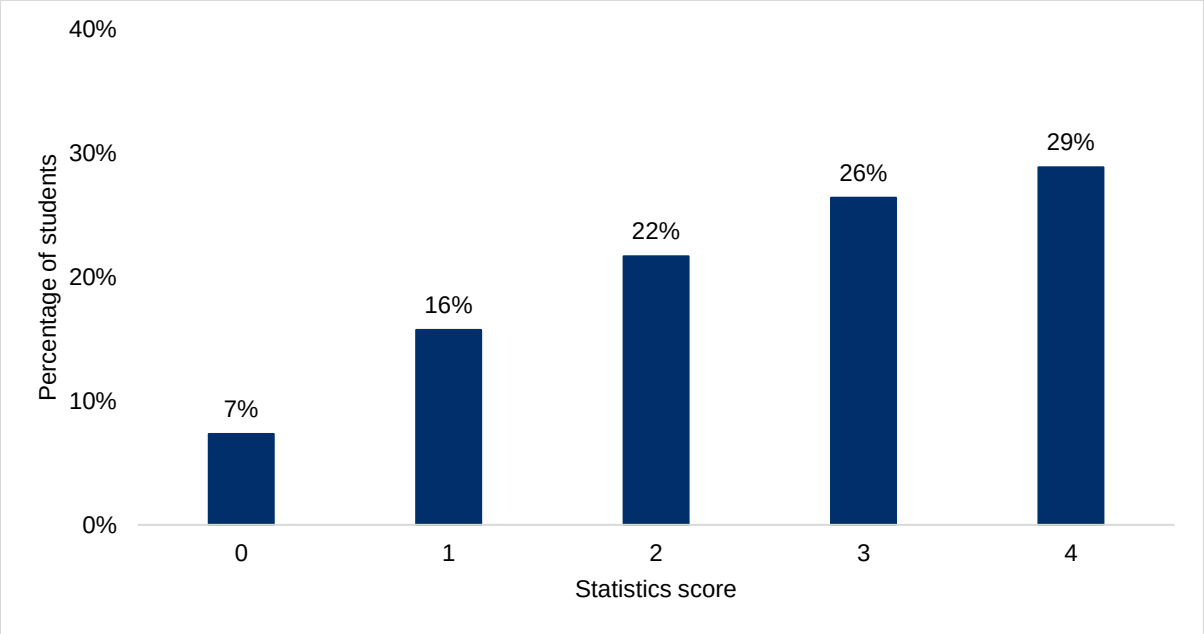

### Exhibit H-2. Geometry Score Distribution, Grade 4



## Exhibit H-3. Measurement Score Distribution, Grade 4




## Exhibit H-4. Statistics Score Distribution, Grade 4



## ANNEX I: Grade 2 and 4 Performance By Gender

**Exhibit I-1. Grade 2 National Uzbekistan Reading and Math Ability**

| | | | Baseline—Boys *N*= 310 | | Baseline—Girls *N*= 308 | | Endline—Boys *N*= 747 | | Endline—Girls *N*= 741 | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Subject** | **Subtask** | **Measure** | **Estimates_ 1** | **Precision_ 1** | **Estimates_ 2** | **Precision_ 2** | **Estimates_ 3** | **Precision_ 3** | **Estimates_ 4** | **Precision_ 4** |
| **Reading** | **Nonwords** | Fluency (correct letters per minute [clpm]) | **29.5** | [±2.1] | 33.7 | [±1.9] | 33.7 | [±2.3] | 37.2 | [±2.2] |
| | **Oral Reading Fluency** | Reading Fluency (correct words per minute [cwpm]) | **34.8** | [±2.8] | 44.7 | [±2.9] | 37.2 | [±2.7] | 43.7 | [±2.3] |
| | **Reading Comprehension** | % correct | 60.3 | [±3.5] | 63 | [±2.7] | 67.3 | [±3.4] | 70.7 | [±3.1] |
| **Mathematics** | **Missing Number [10 items]** | % correct | **66.7** | [±2.9] | 66.2 | [±3.5] | 75.5 | [±2.5] | 73.6 | [±2.8] |
| | **Word Problems [6 Items]** | % correct | **72.8** | [±3.5] | 71.7 | [±3.3] | 64.9 | [±3.2] | 64.7 | [±3.8] |
| | **Addition [5 items]** | % correct | **80.9** | [±3.4] | 79 | [±3.3] | 78.2 | [±3.2] | 75 | [±3.4] |
| | **Subtraction [5 items]** | % correct | **72.6** | [±4.2] | 71.6 | [±4.2] | 66.3 | [±4.0] | 65.7 | [±4.2] |
| | **Relational Reasoning [5 items]** | % correct | **60.1** | [±4.5] | 58 | [±4.9] | 56.5 | [±4.2] | 51.4 | [±5.0] |

## Exhibit I-1. Grade 2 National Uzbekistan Reading and Math Ability

| | | | Baseline—Boys *N*= 310 | | Baseline—Girls *N*= 308 | | Endline—Boys *N*= 747 | | Endline—Girls *N*= 741 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Spatial Thinking [4 items]** | % correct | **66.1** | [±4.2] | 58.8 | [±3.5] | 67.2 | [±2.8] | 60.7 | [±3.3] |

## Exhibit I-2. Grade 4 National Uzbekistan Reading and Math Ability

| | | | Baseline—Boys *N*= 309 | | Baseline—Girls *N*= 312 | | Endline—Boys *N*= 752 | | Endline—Girls *N*= 738 | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Subject** | **Subtask** | **Measure** | **Estimates_ 1** | **Precision_ 1** | **Estimates_ 2** | **Precision_ 2** | **Estimates_ 3** | **Precision_ 3** | **Estimates_ 4** | **Precision_ 4** |
| **Reading** | **Nonwords** | Fluency (clpm) | 39.3 | [±2.2] | 42.9 | [±1.6] | 45.7 | [±1.9] | 49.8 | [±1.9] |
| | **Oral Reading Fluency** | Reading Fluency (cwpm | 53 | [±3.0] | 64.8 | [±2.8] | 65 | [±3] | 76 | [±2.5] |
| | **Silent Reading Comprehension** | % correct | 58 | [±3.3] | 55.6 | [±2.7] | 64.3 | [±2.9] | 65.4 | [±2.7] |
| **Mathematics** | **Overall Score [30 items]** | % correct | 52.9 | [±3.7] | 52.2 | [±3.2] | 60.3 | [±2.5] | 57.8 | [±2.4] |
| | **Number and Operations [18 items]** | % correct | 55.5 | [±3.7] | 54 | [±3.2] | 64.8 | [±2.6] | 61.8 | [±2.4] |
| | **Geometry [4 Items]** | % correct | 38.2 | [±3.2] | 38.7 | [±3.4] | 43.5 | [±2.8] | 42.7 | [±2.9] |
| | **Measure [4 items]** | % correct | 50.7 | [±4.7] | 49.3 | [±4.5] | 53 | [±3.9] | 49.6 | [±3.8] |
| | **Statistics [4 items]** | % correct | 58.3 | [±5.8] | 60.2 | [±4.9] | 63.7 | [±3.9] | 63.1 | [±.6] |