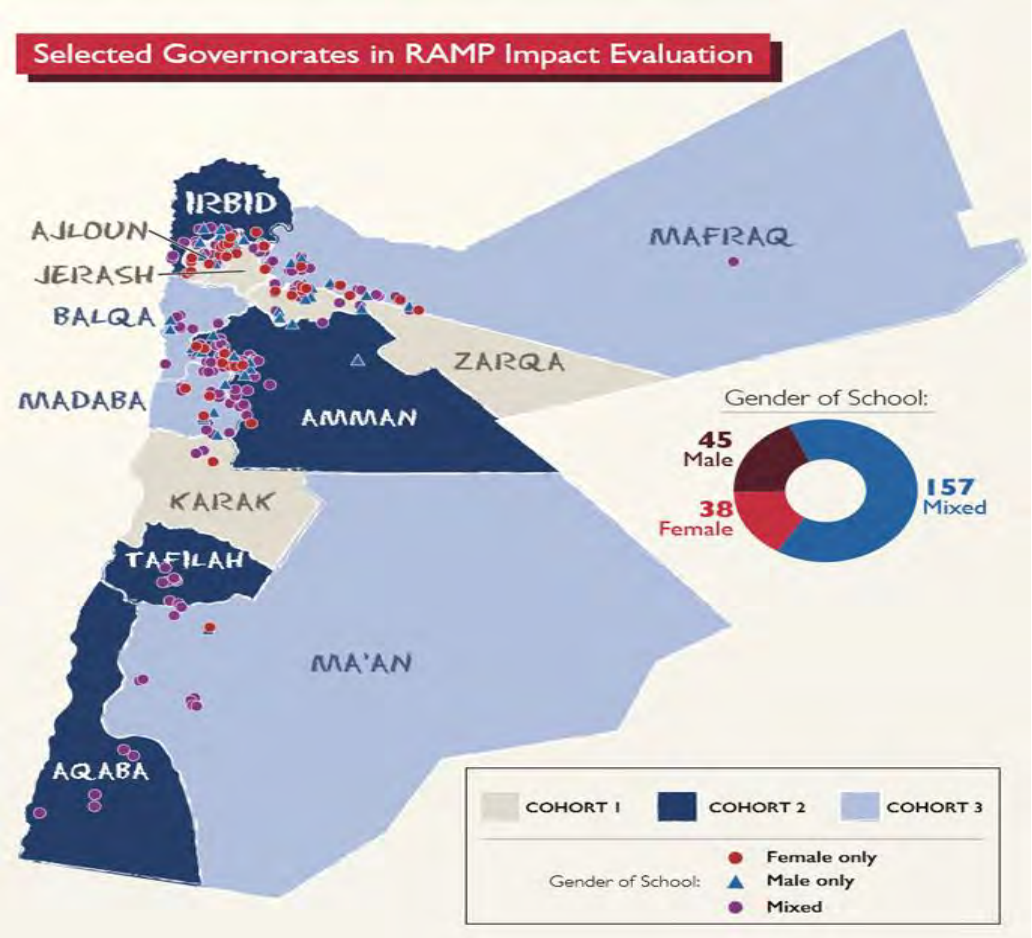
USAID
FROM THE AMERICAN PEOPLE



RAMP IMPACT EVALUATION FINAL REPORT

SEPTEMBER 2019

This publication was produced for review by the United States Agency for International Development. It was prepared by Management Systems International (MSI), A Tetra Tech Company.

RAMP IMPACT EVALUATION FINAL REPORT

Contracted under AID-278-C-13-00009

USAID/Jordan Monitoring and Evaluation Support Project

CONTENTS

ANNEXES (UNDER SEPARATE COVER)

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ADS	Automated Directives System
CADER	Change Agent for Arab Development & Education Reform
COTI	Classroom Observation of Teacher Instruction
CWPM	Correct Words Per Minute
DEC	Development Experience Clearinghouse
EA	Evaluability Assessment
EGMA	Early-Grade Math Assessment
EGRA	Early-Grade Reading Assessment
EMIS	Education Management Information System
EQ	Evaluation Question
FY	Fiscal Year
G1	Grade 1
G2	Grade 2
G3	Grade 3
GoJ	Government of Jordan
IE	Impact Evaluation
KG2	Kindergarten 2
KIIs	Key-Informant Interviews
LQAS	Lot Quality Assurance Sampling
MESP	Monitoring and Evaluation Support Project
MoE	Ministry of Education
MSI	Management Systems International
OLS	Ordinary Least Square
PE	Performance Evaluation
QRTA	Queen Rania Teacher Academy
RAMP	Early Grade Reading and Math Project
RTI	Research Triangle Institute

SOW	Statement of Work
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

PROJECT BACKGROUND

United States Agency for International Development Mission in Jordan awarded RTI International (RTI) the task to implement the Early Grade Reading and Math Project (RAMP) from 2015 to 2019. RAMP aims to improve: 1) teaching and learning by introducing new materials and curricula; 2) teacher and administrator instructional practices through pre-service and in-service training, mentoring and supervision; 3) community engagement activities; and 4) support for Jordan's Ministry of Education (MoE) efforts to institutionalize early grade reading and math policies, standards, and assessment (RTI 2014).

EVALUATION PURPOSE AND METHODS

The RAMP Impact Evaluation (IE) aimed to measure the impact of the intervention on teachers' instructional practices and student learning. The study selected a sample of 240 schools across Jordan (120 intervention schools receiving RAMP and 120 comparison schools that were to receive RAMP one year later). The quantitative analytical strategy of the evaluation used a longitudinal quasi-experimental design incorporating analytic strategies[1] to adjust for preexisting differences between intervention and comparison schools and students. The study created equivalent groups for comparison at baseline allowing measurement of RAMP impacts on students' reading and math scores and teachers' instructional practices. The design permits the study to test whether the intervention is the cause of any changes observed in students and teachers. The quantitative analytic strategy was complemented with a qualitative study using in-depth interviews and discussions with teachers and principals to gauge their general perceptions on early grade education in Jordan and about RAMP. This report presents the results of the IE, which follows the USAID evaluation report guidelines to summarize findings by evaluation questions.

EVALUATION QUESTIONS AND FINDINGS

1. **How does the RAMP training and mentoring support affect teachers' instructional practices? Do teachers' (1) use of instructional time, (2) student engagement, and (3) management of the classroom environment differ based on RAMP training and do they change over time?**

Overall, there were few statistically significant differences between Grade 3 (G3) intervention and comparison teachers on their use of instructional time, student engagement, and management of the classroom environment and no statistically significant impacts on teachers' instructional practices during reading lessons.

During math lessons, intervention teachers spent significantly less time than comparison teachers on number identification and writing. In addition, intervention teachers provided more specific and strategic feedback during math lessons than comparison teachers did. Table 1 summarizes the outcomes that had significant impacts in math and reading (out of the total number that were measured), and the direction of each impact (negative or positive).

G3 intervention teachers participated in RAMP trainings, received at least some coaching or mentoring, and were more likely than comparison teachers to use RAMP assessment tools in the classroom. Some

[1] The analytic strategies included propensity score matching to select a sample of similar intervention and comparison schools and propensity score weights at the student level to improve baseline equivalence between the groups.

G3 comparison teachers also reported participating in training and receiving coaching, as well as using the coarse- and fine-grained assessment tools.

While it is plausible that spillover, from intervention to comparison teachers, may have reduced the analytic ability to detect the impact of RAMP, it is unlikely given the small number of coaching visits to all teachers. RAMP teachers, on average, reported receiving fewer mentoring and coaching sessions than anticipated. Teachers also reported that it was difficult to integrate RAMP strategies with the MoE curriculum given time limitations, a lack of adequate materials, the high workload and requirement for documentation, and no incentives to implement RAMP. Additionally, some teachers admitted to implementing RAMP routines only when being observed.

TABLE 1: SUMMARY OF IMPACTS ON INSTRUCTIONAL PRACTICES AT ENDLINE

During reading lessons, RAMP teachers:	During math lessons, RAMP teachers:
Implemented RAMP in the classroom.	
• Implemented RAMP strategies • Integrated RAMP and MoE curriculum in lessons	• Used RAMP worksheets during lessons • Implemented RAMP strategies • Integrated RAMP and MoE curriculum in lessons
Shifted some lesson time away from basic skills towards more G3-appropriate content.	
• Were less likely to include writing activities in the lesson • Spent less lesson time on reading and identifying written characters • Spent more lesson time on vocabulary	• Were less likely to cover number writing and identification during the lesson • Spent less lesson time on number writing and identification
Had few impacts on classroom management and student engagement.	
• No significant impacts on student engagement. • Were more likely to use whole-class instruction	• Provided better feedback on student participation and written work • No significant impacts on classroom structure.

2. What are the impacts of RAMP on students' proficiency in reading and math?

After approximately two school years of implementation, RAMP had few statistically significant, positive impacts on students' reading and math outcomes (see Table 2). In other words, while students demonstrated growth in learning outcomes from year to year as they progress through grades, we do not see significant differences in the rate of growth between students that could be attributed to the RAMP intervention. Specifically:

- RAMP had a statistically significant positive impact on G2 students'[2] ability to segment words into syllables, but there were no other positive impacts on either G1 or G2 students' math or reading outcomes.
- RAMP had a statistically significant negative impact on G1 and G2 students' knowledge of letter sounds and on the percentage of G1 students who obtained a zero-score in the phoneme isolation and oral passage reading subtasks. The negative impact on knowledge of letter sounds and phoneme isolation may be related to intervention teachers' decreased focus on character reading and identification (see Table 1).
- The MoE set the goal of reducing the proportion of students unable to answer a single reading comprehension question correctly from 34 percent to 13 percent, by 2019. By 2017, when endline data were collected, about 40 percent of G1 students and 20 percent of G2 students (who were

[2] These students were in G3 at endline.

in G2 and G3 at endline, respectively) obtained a score equal to zero in the reading comprehension subtask, in both the intervention and comparison groups.

The overall lack of impacts at endline is consistent with teacher-level findings in this report (see Annex Q).

TABLE 2. SUMMARY OF RAMP READING AND MATH AVERAGE AND SUBGROUP AT MIDLINE AND ENDLINE, BY GRADE

READING	Midline		Endline	
	G1	G2	G1	G2
1. Orientation to print	No impact	NA	NA	NA
2. Phoneme isolation	No impact	NA	No impact	NA
3. Syllable segmentation	Positive impact [S]	Positive impact [S]	No impact	Positive impact
4. Letter sound knowledge	No impact	No impact	Negative impact	Negative impact
5. Non-word decoding	NA	No impact [S]	NA	No impact
6. Reading vocabulary	No impact [S]	No impact	No impact [S]	No impact
7. Passage reading	Positive impact [S]	No impact [S]	No impact [S]	No impact
8. Reading comprehension	No impact [G]	No impact	No impact [S]	No impact

MATHEMATICS	Midline		Endline	
	G1	G2	G1	G2
1. Counting numbers	Negative impact [S]	NA	No impact	NA
2. Counting objects (or enumerating quantities)	No impact	NA	No impact	NA
3. Number identification	No impact	No impact [G]	No impact	No impact
4. Number discrimination	No impact	No impact	No impact	No impact
5. Missing numbers	No impact	No impact	No impact	No impact [S]
6. Addition facts - L1	No impact	No impact	Negative impact	No impact
7. Addition facts – L2	NA	No impact	NA	No impact
8. Subtraction facts - L1	NA	No impact	NA	No impact [S]
9. Subtraction facts – L2	NA	No impact	NA	No impact [S]

Note: G denotes different impacts of RAMP for boys versus girl. S denotes different impacts of RAMP for students in single- versus double-shift schools. L1

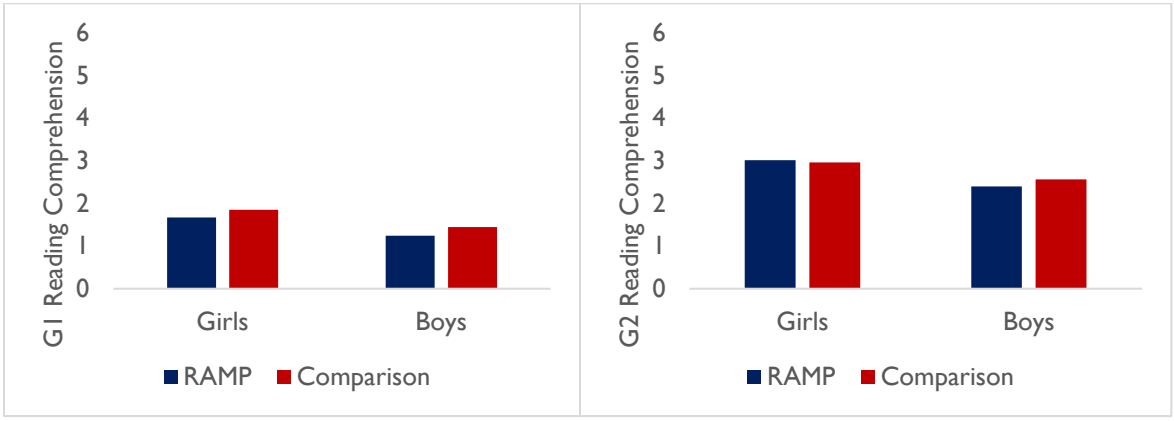
3. Do the RAMP impacts vary by gender, location (urban or rural), nationality (Jordanian or other), session (morning or afternoon), school type (boys, girls, or mixed; single or double shift), or whether the school received infrastructure support from USAID?[3]

- Overall, there were few statistically significant differences in RAMP impacts between boys and girls and no discernible patterns indicating that the intervention was more beneficial for either gender (see Figure 1).
- There were differential impacts on the proportion of students who obtained zero-scores in three reading subtasks (knowledge of letter sounds, invented word reading, and reading comprehension), but neither gender had a consistent advantage.

3 Due to small sample sizes in some cells, several variables could not be included in the analysis. Additionally, there was insufficient data on students' nationality and school infrastructure to include these items in the analysis.

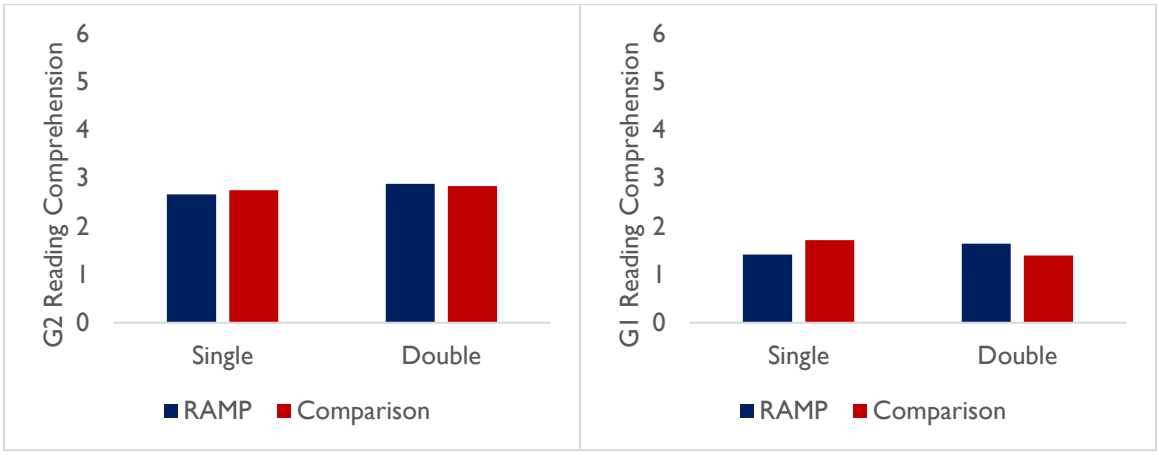
- The endline results reveal a tendency for RAMP to have a negative impact in single-shift schools and a positive impact in double-shift schools. However, most differences between intervention and comparison students in each type of school were not statistically significant (see Figure 2). A similar consistent pattern was not detected at midline (see Annex R)

FIGURE 1. GRADE 1 AND GRADE 2 STUDENTS' READING COMPREHENSION SCORES BY GENDER, FOR INTERVENTION (RAMP) AND COMPARISON SCHOOLS



Source: *RAMP Impact Study - Endline Data 2018 Student Assessments.* The figure shows endline scores equated into the baseline scale.

FIGURE 2. GRADE 1 AND GRADE 2 STUDENTS' READING COMPREHENSION SCORES BY NUMBER OF SCHOOL SHIFTS, FOR INTERVENTION (RAMP) AND COMPARISON SCHOOLS



Source: *RAMP Impact Study - Endline Data 2018 Student Assessments.* The figure shows endline scores equated into the baseline scale.

Overall, the analysis revealed limited RAMP impacts, likely due to implementation issues, including going to scale too quickly or scaling up prior to fully developing intervention components. In addition, for example:

- The school operating environment for RAMP implementation was complex; and
- RAMP program modifications may have affected learning outcomes.

As we have shown, with all such matters considered, the measurable impacts of RAMP are limited.

RECOMMENDATIONS

The study provides the following recommendations for stakeholders working to improve student level outcomes at the early grade level, including USAID, MoE officials, RTI and other implementers, school leaders, teachers, and other practitioners.

Recommendations are organized in five main categories: program design, program implementation, monitoring and evaluation, reporting and dissemination, and evaluation timeline.

RECOMMENDATIONS FOR PROGRAM DESIGN

RAMP could benefit from a more detailed development hypothesis or theory of change (ToC) that is clearly based on evidence of what works in early grade reading and math. In the future, program developers should develop the ToC in collaboration with key stakeholders to critically asses and validate the program's framework and logic model. Programmatic components should be adapted to reflect the local context and it should be clear which components need to be prioritized during implementation, how each component should be implemented, and how each is expected to contribute to achieving the program's goals. The ToC should also specify short, medium, and longer-term indicators that the implementer and evaluator can measure to track progress. Finally, it is advisable that the design call for implementer and evaluation partners to work closely from the beginning. This partnership can ensure that all necessary information is efficiently collected, analyzed, and utilized to improve program implementation. Additionally, such a partnership would be expected to enable an external evaluation team to consider and propose implementation plans that would support a randomized control trial with data collection intervals matched to implementation.

To support the implementer-evaluation partnership, USAID is advised to include in its design a full recognition of planned evaluation efforts, commencing both program implementation and evaluation design at the same time

RECOMMENDATIONS FOR PROGRAM IMPLEMENTATION

Despite the urgency to quickly improve the quality of education throughout Jordan, it is recommended stakeholders extend the implementation timeframe to ensure quality rollout of the program. A country-wide effort is too large in scale to implement with fidelity when the intervention is still adapting to challenges and not externally tested. An intervention can only be scaled with fidelity and sustainability after reaching success at a small scale. In future programming, it is recommended that stakeholders consider conducting an evaluability assessment prior to large scale program implementation to ensure the intervention has the potential to achieve its goals. Another way to improve RAMP implementation is to conduct frequent analysis of monitoring and evaluation (M&E) data collected during teacher coaching and mentoring visits. These data can inform stakeholders of the implementation gaps in real time and support improved fidelity to the program design.

RECOMMENDATIONS FOR MONITORING AND EVALUATION

Program refinement requires collecting high-quality M&E data to inform and guide implementers on which components need improvements. Implementers should align their M&E plan, staffing time, and technical capacity to ensure that data are used in a timely manner. It is essential to articulate a clear data collection and analysis plan to ensure that the data can provide valid answers to the implementer team's questions. For example, implementers should guard against bias during data collection due to over sampling from stronger regions or schools. Further, measurement instruments should be chosen that are valid and

reliable methods to assess the outcomes of interest in the target populations. For example, grade-specific EGRA and EGMA tools should be developed and tested for validity and reliability, rather than relying on one tool to measure learning in multiple grades, particularly given the expected growth and changes in curriculum from year to year.

RECOMMENDATIONS FOR REPORTING AND DISSEMINATION

Sharing program and M&E information is essential to developing fruitful collaborations among stakeholders. From the evaluation perspective, it would have been beneficial to have access to program materials at the onset of the project to inform measurement decisions. Also, to better coordinate data collection and dissemination efforts, especially in the face of disparate findings, the evaluation team would have benefitted from detailed and timely information about data collection efforts by the implementing team (for example, regarding instrumentation, sampling approach, and analytic strategy).

RECOMMENDATIONS FOR EVALUATION TIMELINE

Funders and stakeholders desire quick results; however, evaluations need to be designed to allow a longer exposure period. As mentioned, teachers may need more time to fully adopt and implement RAMP instructional strategies and students may require more years of teachers trained on RAMP to demonstrate improved performance. If funders, implementers, evaluators and other stakeholders work in partnership to carefully determine the length of exposure needed to meet short, medium, and long-term program goals, then the evaluation should be designed with a long enough timeline to measure these outcomes. The evaluation team recommends a second endline to test whether RAMP had impacts after a longer period of exposure.

INTRODUCTION

The Hashemite Kingdom of Jordan, through the Ministry of Education (MoE) and USAID, began a partnership in the 1950s to improve human potential by investing in public education. Since then, Jordan has achieved elevated levels of net enrollment in primary and secondary schools for both boys and girls (United Nations Educational, Scientific and Cultural Organization 2015).

Significant challenges remain, however. The quality of education remains weak, particularly in the early grades. Teachers' instructional practices have not kept up with best practices for student learning. Teachers have also lacked curriculum to guide their lessons and materials to support instruction. In addition, given the growing population of students and an influx of Syrian refugees, overcrowded classrooms pose an important challenge to teacher effectiveness and student learning (USAID 2011).

In the face of these challenges, the Government of Jordan (GoJ) has shown persistent commitment to investing in education, and the USAID Jordan Mission has supported the GoJ's efforts to address the continuing challenges. Current USAID investments aim to continue improving the quality of public education by supporting an evidence-based intervention and a rigorous evaluation, both of which were designed to advance progress in early grade reading and math. These investments support the MoE's efforts to increase literacy and numeracy, school completion rates, and access to schools, as well as to decrease gender disparities in education.

EARLY GRADE READING AND MATH PROJECT (RAMP)

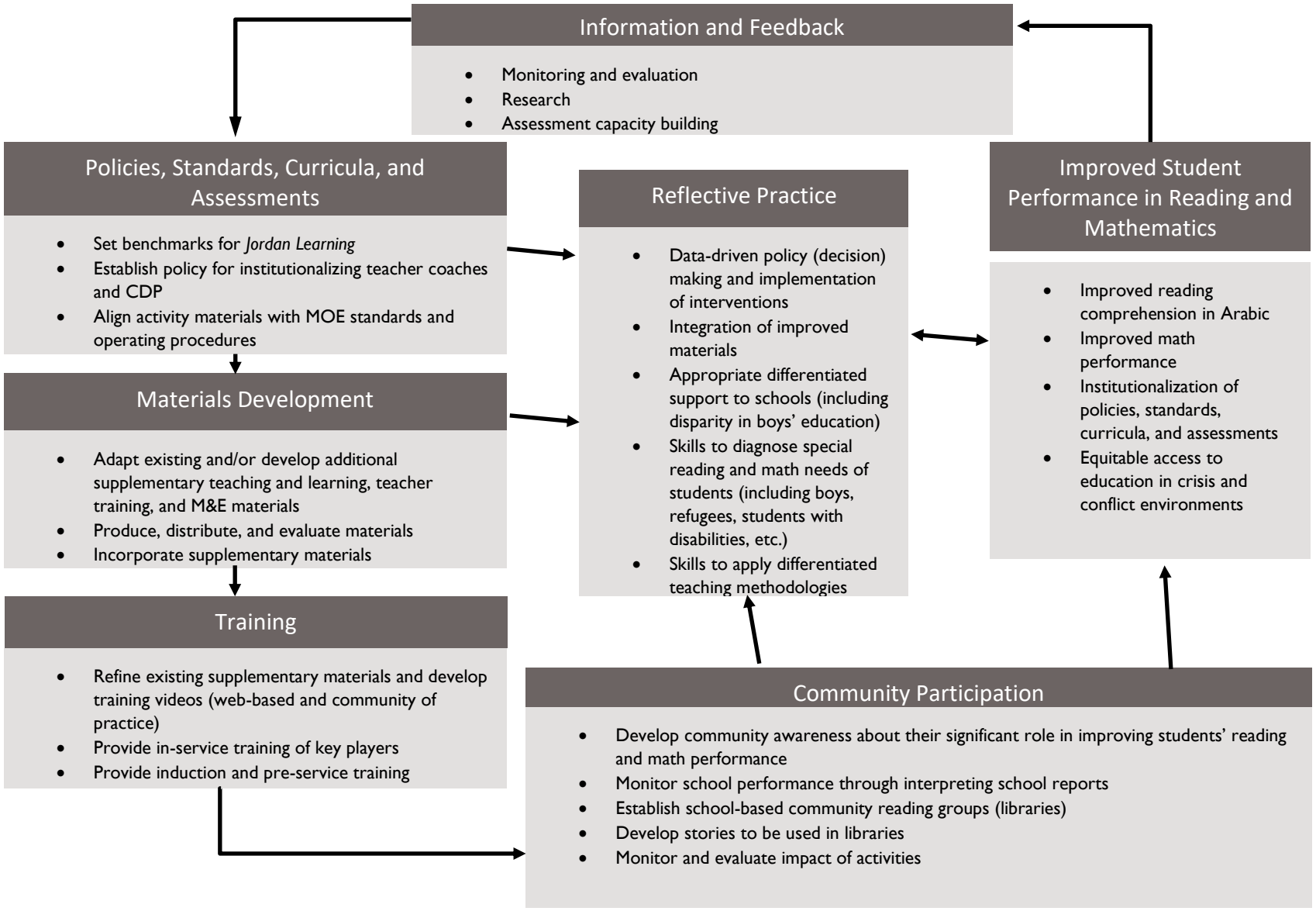
USAID contracted RTI to implement the early grade reading and math program (RAMP). According to project documents, the RAMP intervention consists of: 1) developing evidence-based early grade learning materials to be integrated into every Kindergarten 2 (KG2) to Grade 3 (G3) classroom; 2) improving teacher and administrator instructional skills through training, mentoring, and supervision; 3) mobilizing communities and parents to participate in children's learning; and 4) supporting MoE efforts to institutionalize early grade reading and mathematics policies, standards, and assessment (RTI 2014). The RAMP development hypothesis is as follows:

> The RAMP Initiative is a response to [these] challenges for early grade education in Jordan. Its primary goal is to improve or change the learning outcomes for reading and mathematics from K2 to G3 (Brombacher and Gargano 2017). The activity is constructed based on the following development hypothesis:
>
> *By investing in building MoE staff capacity, especially that of public-school teachers and supervisors, to use appropriate materials; research-based pedagogies; and, differentiated support to students according to their needs, RAMP will contribute to a substantially higher proportion of grade 2 and grade 3 public school students being able to read with comprehension and do mathematics with understanding by the end of the initiative. By also involving parents and communities in general in support of RAMP, the impact of the initiative will be significantly enhanced. Also, these gains will be sustained and built upon beyond the life of the initiative through the institutionalization of RAMP's research-based methodologies within a context of strong reflective practice.*

The development hypothesis is also presented in Figure 1. Note that the development hypothesis and other RAMP project descriptions present the intervention somewhat differently. Further, not all project activities listed in the hypothesis and project descriptions were implemented. The evaluation results identified weaknesses in the RAMP model (Figure 1) primarily in information/feedback portion and the mentoring component. There were several problems discovered in the quality and processes to monitor and assess RAMP. These are discussed in the qualitative report. RTI's own mentoring data was not complete and presented significant challenges for the evaluation team to use in analysis (as discussed in the qualitative report, Annex V). However, during the time of data collection, the team learned of the significant changes being carried out by RTI to create a modern robust M&E platform to track all indicators/metrics related to mentoring and trainings.

The IE did not address issues related to community participation as this was not part of the USAID-approved IE design. However, the qualitative report did go into some detail about the positive strides the IP had made on this front.

FIGURE 1: USAID RAMP DEVELOPMENT HYPOTHESIS



While the RAMP intervention did not specify the change in teacher and student outcomes expected in the short-, medium-, and long-term, a 2017 presentation by RTI states that the 5-year goal of RAMP is: "By 2019, 55 percent of early grade students in Jordanian public schools will be reading with comprehension and doing mathematics with understanding."

RAMP EVALUATION

To assess the contributions of RAMP to improve student learning outcomes, USAID contracted MSI, through the Jordan Monitoring and Evaluation Support Project (MESP), with support from its subcontractors Mathematica, Integrated, and Mindset, to conduct an impact evaluation (IE) of RAMP. Using both quantitative and qualitative approaches, the evaluation aims to measure RAMP's impacts on teachers' instructional practices and students' learning outcomes after one and two years of exposure to the intervention, as well as determine the cost effectiveness of the RAMP approach. After revisions to the evaluation's scope, per USAID's request, a cost-effectiveness analysis was not included as part of the evaluation.

This report describes the endline evaluation findings, synthesizes the findings of the evaluation as a whole, and provides an overview of the steps implemented for data collection and analysis, both for the quantitative and qualitative approaches.

EVALUATION PURPOSE AND QUESTIONS

PURPOSE

Results from the comprehensive quasi-experimental impact evaluation and the qualitative evaluation are intended to provide recommendations to improve the implementation of RAMP and its impact on school children. The findings will also guide future investments in education interventions for early grade reading and mathematics, as well as future RAMP evaluations.

The evaluation aimed to estimate the impact of the RTI-implemented RAMP intervention on students' literacy and numeracy and teachers' instructional practices. The evaluation team estimated the impacts of RAMP using a quasi-experimental design to compare students and teachers in two groups of schools that were as similar as possible—except for their exposure to RAMP. This design provided a robust test of differences between groups that result from the RAMP[4] intervention.

RTI worked to quickly scale the intervention across Jordan, assigning governorates and devising a schedule to implement RAMP in three phases across Cohorts 1, 2, or 3. RTI implemented a pilot version of RAMP in Cohort 1 schools prior to this evaluation. Cohort 2 schools were scheduled for implementation in August 2016, followed by Cohort 3 schools one year later. Therefore, for the impact evaluation, a sample of Cohort 2 schools were selected for the **intervention group (T)** and similar schools in Cohort 3 were selected for the **comparison group (C)**.

Using a matched comparison design, with data collection at baseline, midline, and endline, the research team tested the causal relationships between the RAMP intervention, teachers' instructional practices, and math and reading performance for students in grades one through three. Note that a randomized control design, in which schools are randomly assigned to the intervention, was not feasible because RAMP had already determined which schools would receive the intervention first. Further note that Cohort 2 results can be generalized throughout Cohort 2, as the sample is representative of Cohort 2 governorates.

[4] For details on the IE design and sampling process see Annex B.

Cohort 3 schools and students were matched to Cohort 2, so results are not generalizable throughout Cohort 3. Given that some districts in governorates were not reachable, the results likely capture some of the strongest performing schools in each Cohort, omitting hard-to-reach schools where systems and teaching may be lower quality, such as in Cohort 3.

EVALUATION QUESTIONS

The evaluation questions are as follows:

1. How does the RAMP training and mentoring support affect teachers' instructional practices? Do teachers' (1) use of instructional time, (2) student engagement, and (3) management of the classroom environment differ based on RAMP training and do they change over time?
    a. What practices do trained and non-trained G1 and G2 teachers implement with regard to teaching reading, writing, language and mathematics?**[5]
    b. What are the impacts of RAMP on G3 teachers' instructional practices in reading and math?
    c. How do practices change over time given ongoing mentoring and support for RAMP trained teachers?
2. What are the impacts of RAMP on G1, G2, and G3 students' proficiency in reading and math?
3. Did the RAMP impacts vary by gender, location (urban or rural), school type (boys, girls, or mixed; single or double shift), or whether the school received infrastructure support from USAID? (Note that the analysis focuses on the subgroup findings for child's gender and the number of school shifts.)
4. How cost-effective is the RAMP intervention for impacts on reading and math outcomes? **

Impact Evaluation Approach:

Test the **causal relationship** between the RAMP intervention and students' learning and teacher instructional practices using:

1. **A quasi-experimental design** because an RCT was not feasible due to the RAMP implementation plan.
2. **A matched comparison design** at the school level.
3. **Propensity score weights** at the student level.
4. **Data collected in three rounds**: baseline, midline and endline.

BACKGROUND

EGRA AND EGMA PILOT

From 2002 to 2014, USAID has invested $458 million in the Jordanian education sector, with programming focused on early childhood education (ECE), teacher training, technical training, and school infrastructure.

In 2011, as part of this commitment, USAID/Jordan commissioned a study, in partnership with the MoE, to measure basic reading (in Modern Standard Arabic) and mathematics skills among early-grade students. The study assessed students from a nationally representative sample of 156 schools using the Early Grade Reading and Early Grade Math Assessment tools (EGRA and EGMA). The findings revealed important

[5] Note that the questions marked with "**" were removed. Question 1a required data collection among G1 and G2 teachers; however, it was determined that this effort would not yield enough valuable information to observe these teachers a third time. Upon USAID's advice, the team removed Evaluation Question 4 due to limitations in accessing the required data.

deficiencies in both reading and mathematics skills among most students in G2 and G3 (Brombacher et al. 2012). The average G2 and G3 students had not mastered reading at their respective grade levels. Further, by the end of G3, most students had not yet acquired sufficient fluency skills to read with comprehension. The study also found that G2 and G3 students scored well on simple addition and subtraction but performed poorly on mathematics tasks that require both the understanding and application of procedural knowledge and problem solving (RTI 2014).

Following this assessment, RTI piloted an early grade reading and math intervention from 2011 to 2014. The pilot required teachers to implement brief daily sessions to reinforce past lessons. Also, students received workbooks to practice basic reading and mathematics skills. By the 2013-2014 school year, the pilot project trained more than 400 teachers in 43 schools serving nearly 12,000 students (Brombacher et al. 2014).

RTI conducted an internal evaluation of the pilot project using purposive sampling to construct intervention and comparison groups of schools. It is not clear whether the study groups were similar enough at baseline for the comparison group to accurately represent what would have happened without the pilot project. Nevertheless, RTI reported that students exposed to the pilot improved their reading and mathematics scores. However, these differences were not consistent across gender. Girls – particularly those in all-girls schools – achieved significant gains across every subtask, whereas boys in intervention schools did not perform significantly better than boys in comparison schools on any subtask. In Jordan, girls tend to outperform boys academically, so the intervention may have increased the gap between boys' and girls' learning outcomes (Brombacher et al. 2014).

Given RTI's interpretation of promising results, USAID/Jordan contracted with RTI in 2015 to develop and implement the RAMP early grade reading and math program countrywide. The RAMP activity built on, and was expected to improve upon, the results from the EGRA/EGMA pilot activity.

READING AND MATH PROJECT (RAMP)

USAID Jordan contracted with RTI to implement RAMP from 2015 to 2019. The Mission's $47.8 million investment in this project focused on improving learning outcomes for students in KG2 (the second year of a two-year preschool education), and in first through third grades (G1, G2, and G3).

The RAMP intervention aimed to improve Jordanian students' reading and math performance by providing teachers with (1) in-service and pre-service training opportunities, (2) mentoring, supervision, and coaching, (3) developing materials to support teaching and learning, training, and monitoring and evaluation efforts, and (4) encouraging community participation in students' learning process. Teacher training, in particular, aimed to develop teachers' skills to identify students' learning needs in reading and math and to apply differentiated teaching methodologies to respond to those needs. The main goal of RAMP was to build students' foundational skills in reading and math, so they demonstrate improved learning outcomes from KG2 to G3. RAMP's development hypothesis was that mastering early grade reading and math skills is essential to students' academic success and important to future economic opportunities (see Figure 1). Targeted investments in teacher training and mentoring, materials and other supplementary interventions will significantly increase the number of children performing at grade-level in reading and math.

During the period of the evaluation, the RAMP intervention consisted of:

- The development of evidence-based early grade reading and mathematics learning materials to be integrated into every KG2 to G3 classroom;
- Improved preparation of teachers and administrators to provide effective reading and mathematics instruction through pre-service training, in-service training, induction support, mentoring, and supervision;
- Community engagement activities to mobilize parents to participate in reading and mathematics education for all children and to hold schools accountable for results; and
- Advocacy for nationwide adoption of early grade reading and mathematics policies, standards, curricula and assessments (RTI 2014).

RTI implemented RAMP in all public schools throughout Jordan in three phases:

- **Phase 1** (Cohort 1) began in January 2016 and included the training of 3,718 KG2–G2 teachers in **623 schools**; G3 teachers were trained in July-August 2016, before the start of the 2016-2017 academic year.[6]
- **Phase 2** (Cohort 2) was launched in July-August 2016 and included the training of 4,509 KG2–G2 teachers in **1,087 schools**; G3 teachers were trained in July-August 2017.
- **Phase 3** (Cohort 3) was launched in July-August 2017 and included the training of 2,458 KG2–G2 teachers in **749 schools**; G3 teachers were trained in July-August 2018.

For details on the implementation of RAMP by governorate, see Annex A.

Given the development hypothesis and program parameters, the evaluation includes activities to assess and measure both:

1. **Teachers' instructional practices** and
2. **Student learning outcomes**

The research team observed teachers during reading and mathematics lessons to assess their instructional practices and administered learning assessment tools to measure students' mastery of foundational reading and math skills. Additional data collection was conducted to understand implementation and interpret study findings. The study methods are described in more detail in the next section.

EVALUATION METHODS

DESIGN

To answer the evaluation questions, the team used the following complementary approaches and data collection methods:

1. A **quasi-experimental impact study** to estimate the impacts of the intervention on G3 teachers' practices, using classroom observations. Impacts were estimated after teachers had one year of exposure to RAMP.
2. A **descriptive study** that explored Grade 1 (G1) and grade 2 (G2) teachers' instructional practices using classroom observations. This study was discontinued at endline because the completed impact study provides more rigorous evidence of RAMP implementation in the classroom.

[6] The academic year in Jordan starts in September and ends in June.

3. A **qualitative study** exploring adherence to the RAMP design and exposure/dosage issues that could potentially influence effectiveness of RAMP on students. Teachers, principals, RAMP staff, trainers, mentors and coaches were interviewed to help explain the "how" and "why" driving the observed quantitative outcomes. The full qualitative study is provided in Annex V.
4. A **quasi-experimental longitudinal study** to estimate the impacts of the intervention on students' learning outcomes, using the Early Grade Reading and Early Grade Math Assessments (EGRA and EGMA). Two impact estimates were calculated: impacts of two years versus one year of RAMP for students who were in G1 at baseline and impacts of two years versus no exposure to RAMP for students who were in G2 at baseline.

The evaluation questions are in Table 3 including their associated approaches and data sources (Figure 2).

FIGURE 2. RAMP IE DESIGN AND TIMELINE

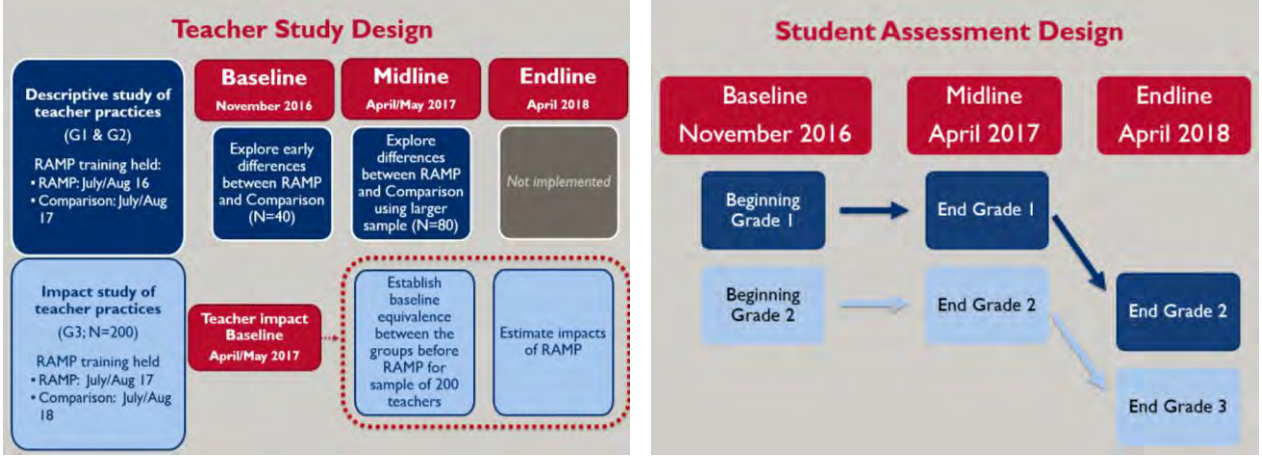


TABLE 3. EVALUATION QUESTIONS, APPROACH, AND DATA SOURCE

Evaluation Question (EQ)	Approach	Data Source
1. How does RAMP training and mentoring support affect teachers' instructional practices? Do trained and non-trained teachers' 1) use of instructional time, 2) student engagement, and 3) management of the classroom environment and do these differences change over time?	(See details for each sub-question below)	
1a. What practices do trained and non-trained G1 and G2 teachers implement with regard to teaching reading, writing, language and math?	Descriptive Study	Classroom observations of G1 and G2 teachers (baseline and midline only)
1b. What are the impacts of RAMP on G3 teachers' instructional practices in reading and math?	Quasi-experimental Impact Study of Teacher's Practices	Classroom observations of G3 teachers
1c. How do practices change over time given ongoing mentoring and support for RAMP trained teachers?	Descriptive Study	Classroom observations of G1 and G2 teachers (baseline and midline only)
	Quasi-experimental Impact Study of Teachers' Practices	Classroom observations of G3 teachers

Evaluation Question (EQ)	Approach	Data Source
	Qualitative Assessment	In-depth interviews with teachers, principals, and education leaders
2. What are the impacts of RAMP on G1, G2, and G3 students' proficiency in reading and math?	Quasi-experimental Impact Study of Students' Learning	Student academic assessments and surveys of students, teachers, and principals
3. Do the RAMP impacts vary by gender, location (urban or rural), nationality (Jordanian or other), session (morning or afternoon), school type (boys, girls, or mixed; single or double shift), or whether the school received infrastructure support from USAID?	Quasi-experimental Impact Study of Students' Learning	Student academic assessments and surveys of students, teachers, and principals
4. How cost-effective is the RAMP intervention for impacts on reading and math outcomes?	Cost Effectiveness Analysis	Per USAID's request, a cost-effectiveness analysis was not conducted.

DESCRIPTIVE AND IMPACT STUDIES OF TEACHERS' PRACTICES

The evaluation aimed to measure changes in teachers' instructional practices by conducting two systematic classroom observation studies:

- First, a small descriptive study was conducted of a sample **G1 and G2** teachers in intervention and comparison schools at two-time points:
    - Several months following RAMP training in 2016 (base year); and
    - One school year post-RAMP training in 2017 (midline)
- Second, an impact study was conducted of G3 teachers' instructional practices and classroom management at two-time points.
    - Prior to RAMP training for Cohort 2 in 2017 (baseline); and
    - One school year post-RAMP training for intervention teachers (endline)
- Finally, a detailed qualitative study was carried out, with focus on teacher training and mentoring, addressing (a) the extent and modes of adherence to the RAMP design; and (b) exposure/dosage issues that could potentially influence the effectiveness of RAMP on students. Interviews with teachers, principals, RAMP staff, trainers, mentors and coaches were conducted in the spring of 2018. The full qualitative study is provided in a separate report.[7]

For the **descriptive study of G1 and G2 teachers' practices,** classrooms were observed, and teacher practices were assessed at two time points (in December 2016 and April-May of 2017, respectively four and nine months after the implementation of RAMP had begun). Results from this descriptive study are available in Annex I.

This report focuses on results from the **quasi-experimental impact study of G3 teacher's practices,** which takes advantage of the staggered rollout of RAMP to observe teachers with a true baseline prior to RAMP training. Baseline data in G3 classrooms coincided with midline data for the other studies in April 2017, prior to RAMP training (conducted in the summer of 2017). G3 teachers in comparison schools were trained in the summer of 2018. This design generates rigorous evidence on the

[7] *Teacher Training and Mentoring: A Qualitative Study of RAMP* (May 2019), Annex V to this report.

impact of one year of RAMP on instructional practices, provides insights into short-term outcomes along the causal pathway, and presents context to help interpret findings from the student-level impact study.

A separate teacher survey was also conducted to gather data on teachers' background characteristics and implementation of RAMP in their classrooms.

One external factor to note is that the MoE disseminated a new national curriculum booklet and conducted teacher training during the 2016-2017 school year. **This national effort is expected to affect both the intervention and comparison schools similarly.** Because findings at midline showed that intervention and comparison G3 teachers demonstrated similar instructional practices before RAMP was introduced to Cohort 2, the evaluation was designed to capture impacts attributable to RAMP.

Principles of an Impact Evaluation (IE)

- IEs rely on comparing similar groups at baseline, prior to an intervention
- The two study groups must be similar at baseline so the comparison group serves as a an estimation of what would have happened without the intervention (counterfactual).
- Similar groups enable evaluators to confidently attribute any observed difference in outcomes to the intervention
- Alternative explanations—such as that differences were due to pre-existing between-group differences—can be ruled out
- At times an IE uses baseline data that captures status of an intervention some time later than the actual beginning of the intervention. This is the case for this evaluation. In such cases, the evaluation team uses various appropriate and accepted techniques to mitigate against technically inappropriate analysis of baseline data.

QUALITATIVE FINDINGS

The evaluation report includes the key highlights from qualitative data collected over the course of the evaluation to better understand how and why there might be change/or no change in student performance and changes in teacher's approaches to math and reading. Qualitative data collection efforts have included:

- In-depth interviews with teachers and principals to understand their perspectives on RAMP and its application in classrooms. The team also attended several RAMP teacher training sessions across the whole country in August of 2017 to observe the intervention for both Math and Reading components.
- Key-informant-interviews (KIIs) of stakeholders to capture their perspectives and views of RAMP, as well as to understand key aspects of implementation fidelity. This included interviews with RTI, USAID, Dajani, CADER, QRTA, Ministry of Education, and other partners.
- In-depth discussions and classroom observations focused on the mentoring/coaching component of RAMP. The team interviewed teachers, principals, RAMP staff, trainers, coaches and mentors. Additionally, the team analyzes some of the monitoring data on mentoring collected by RTI. Results are provided in the qualitative study report (Annex V).

QUASI-EXPERIMENTAL IMPACT STUDY OF STUDENTS' LEARNING

The evaluation used a quasi-experimental design to estimate the impact of RAMP on the learning outcomes of G1 and G2 students who received RAMP starting in different years (see Table A.1. in Annex A).

- Specifically, midline analysis estimated the impact of approximately one year of exposure to RAMP versus no exposure[8], for both G1 and G2 students.
- At endline, for students who were in G1 at baseline, estimates reflect **two years** of exposure for the intervention group versus **one year** of exposure for the comparison group, as opposed to two years versus no exposure.
- At endline, for students who were in G2 at baseline, impact estimates reflect **two years** of exposure for the intervention group versus **no exposure** for the comparison group.

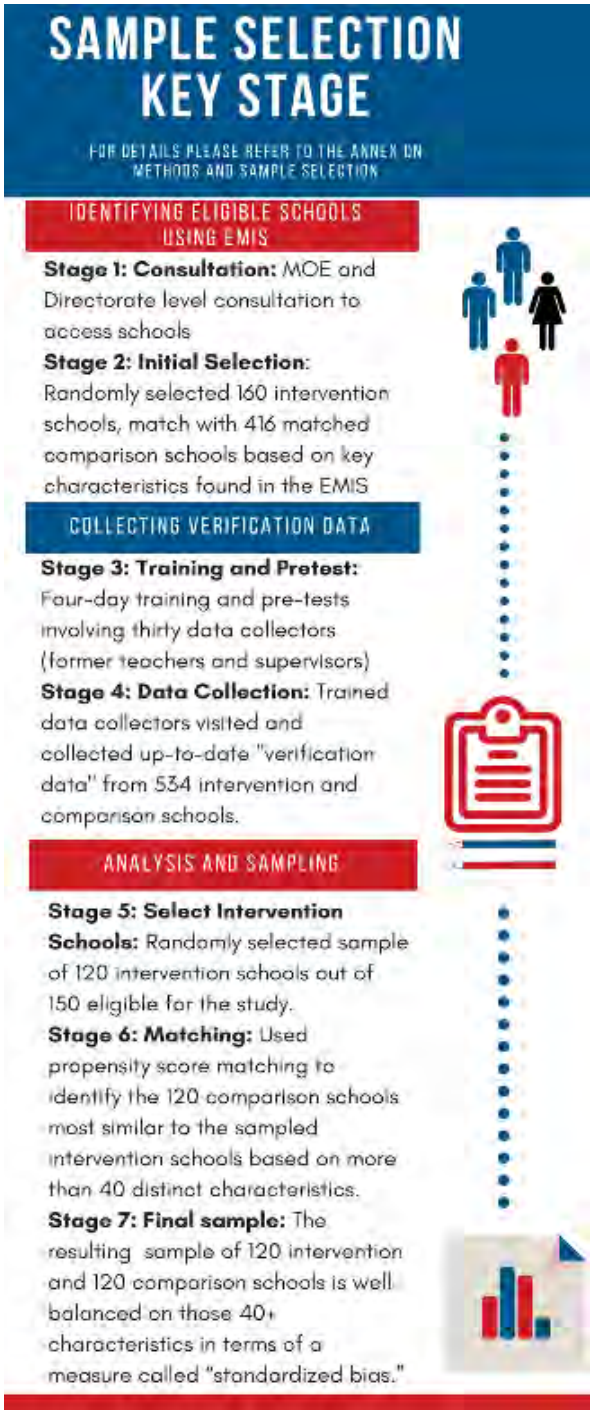
The study used the EGRA and EGMA to measure students' foundational reading and mathematics skills, such as the knowledge of letter sounds and the ability to identify numbers, which contribute to the development of more advanced skills such as reading with understanding and arithmetic.

SAMPLE

A multistage sampling process was implemented (Figure 3) to select the schools to conduct the teacher observations and student assessments (See Annex B for more information).

- First, the research team examined Education Management Information System (EMIS) data to understand the geographical distribution, characteristics, and the number of schools across governorates. Cohort 2, as assigned by RTI, consisted of 1,028 schools, which served as the intervention group sampling frame. Cohort 3 was comprised of 679 schools and served as the comparison group sampling frame.

FIGURE 3. DIAGRAM OF THE SAMPLE SELECTION STAGES.



8 Estimates were actually computed after 6 months of exposure to RAMP.

- Schools in Cohorts 2 and 3 were classified into six strata based on shift status and gender of the students (i.e., single-shift all girls, all boys, or mixed-gender schools, or double-shift all girls, all boys, or mixed-gender schools).
- Next, the research team randomly selected an initial sample of 160 intervention schools and 416 potential comparison schools, proportional to the size of each stratum in the intervention group.
- Given concerns with EMIS data quality, a verification activity was conducted in which schools were visited to obtain accurate, up-to-date school level information. This verification data informed the school matching process.
- The research team used the verified school data to select a final stratified sample of 120 schools, representative of the intervention group (Cohort 2).
- Then, using a combination of EMIS and verification data, a sample of 120 comparison schools (Cohort 3) was selected based on their similarity to intervention schools. **This multi-stage sampling process yielded 120 intervention schools and 120 matched comparison schools suitable for measuring the impacts of RAMP** (Table 4).
- Once in schools, the team randomly selected one G1 and one G2 classroom and, within each classroom, a random sample of 10 students to assess with the EGRA and EGMA. At endline, EGRA and EGMA were administered to the same students who were assessed at baseline and midline and who had advanced to G2 and G3, respectively.

Subsamples were selected from the above school sample for the descriptive and the impact study of teachers' instructional practices:

- For the descriptive study of G1 and G2 teachers and qualitative interviews, the team purposively selected a subsample of 40 schools (20 intervention and 20 comparison) from the 240 schools. Within schools, the team randomly selected one G1 and one G2 teacher per grade to observe and interview.
- In addition, 200 schools (100 intervention and 100 comparison) were randomly selected from the 240 schools for the impact study of G3 teachers. At the baseline of the teacher impact study, which coincided with midline data collection for the student study, teachers were randomly selected from all G3 teachers in each school. The team tried to observe the same teachers at endline. However, some teachers were not available because they moved, retired, or were assigned to another grade. In these cases, a replacement teacher was randomly selected from the same grade and school.

TABLE 4. STUDY SAMPLE AND DATA COLLECTION TIMING

Study	Intervention Schools	Comparison Schools	Intervention Students/ Teachers	Comparison Students/ Teachers	Was data collected at this time for this group? Baseline (October 2016)[9]	Midline (May 2017)	Endline (May 2018)
Longitudinal impact study of student learning	120	120	1,939 students at endline	1,954 students at endline	Yes	Yes	Yes
Descriptive study of G1 and G2 teachers' practices & qualitative study*	20	20	40 teachers at midline	40 teachers at midline	Yes (after teachers had been trained)	Yes	No
Impact study of G3 teachers' practices**	100	100	100 teachers	100 teachers	No	Yes, data collected in 2017 served as baseline, prior to training	Yes, teachers trained after the endline data collection

* Denotes purposive subsample of sample for impact study of students' learning
** Denotes random subsample of sample for impact study of students' learning

[9] Baseline data was collected six to eight weeks after the 2016-17 school year began. For further information, see the Limitations section below (p. 27) and Annexes G and P.

MEASUREMENT INSTRUMENTS

CLASSROOM OBSERVATION OF TEACHERS' INSTRUCTION (COTI)

Teachers and classrooms were observed using the COTI to capture information on (1) instructional practices and resources, (2) use of instructional time, (3) student engagement, and (4) the classroom environment. These evidence-based domains have been widely shown to positively affect student learning in countries around the world (Hamre et al 2013; details on the development of COTI are provided in Annex H. Observers collected data on all math and reading lesson segments that occurred on the day of the observation, which ranged between 1 and 3 per school. A standard number of lessons could have been observed across schools, but the team chose to capture RAMP's influence on additional weekly math and Arabic lessons that were implemented under directions from the MoE.

COTI was also used to capture RAMP implementation in the classroom. During training, classroom observers were sensitized to a wide range of RAMP instructional strategies, such as using clapping and hand gestures to emphasize and differentiate phonemes and syllables, asking reading comprehension questions after each paragraph, using songs and manipulatives to teach counting skills, and teaching multiplication as repeated addition (see Annex H). Teachers were rated on the frequency with which they demonstrated RAMP strategies and integrated RAMP and MoE curriculum on a scale from zero to two, where zero indicated that none of the strategies was used and two indicated that multiple strategies were used or integrated throughout the lessons.

Per midline results, the team compiled feedback from RTI and USAID to revise the tool for endline data collection. Items on the COTI tool that were not used in the midline analysis were dropped, and questions on math and reading lesson content were broken into 15-minute segments that provide additional detail on the content covered.

FIGURE 4. TOOL DEVELOPMENT PARTNERS AND STAGES.

EARLY GRADE READING AND MATH ASSESSMENTS (EGRA AND EGMA)

Students' foundational reading and mathematics skills were assessed using the EGRA and EGMA at baseline, midline, and endline (see Figure 4 and Annex E for additional details on these measurement tools).

The G2 endline EGRA was like the G1 version administered at midline, but it did not include the knowledge of early print concepts (or orientation to print) task. The G2 endline EGMA and G3 endline EGRA and EGRA were the same as the G2 versions administered at midline.

To account for the difference between the tools from baseline to endline when estimating students' scores at endline, a psychometric process called test equating was used (Annex F provides more details on the equating process and on the instrument development process).

ANALYTIC APPROACH

QUANTITATIVE

A difference-in-differences approach was employed to estimate impacts on teacher practices because teachers were more likely than students to move to another school or teach a different grade, making it difficult to observe the same teachers over time. Ordinary least square (OLS) regressions were used to estimate the impacts of RAMP on student outcomes. To account for the nesting of students within schools, and teacher observation segments within classrooms, robust clustered standard errors were used. Annex G provides a full description of the analytic approach.

To improve the study's ability to estimate impacts on students' learning, the sample of comparison schools was selected using propensity score matching and the student sample was weighted using student-level propensity score weights. These two strategies improved the similarity between the intervention and comparison groups on key school and student characteristics thought to affect student learning outcomes and allow the attribution of post-intervention differences in outcomes to the intervention and not to pre-existing differences between the groups.

QUALITATIVE

The team conducted content analysis of all the qualitative data (transcripts of the in-depth interviews and discussions with teachers and principals). Where appropriate and when large samples were present, frequency tables were developed to quantify responses against key themes, challenges and opportunities.

HUMAN SUBJECTS AND CONSENT

The evaluation followed standard practices in the protection of human subjects, including standards for working with children. The MoE approved access to the schools, principals, teachers, and students (see consent letter in Annex W). Consent from individual participants was requested at the time of data collection. Participants were read the list of potential benefits and harm and were told about their rights to discontinue involvement at any point during the evaluation. There was minimal risk associated with participating in this study.

LIMITATIONS

This mixed-methods evaluation offers evidence not available from any other sources and has several notable strengths: 1) a quasi-experimental design with matched schools, rigorous sampling and analytic approaches, and 2) the use of multiple data sources that can be triangulated, including student assessments, teacher and principal qualitative interviews, and classroom observation data. However, like all studies, this evaluation has limitations to consider when interpreting the results.

First, as with all quasi-experimental impact evaluations, there may be observed and unobserved differences between schools, teachers, and students in the intervention and comparison groups, which may pre-date RAMP and bias the impact estimates. To address differences in *observed* characteristics, a propensity score matching procedure was implemented using administrative and school verification data, which permitted selecting a sample of comparison schools as similar as possible to the sample of intervention schools. The data available for matching was limited, however. Notably, the Jordanian educational system does not have a standardized measure of student achievement in the early grades[10], which prevented the evaluation team from matching schools on individual student achievement, a key outcome of the evaluation. Therefore, the study groups had similar school characteristics at baseline, but analysis revealed statistically significant differences in students' academic performance. Student-level propensity score weights were used to rebalance the sample (Golinelli, Ridgeway, Rhoades, Tucker, & Wenzel, 2012; McCaffrey, Ridgeway, & Morral, 2004). The reweighted student sample has improved comparability in observed baseline student demographic characteristics and academic outcomes. There may be, nonetheless, differences in *unobserved* characteristics that cannot be directly addressed and that could bias the impact estimates. In short, the evaluation team reduced the threat of bias due to observed differences between the groups, but the threat of bias due to unobserved differences remains.

Second, although Cohort 2 teachers were trained in the summer of 2016, baseline data collection took place six to eight weeks into the 2016-2017 school year. The delayed baseline raises the question of whether differences between the two groups existed prior to RAMP implementation or resulted from up to six weeks of early RAMP implementation. Classroom observations and interviews with teachers and principals offered no persuasive evidence to suggest that RAMP caused notable differences in the relatively short period between the beginning of the school year and baseline data collection for this evaluation.[11] Further, those early weeks of school were interrupted by holidays and teacher strikes. Nevertheless, the process of analytically adjusting for baseline scores, even if data were collected after the intervention had begun, would contribute to improving precision in the impact estimates. It is also a requirement for quasi-experimental designs (Schochet 2008). The results presented in Annex P indicate that the conclusions do not change when excluding baseline scores from the analysis.

Third, the evaluation timeline may have been too short to detect intervention impacts, should they emerge. The evaluation followed teachers for only one school year after exposure to RAMP and students for only two school years of exposure to RAMP trained teachers. It may be that changes in teacher instructional practices require a longer timeline and the adoption and mastery of RAMP will occur over more than a year. Further, assuming the development hypothesis of RAMP is correct, students' performance improve will only improve with full adoption and integration of these practices. The overall pattern of results from this evaluation, however, does not support the hypothesis that RAMP impacts grew incrementally over time. There were few statistically significant positive impacts—and even some negative impacts detected—when comparing G1 students exposed to one versus two years of RAMP and

[10] Principals' reports of average achievement were used.

[11] In response to a draft of this report, RAMP expressed concerns about proper analysis of the baseline data. The evaluation team carefully considered these concerns, discussing them extensively with RAMP and USAID. Annexes G and P describe the evaluation team's analyses and actions to address the concerns.

G2 students exposed to two years of RAMP versus students who had not been exposed. Despite the short timeframe of this study, interventions focused on building teachers' capacity, which have been evaluated with comparable time frames to the RAMP intervention and evaluation, have been proven effective in other contexts (McEwan, 2015).

A final key limitation of the study is the role of group contamination and intervention spillover, which may attenuate the contrast between the intervention and comparison groups. At endline, a non-trivial number of G3 teachers in the comparison group reported having participated in RAMP trainings (17 percent) and receiving support from RAMP coaches (5 percent). About half of teachers in the comparison group also reported being familiar with the coarse-grain tools and a fifth reported using the fine-grain tools. Comparison teachers may have confused resources provided by the MoE with those provided by the RAMP intervention, but contamination and spillover cannot be ruled out. Further, RAMP started training KG2-G2 teachers in Cohort 3 (the comparison group) in the 2017-2018 school year. Therefore, the design of the implementation plan may have reduced the contrast between intervention and comparison students who were in G1 at baseline and G2 at endline. Additionally, even though G3 teachers in Cohort 3 were supposed to start participating in RAMP one year after the evaluation was completed (2018-2019), they may have been exposed to RAMP resources available to other teachers at their school during the year of endline data collection. Nevertheless, there were statistically significant differences between the intervention and comparison groups in the proportion of teachers who were trained and who reported receiving support from RAMP and using RAMP's tools. Despite contamination and spillover, it was still feasible for the evaluation to have detected impacts of RAMP, had they emerged. Further, given the low number of coaching visits to intervention and comparison teachers, it is more likely that the intervention was not implemented as intended or with the fidelity necessary to change teaching practices within two school years.

KEY FINDINGS

EVALUATION QUESTION (EQ) 1

EQ 1.1 – WHAT PRACTICES DO TRAINED AND NON-TRAINED G1 AND G2 TEACHERS IMPLEMENT WITH REGARD TO TEACHING READING, WRITING, LANGUAGE AND MATH?

Classroom observation data were collected from a small sample of G1 and G2 teachers in intervention and comparison schools to answer this question and provide context to the assessment of the impacts of RAMP on students' learning. This activity was conducted approximately six months after RAMP implementation had begun (corresponding to midline data collection for the full evaluation). While the observations revealed that RAMP trained teachers were more likely than comparison teachers to use RAMP materials, there were few significant differences among G1 and G2 trained teachers indicating that teachers may not have fully adopted the key components of RAMP training.

The sample consisted of only 79 teachers (40 or fewer in each grade), which limits the power to detect differences between the groups. Therefore, the analysis examined both differences between the groups that were statistically significant, and those with large effect sizes. Statistically significant means that it is unlikely that differences are due to random chance; differences with a large "effect size" exceed 0.25 standard deviations in magnitude and are considered substantively large (What Works Clearinghouse, 2015). Differences can be statistically significant, have a large effect size, both, or neither. The full results are provided in Annex I. The main findings from these descriptive analyses indicate that after approximately six months of implementation:

- Intervention teachers were **implementing RAMP** in the classroom. Intervention teachers in G1 and G2 showed a greater propensity than comparison teachers to use RAMP worksheets and RAMP instructional strategies in their classrooms during both math and reading lessons. These differences were statistically significant.

There were also some differences in **lesson content** that varied in G1 versus G2:

- During reading lessons, G2 intervention teachers were 18 percentage points more likely to **include writing activities** in the reading lesson. This difference was large and statistically significant. However, the evaluation team could not assess the quality of the writing activity to determine the practical importance of this finding.
- Also, during reading lessons, more G2 intervention teachers covered **phonemic awareness**, a 21percentage point difference. This difference was large but not statistically significant, likely due to the small sample size.
- During mathematics lessons, G2 intervention teachers were 20 percentage points more likely than comparison teachers to cover foundational skills, such as **identifying and writing numbers**. G2 intervention teachers were also 15 percentage points more likely to cover **advanced arithmetic** skills, although the reverse was true in G1. These differences were large but not statistically significant, again, likely due to the small sample size.

There were, however, no clear findings across the 12 measured **instructional practices**. The only statistically significant difference was that intervention teachers were less likely than comparison teachers to use **assessment** practices during reading lessons, which was surprising given RAMP's emphasis on assessing student learning levels. There were also some large but statistically insignificant differences:

- Intervention teachers also scored higher on **student engagement practices** and demonstrated fewer behaviors that cultivate a **negative classroom climate** during reading lessons**.**
- During math lessons, teachers in the intervention group spent an additional four percentage points of the lesson **time on task,** versus time spent managing disruptions, transitions between activities, and student behavior**.**

Data from the descriptive study can provide insights into classroom instruction but cannot be used to draw conclusions about the impact of RAMP on lesson content or instructional practices. Again, baseline data were not collected that would be needed to assess equivalency before the introduction of RAMP. Therefore, this report focuses on EQ.1.2, for which the evaluation team used baseline and endline classroom observations to estimate RAMP impacts.

EQ 1.2. – WHAT ARE THE IMPACTS OF RAMP ON G3 TEACHERS' INSTRUCTIONAL PRACTICES IN READING AND MATH?

Classroom observation data were also collected for a sample of G3 intervention and comparison teachers to assess their instructional practices at baseline, prior to RAMP training, and after a year of the intervention. The analysis of G3 teachers' instruction does not indicate that trained teachers regularly implemented improved instructional practices in line with the RAMP training.

At baseline, before RAMP was implemented, G3 intervention and comparison teachers were similar in most characteristics and instructional practices (see Annex J). By endline, nearly one year after G3 teachers were trained, intervention teachers were more likely to use RAMP materials and some RAMP

strategies than comparison teachers. **RAMP trained** teachers allocated more lesson time to G3-appropriate content, such as vocabulary, compared to teachers that were not trained. However, there were no differences in the amount of classroom time spent on more sophisticated content such as reading comprehension and advanced arithmetic skills. **While RAMP teachers provided higher quality feedback during math lessons**, there were no other significant impacts on teachers' instructional practices, even when examining a wide range of outcomes. Table 1 summarizes the outcomes that had significant impacts in math and reading (out of the total number that were measured), and the direction of each impact (negative or positive).

TABLE 5. SUMMARY OF INSTRUCTIONAL PRACTICES AT ENDLINE

During reading lessons, RAMP teachers:	During math lessons, RAMP teachers:
Implemented RAMP in the Classroom.	
+Implemented RAMP strategies +Integrated RAMP and MoE curriculum in lessons	+Used RAMP worksheets during lessons +Implemented RAMP strategies +Integrated RAMP and MoE curriculum in lessons
Shifted some lesson time away from basic skills towards other content.	
-Were less likely to include writing activities in the lesson -Spent less lesson time on reading and identifying written characters +Spent more lesson time on vocabulary	-Were less likely to cover number writing and identification during the lesson -Spent less lesson time on number writing and identification
Had few impacts on classroom management and student engagement.	
No significant impacts on student engagement. +Were more likely to use whole-class instruction	+Provided better feedback on student participation and written work *No significant impacts on classroom structure.*

Note: Red font indicates a negative impact and blue font indicates a positive impact.

The following sections describe results from a teacher survey and classroom observations of G3 teachers. Additional details on teacher characteristics are provided in Annex J.

RESULTS FROM THE TEACHER SURVEY: RAMP IMPLEMENTATION

G3 intervention and some comparison teachers that were observed and surveyed at each data collection interval reported receiving **RAMP training and coaching.**

- 86 percent of G3 intervention teachers reported receiving RAMP training, as did 17 percent of comparison teachers.
    - Comparison teachers, however, reported significantly fewer days of training and fewer coaching sessions.
    - The comparison teachers may be confusing a short RAMP orientation, implemented countrywide, with the full training. They may also have participated in training earlier than intended.
- More than 80 percent of trained intervention teachers and five percent of comparison teachers reported that a **RAMP coach observed their lesson** during the current school year.
    - Most intervention teachers reported fewer mentoring sessions than RAMP had intended to provide based on the implementation plan, which specified six visits per intervention teacher. Only 20 percent of teachers reported four to six visits and 5 percent of teachers reported six or more visits. Fifty-five percent of intervention and 5 percent of comparison teachers reported between **one and three days of coaching or mentoring per academic year**.

Teachers self-reported that they used the **RAMP assessment tools.**

- Most intervention teachers (nearly 90 percent) reported having used the **coarse-grain assessment** tool in math and reading in the current school year, compared to about half of teachers in the comparison group. It is not clear whether all teachers were familiar with the RAMP tools or if there was confusion with MoE curriculum materials.
- Slightly more than 70 percent of intervention teachers reported using the corresponding **fine-grain assessment tools**, compared to about 20 percent of teachers in the comparison group. Differences in take-up of the two tools may be consistent with the guidance provided by RAMP that teachers should implement the fine grain tool only with students who performed poorly on the coarse-grain tool.

Most trained teachers in both groups were able to print out RAMP materials when needed, despite some initial challenges that were reported earlier in the implementation process. The reported spillover of RAMP training, coaching, and assessments from intervention to comparison teachers, although modest, could reduce the contrast between the intervention and comparison groups and make it more difficult to detect the intervention's impacts during observations (see Annex Q). Still, the small number of coaching visits that both intervention and comparison teachers received suggests that implementation goals were not met, rather than spillover of the intervention to comparison teachers.

TABLE 6. TEACHER SELF-REPORTS OF RAMP IMPLEMENTATION

Variable	Intervention (T)	Comparison (C)	Impact (T-C) *	*p*-value	Effect Size	Number of Teachers
RAMP Training (%)						
Received training from RAMP	86.2	17.1	69.1*	0.000	2.06	200
4-5 days of training	0	5.0	-5.0*	0.024	NA	200
6-7 days of training	3.3	0.9	2.4	0.24	0.82	200
More than 7 days of training	82.9	11.3	71.6*	0.000	2.2	200
RAMP Coaching (%)						
RAMP coach observed classroom this school year (%)	80.1	5.3	74.9*	0.000	2.59	200
Coach observed lesson 1-3 times	55.4	5.3	50.1*	0.000	1.87	200
Coach observed lesson 4-6 times	19.9	0	19.9*	0.000	NA	200
Coach observed lesson more than 6 times	4.9	0	4.9*	0.024	NA	200
Used RAMP Assessment Tools						
RAMP math coarse-grain tool	89.3	51.5	37.8*	0.000	1.24	200
RAMP reading coarse-grain tool	89.3	49.5	39.8*	0.000	1.29	200
RAMP math fine-grain tool	71.7	22.8	48.9*	0.000	1.3	200
RAMP reading fine-grain tool	73.7	23.8	49.8*	0.000	1.32	200
Trained teachers able to print worksheets for RAMP	80.5	76.9	3.6	0.75	0.13	103

Source: RAMP Impact Study - Endline Data 2018 COTI Tool

Note: Columns T and C present ordinary least squares regression-adjusted means (or percentages), with weights accounting for school sampling probabilities. Effect sizes were calculated as the Cox Index. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. NA stands for Not Applicable. Effect sizes are "NA" when the standard deviation for one or both groups equals zero.

*Difference in differences is statistically significant at the .05 level.

RESULTS FROM READING LESSON OBSERVATIONS: RAMP IMPLEMENTATION, READING LESSON CONTENT, AND INSTRUCTIONAL PRACTICES

RAMP Implementation and Impacts on Lesson Content

At endline, G3 intervention teachers demonstrated few significant differences in instructional practices during reading lessons compared to teachers that did not participate in RAMP training. While there were some differences, there were no clear patterns indicating that the program was implemented with fidelity. First, G3 intervention teachers were significantly more likely than comparison teachers to use a range of RAMP strategies during the lesson, such as asking reading comprehension questions, summarizing text after each paragraph, and integrating RAMP strategies with MoE lessons and strategies.

RAMP had a large, statistically significant, negative impact of 18 percentage points on whether teachers included **writing activities** at any point in the lesson, and a large but insignificant effect on character reading and identification. The reductions in the percent of lessons spent on writing relative to the comparison group happened across most of the types of writing that were examined, including writing words, sentences and phrases, practicing grammar and spelling, and functional writing (See Annex K, Table K.1). While interpretation of this finding is unclear, this shift away from writing may have been due to teachers using more interactive methods of practice and instruction and other G3-appropriate foundational skills such as vocabulary (see Table 7).

TABLE 7. RAMP IMPLEMENTATION AND LESSON CONTENT DURING READING LESSONS

Variable	T Group			C Group			RAMP Impacts*	Effect Size
	Baseline Mean (T_1)	Endline Mean (T_2)	Change Over Time (T_2-T_1)	Baseline Mean (C_1)	Endline Mean (C_2)	Change Over Time (C_2-C_1)	(T_2-T_1) - (C_2-C_1)	
RAMP Implementation								
Teachers or students used a RAMP worksheet (%)	1.5	4.6	3.2	0.6	1.8	1.2	2.0	0.05
Teacher demonstrated RAMP strategies *(0: not used -2: multiple strategies used)*	0.3	1.7	1.4	0.2	1.1	0.9	0.4*	0.53
Teacher integrated MoE and RAMP strategies *(0: not used -2: multiple strategies used)*	0.3	1.7	1.4	0.2	1.1	1.0	0.4*	0.55
Lesson Content (% of lessons including a content area)								
Character reading and identification	35.7	10.2	-25.5	40.1	21.5	-18.6	-6.9	-0.42
Phonemic awareness	59.8	47.2	-12.6	60.6	45.5	-15.0	2.5	0.05
Vocabulary	61.1	64.6	3.5	56.6	52.2	-4.4	7.9	0.21
Writing	92.8	81.5	-11.3	87.6	94.7	7.1	-18.4*	-1.20
Reading comprehension	59.6	64.4	4.9	52.2	58.1	5.9	-1.1	-0.03
Number of schools	100	100		100	100			
Number of lessons	135	106		161	109			

Source: RAMP Impact Study – Midline and Endline Data, 2017/2018 COTI Tool.

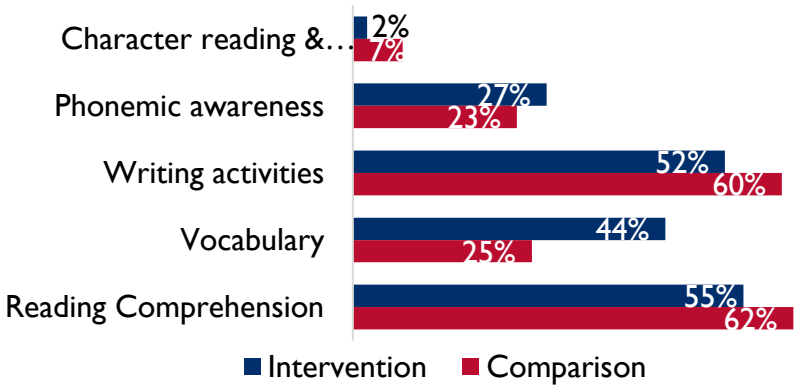
Note: Columns T_1, T_2, C_1, and C_2 present ordinary least squares regression-adjusted group means (or percentages) at each time point, with weights accounting for school sampling probabilities. The RAMP Impacts column reflects the coefficient and significance of the difference-in-differences estimator. Errors are clustered at the classroom level to account for the clustering of observations within some classrooms. Teachers were rated on the frequency with which they demonstrated RAMP strategies and integrated RAMP and MoE strategies on a scale from 0 to 2, where 0 indicates that strategies were not used, and 2 indicates that multiple strategies were used or integrated throughout the lesson. Effect sizes are the difference in effect sizes between endline and baseline, calculated using Hedges' g for continuous outcomes and the Cox index for dichotomous outcomes. Content areas covered during a lesson are not mutually exclusive.

*Difference in differences is statistically significant at the .05 level.

At endline, lesson content during G3 classroom observations was recorded in 15-minute increments. This allows more nuanced descriptions of the use of instructional time than the results in Table 7. However, it does not permit direct comparisons with baseline results. Therefore, this report will present both sets of results for reading and math lessons.

When comparing the percent of 15-minute segments that included a given content area, RAMP statistically significantly reduced the number of segments that included **identifying and reading written characters** (as opposed to words and sentences) by five percentage points (Figure 5). In contrast, RAMP increased the amount of time spent on building students' **vocabulary**. For example, additional analysis of the specific topics in each category showed that G3 intervention teachers were 23 percentage points more likely than comparison teachers to cover word families during the lesson (see Table K.1 in Annex K).

FIGURE 5. PERCENT OF LESSON TIME G3 TEACHERS SPENT ON READING LESSON CONTENT CATEGORIES.



This figure shows the average proportion of 15-minute lesson segments that included a content area. Content areas covered during a segment were not mutually exclusive.

Source: *RAMP Impact Study - Endline Data 2018 COTI Tool.*

Note: Figure presents ordinary least squares regression-adjusted group percentages. All regressions include school-level propensity score weights accounting for school sampling probabilities.

*Difference in group means is statistically significant at the .05 level.

RAMP IMPACTS ON TEACHERS' INSTRUCTIONAL PRACTICES

RAMP did not have a statistically significant impact on any of the instructional practices of G3 teachers measured during reading lessons. Some outcomes, such as teacher engagement and student participation, showed substantively large but not statistically significant trends (negative and positive respectively) in the intervention compared to the comparison group.

TABLE 8. INSTRUCTIONAL PRACTICE SCORES DURING READING LESSONS

Variable	T Group			C Group			RAMP Impacts* (T2-T1)-(C2-C1)	Effect Size
	Baseline Mean (T_1)	Endline Mean (T_2)	Change Over Time (T_2-T_1)	Baseline Mean (C_1)	Endline Mean (C_2)	Change Over Time (C_2-C_1)		
Percent of time on task	96.5	96.9	0.5	95.3	94.1	-1.2	1.7	0.21
Quality instructional practices	86.4	90.8	4.5	80.1	86.0	5.9	-1.4	0.07
Teacher engagement practices	35.8	43.3	7.6	34.3	46.2	12.0	-4.4	-0.27
Teacher encouragement	48.7	54.1	5.4	51.7	52.5	0.8	4.6	0.12
Teacher feedback	38.6	40.8	2.2	35.1	35.9	0.9	1.3	0.06
Materials to support instruction	14.3	21.0	6.7	8.7	15.3	6.6	0.0	-0.05
Assessment to support instruction	31.9	45.0	13.1	39.1	50.0	10.9	2.1	0.14
Differential instruction	28.9	40.0	11.1	30.4	38.0	7.5	3.5	0.14
Student engagement	67.8	68.4	0.6	68.3	66.0	-2.3	2.8	0.08
Student participation	16.4	28.8	12.4	11.8	13.8	2.0	10.4	0.27
Positive classroom climate	65.9	74.9	9.0	64.2	69.6	5.3	3.7	0.18
Negative classroom climate	14.6	28.4	13.8	18.7	33.3	14.5	-0.7	-0.01
Number of Schools	**100**	**100**		**100**	**100**			
Number of Lessons	**135**	**106**		**161**	**109**			

Source: RAMP Impact Study – Midline and Endline Data, 2017/2018 COTI Tool

Note: Columns T_1, T_2, C_1, and C_2 present ordinary least squares regression-adjusted group means (or percentages) at each time point, with weights accounting for school sampling probabilities. The RAMP Impacts column reflects the coefficient and significance of the difference-in-differences estimator. Errors are clustered at the classroom level to account for the clustering of observations within some classrooms. Effect sizes are the difference in effect sizes between endline and baseline, calculated using Hedges' g for continuous outcomes and the Cox index for dichotomous outcomes. (T = intervention group, and C = comparison group).

*Difference in differences is statistically significant at the .05 level.

Intervention teachers increased the proportion of lessons that incorporated work in **small groups** by 16 percentage points relative to the comparison group. This difference was only significant at the 10 percent level (p=0.07), but it is worth noting because it is a substantively large difference and many RAMP training videos demonstrate students working in small groups to build foundational skills (see Table 9).

TABLE 9. CLASS SIZE AND STRUCTURE DURING READING LESSONS FOR GRADE 3 TEACHERS

	T Group			C Group				
Variable	**Baseline Mean (T_1)**	**Endline Mean (T_2)**	**Change Over Time (T_2-T_1)**	**Baseline Mean (C_1)**	**Endline Mean (C_2)**	**Change Over Time (C_2-C_1)**	**RAMP Impacts* (T_2-T_1)- (C_2-C_1)**	**Effect Size**
Class Size	24.3	23.4	-0.9	23.1	21.9	-1.2	0.3	0.03
Whole Group/Whole Class	97.8	99.1	1.3	99.4	91.8	-7.6	**8.8***	2.14
Large group (more than half the class)	2.2	15.0	12.8	0	13.7	13.7	-0.9	NA
Small group (half or less than half the class)	22.6	42.4	19.9	21.5	25.7	4.3	15.6	0.42
Partners/Pairs	20.5	25.9	5.4	10.4	14.0	3.6	1.8	-0.02
Individual	100.0	96.3	-3.7	100.0	93.3	-6.7	2.9	NA
Number of Schools	**100**	**100**		**100**	**100**			
Number of Lessons	**135**	**106**		**161**	**109**			

Source: RAMP Impact Study – Midline and Endline Data, 2017/2018 COTI Tool.

Note: Columns T_1, T_2, C_1, and C_2 present ordinary least squares regression-adjusted group means (or percentages) at each time point, with weights accounting for school sampling probabilities. The RAMP Impacts column reflects the coefficient and significance of the difference-in-differences estimator. Errors are clustered at the classroom level to account for the clustering of observations within some classrooms. Effect sizes are the difference in effect sizes between endline and baseline, calculated using Hedges' g for continuous outcomes and the Cox index for dichotomous outcomes.

*Difference in differences is statistically significant at the .05 level.

RESULTS FROM MATH LESSON OBSERVATIONS

RAMP IMPLEMENTATION AND IMPACTS ON LESSON CONTENT

Overall, like the findings on reading lessons, the observations of math lessons yield mixed results such that RAMP had both positive and negative impacts on teachers' math instruction. These differences are difficult to interpret because of the size and direction of the impacts but suggest that teachers may not have implemented RAMP strategies as the program intended.

Similar to the results for reading lessons, RAMP had a significant positive impact on G3 teachers' use of RAMP strategies (for example, introducing teaching multiplication as a process of repeated addition; see Annex K) during the lesson and on the integration of RAMP strategies with MoE lessons and strategies. Math lessons were also more likely than reading lessons to include the use of a RAMP-branded worksheets by students or teachers; 12 percent of lessons used RAMP worksheets, compared to only one percent in the comparison group (see Table 10).

RAMP significantly reduced the percent of lessons that covered number identification and writing by more than 30 percentage points, relative to the comparison group. An examination of the specific topics in this category showed that lessons were less likely to include writing numbers on the board or on paper (see Table K.2 in Annex K). Since G3 students in both groups scored highly on the number identification EGMA task (see Table 21 in the next section), this may indicate that assessment tools helped teachers to tailor lesson content to students' skills and abilities.

TABLE 10. RAMP IMPLEMENTATION AND LESSON CONTENT DURING MATH LESSONS

Variable	T Group			C Group			RAMP Impacts* (T2-T1)-(C2-C1)	Effect Size
	Baseline Mean (T_1)	Endline Mean (T_2)	Change Over Time (T_2-T_1)	Baseline Mean (C_1)	Endline Mean (C_2)	Change Over Time (C_2-C_1)		
RAMP Implementation								
Teachers or students used a RAMP worksheet (% of lessons)	2.7	12.3	9.6	2.0	1.0	-1.0	10.6**	1.41
Teacher demonstrated RAMP strategies *(0: not used -2: multiple strategies used)*	0.2	1.8	1.6	0.2	1.2	1.0	0.7**	1.05
Teacher integrated MoE and RAMP strategies *(0: not used -2: multiple strategies used)*	0.2	1.8	1.6	0.2	1.2	1.0	0.6**	0.95
Lesson Content (% of lessons including a content area)								
Number identification and writing	70.8	42.7	-28.1	63.4	67.9	4.5	-32.6*	-0.83
Counting	48.3	45.1	-3.1	44.3	50.5	6.2	-9.3	-0.23
Basic arithmetic	44.0	67.7	23.7	33.1	55.0	21.9	1.8	0.05
Advanced arithmetic	70.0	64.6	-5.4	63.2	57.0	-6.3	0.8	0.01
Number of schools	100	100		100	100			
Number of lessons	100	108		101	104			

Source: RAMP Impact Study – Midline and Endline Data, 2017/2018 COTI Tool.

Note: Columns T_1, T_2, C_1, and C_2 present ordinary least squares regression-adjusted group means (or percentages) at each time point, with weights accounting for school sampling probabilities. The RAMP Impacts column reflects the coefficient and significance of the difference-in-differences estimator. Errors are clustered at the classroom level to account for the clustering of observations within some classrooms. Teachers were rated on the frequency with which they demonstrated RAMP strategies and integrated RAMP

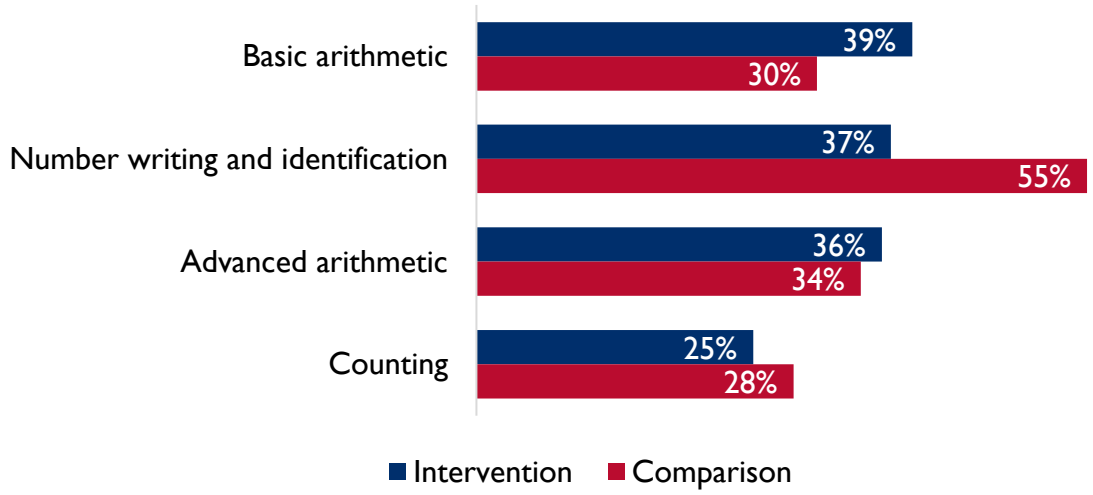
and MoE strategies on a scale from 0 to 2, where 0 indicates that strategies were not used, and 2 indicates that multiple strategies were used or integrated throughout the lesson. Effect sizes are the difference in effect sizes between endline and baseline, calculated using Hedges' g for continuous outcomes and the Cox index for dichotomous outcomes. Content areas covered during a lesson are not mutually exclusive.

*Difference in group means is statistically significant at the .05 level

When examining the percent of lesson time spent on each topic, measured in terms of 15-minute lesson segments that included different content areas, the intervention group also spent less time (a difference of 18 percentage points) covering **number identification and writing** than the comparison group (see Figure 6). Intervention teachers spent nine percentage points more lesson time on **basic arithmetic skills**, although this was only significant at the 10 percent level (p=0.08). The likely interpretation is that intervention teachers were focused on more grade-appropriate skills.

This figure shows the percent of 15-minute lesson segments spent on a content area across all observed lessons. Content areas during a given lesson segment are not mutually exclusive.

FIGURE 6. PERCENT OF LESSON TIME SPENT ON MATH LESSON CONTENT CATEGORIES FOR TEACHERS IN GRADE 3.



Source: RAMP Impact Study – Midline and Endline Data, 2017/2018 COTI Tool.

Note: Figure presents ordinary least squares regression-adjusted group percentages. All regressions include school-level propensity score weights accounting for school sampling probabilities.

*Difference in group means is statistically significant at the .05 level.

RAMP IMPACTS ON TEACHERS' INSTRUCTIONAL PRACTICES

RAMP had a positive impact on the quality of **feedback** teachers provided during math lessons (Table 11). Teachers in the intervention group provided more specific feedback, such as "You used those counters correctly to solve the problem," and more strategic feedback, such as "How did you figure out that the answer was 6?" (see Table K.3 in Annex K). These types of feedback help students learn how to solve problems correctly, rather than merely confirming whether an answer is correct or incorrect. Other instructional practices, such as **student participation** and **negative classroom climate** also showed large, promising trends, but the differences were not statistically significant. There was also a large but insignificant decrease in **teacher engagement practices** such as asking questions and emphasizing

content, due to large increases in the comparison group and relatively small growth in the intervention group.

TABLE 11. INSTRUCTIONAL PRACTICE SCORES DURING MATH LESSONS

Variable	T Group Baseline Mean (T_1)	T Group Endline Mean (T_2)	T Group Change Over Time (T_2-T_1)	C Group Baseline Mean (C_1)	C Group Endline Mean (C_2)	C Group Change Over Time (C_2-C_1)	RAMP Impacts* (T_2-T_1)-(C_2-C_1)	Effect Size
Percent of time on task	95.5	95.7	0.2	94.6	92.9	-1.6	1.9	0.21
Quality instructional practices	87.8	93.0	5.2	87.9	89.7	1.8	3.5	0.24
Teacher engagement practices	42.2	44.2	2.0	44.5	51.5	7.0	-5.0	-0.27
Teacher encouragement	49.0	58.1	9.1	52.2	55.4	3.2	5.9	0.17
Teacher Feedback	45.9	59.4	13.5	49.7	46.9	-2.8	**16.3***	0.63
Materials to support instruction	31.1	30.4	-0.7	23.1	24.6	1.5	-2.1	-0.06
Assessment to support instruction	34.1	48.3	14.2	39.8	53.8	14.0	0.2	-0.02
Differential instruction	29.0	40.3	11.3	33.1	39.7	6.6	4.7	0.19
Student engagement	77.2	79.0	1.7	81.2	80.6	-0.6	2.3	0.07
Student participation	30.5	38.8	8.3	23.1	20.7	-2.4	10.7	0.26
Positive classroom climate	65.4	73.0	7.6	66.2	72.0	5.8	1.7	0.09
Negative classroom climate	16.6	25.6	9.0	17.4	34.7	17.3	-8.4	-0.31
Number of Schools	**100**	**100**		**100**	**100**			
Number of Lessons	**108**	**101**		**104**	**100**			

Source: RAMP Impact Study – Midline and Endline Data, 2017/2018 COTI Tool

Note: Columns T_1, T_2, C_1, and C_2 present ordinary least squares regression-adjusted group means (or percentages) at each time point, with weights accounting for school sampling probabilities. The RAMP Impacts column reflects the coefficient and significance of the difference-in-differences estimator. Errors are clustered at the classroom level to account

for the clustering of observations within some classrooms. Effect sizes are the difference in effect sizes between endline and baseline, calculated using Hedges' g for continuous outcomes and the Cox index for dichotomous outcomes.

*Difference in group means is statistically significant at the .05 level

Although the intervention group showed an increase of 13 percentage points in the number of lessons that incorporated **small group** work relative to the comparison group, this difference was not statistically significant, and was partially offset by a decrease in lessons that included partner/pair work (Table 12).

TABLE 12. CLASS SIZE AND STRUCTURE DURING MATH LESSONS

Variable	T Group Baseline Mean (T_1)	T Group Endline Mean (T_2)	T Group Change Over Time (T_2-T_1)	C Group Baseline Mean (C_1)	C Group Endline Mean (C_2)	C Group Change Over Time (C_2-C_1)	RAMP Impacts* (T_2-T_1)-(C_2-C_1)	Effect Size
Class Size	25.0	23.5	-1.5	22.8	21.9	-0.9	-0.6	-0.08
Whole Group/Whole Class	100.0	95.7	-4.3	99.1	95.1	-4.0	-0.3	NA
Large Group (more than half the class)	4.6	21.4	16.8	0.9	11.9	11.0	5.9	-0.56
Small Group (half or less than half the class)	37.3	44.6	7.4	34.7	29.2	-5.5	12.8	0.34
Partners/Pairs	23.9	20.9	-3.0	12.0	17.9	6.0	-9.0	-0.39
Individual	99.1	96.1	-3.0	99.1	94.1	-5.0	2.0	0.24
Number of Schools	**100**	**100**		**100**	**100**			
Number of Lessons	**108**	**101**		**104**	**100**			

Source: RAMP Impact Study – Midline and Endline Data, 2017/2018 COTI Tool.

Note: Columns T_1, T_2, C_1, and C_2 present ordinary least squares regression-adjusted group means (or percentages) at each time point, with weights accounting for school sampling probabilities. The RAMP Impacts column reflects the coefficient and significance of the difference-in-differences estimator. Errors are clustered at the classroom level to account for the clustering of observations within some classrooms. Effect sizes are the difference in effect sizes between endline and baseline, calculated using Hedges' g for continuous outcomes and the Cox index for dichotomous outcomes.

*Difference in group means is statistically significant at the .05 level

EVALUATION QUESTIONS 2 AND 3

EQ 2.0 AND 3.0– WHAT ARE THE IMPACTS OF RAMP ON G1, G2, AND G3 STUDENTS' PROFICIENCY IN READING AND MATH? DO THE IMPACTS VARY BY GENDER AND BY THE NUMBER OF SCHOOL SHIFTS?

Overall, the student assessments, using EGRA and EGMA tools to assess literacy and numeracy, indicate there were few impacts of RAMP on G1 and G2 students in reading and math outcomes over the course of two school years of RAMP implementation. This pattern of minimal impacts for both G1 and G2 students at midline and endline is consistent. These findings are not surprising, given the minimal impacts on teachers' instructional practices.

EVALUATION QUESTIONS 2 AND 3: SUMMARY OF KEY FINDINGS

- Endline analysis revealed few substantive differences in reading and math proficiency between students in the intervention and comparison groups, after approximately one and a half years of RAMP implementation. This is consistent with the results from midline analyses.
- For G1 students in T, two years of exposure to RAMP (compared to one year of exposure for students in C) had a negative statistically significant impact on **letter-sound knowledge** and no impact on other skills. RAMP also had a negative impact on addition and no impact on other mathematics skills.
- For G2 students in T, two years of exposure to RAMP (compared to no exposure for students in C) had a positive impact on students' ability to **segment words into syllables**. However, as with G1 students, it had a negative impact on students' **knowledge of letter-sound correspondence.** RAMP had no impacts on G2 students' mathematics skills.
- With a few exceptions, there were no statistically significant differences in RAMP impacts between boys and girls on either reading or math. There was, however, a tendency for RAMP to have some negative impacts for students in single-shift schools and positive impacts for students in double-shift schools. Most of these differences were not statistically significant.

Table 13 summarizes the average impacts of RAMP on G1 and G2 students' reading and math scores, as well as differences in impacts by students' gender and the number of school shifts, at both midline and endline. Detailed results are described in the remainder of this section. Results on students' home learning environments and reading habits are provided in Annex L.

TABLE 13. SUMMARY OF RAMP READING AND MATH IMPACTS AT MIDLINE AND ENDLINE, BY GRADE

READING	Midline		Endline	
	G1 at baseline and midline	G2 at baseline and midline	G1 at base- and midline; G2 at endline	G2 at base- and midline; G3 at endline
1. Orientation to Print	No impact	NA	NA	NA
2. Phoneme Isolation	No impact	NA	No impact	NA
3. **Syllable Segmentation**	Positive impact [S]	Positive impact [S]	No impact	Positive impact
4. **Letter Sound Knowledge**	No impact	No impact	Negative impact	Negative impact
5. Non-Word Decoding	NA	No impact [S]	NA	No impact
6. Reading Vocabulary	No impact [S]	No impact	No impact [S]	No impact
7. **Passage Reading**	Positive impact [S]	No impact [S]	No impact [S]	No impact
8. **Reading Comprehension**	No impact [G]	No impact	No impact [S]	No impact
MATHEMATICS	**Midline**		**Endline**	
	G1 at baseline and midline	**G2 at baseline and midline**	**G1 at base- and midline; G2 at endline**	**G2 at base- and midline; G3 at endline**
1. Counting Numbers	Negative impact [S]	NA	No impact	NA
2. Counting Objects (or enumerating quantities)	No impact	NA	No impact	NA
3. Number Identification	No impact	No impact [G]	No impact	No impact
4. Number Discrimination	No impact	No impact	No impact	No impact
5. Missing Numbers	No impact	No impact	No impact	No impact [S]
6. Addition Facts - L1	No impact	No impact	Negative impact	No impact
7. Addition Facts – L2	NA	No impact	NA	No impact
8. Subtraction Facts - L1	NA	No impact	NA	No impact [S]
9. Subtraction Facts – L2	NA	No impact	NA	No impact [S]

Note: G denotes different impacts of RAMP for boys versus girl. S denotes different impacts of RAMP for students in single- versus double-shift schools. See details about differences below. L1 and L2 denote Levels 1 and 2, respectively, for both addition and subtraction facts.

EGRA RESULTS: READING BY GRADE, GENDER, AND NUMBER OF SCHOOL SHIFTS

Grade 1 Reading Results

At the time of endline data collection, students in G1 at baseline (who had advanced to G2 by endline) were exposed to RAMP for approximately two years, while G1 students in the comparison group were exposed for about one year. Therefore, impact estimates for G1 students reflect the contrast between one and two years of RAMP, instead of the contrast between two years and no exposure.

Students who were in G1 at baseline were assessed on six early grade reading skills, four of which were foundational literacy skills required to master fluent reading and comprehension (Kim, Boyle, Simmons Zuilkowski, & Nakamura, 2017): (1) ability to segment a word into individual phonemes and (2) syllables;

(3) knowledge of letter-sound correspondence; and (4) reading vocabulary. Students were assessed also on their ability to read fluently and comprehend a short passage.

Two years of exposure to RAMP (compared to one year of exposure) had a negative, statistically significant impact on students' **letter-sound knowledge** and no impact on other skills.

- G1 intervention students (who were in G2 at endline) identified about 30 sounds correctly on average, whereas comparison students identified an average of 34 sounds (a difference of about four sounds, equivalent to -0.2 standard deviations; see Table 14).
- RAMP did not have a statistically significant impact on the average G1 student's ability to segment words into phonemes or syllables, reading vocabulary, or reading fluency and comprehension.
- RAMP also had a negative impact on the percentage of students who obtained a **score equal to zero** in the **phoneme isolation** and **oral passage reading** subtasks, and no impact on reading comprehension. The two groups significantly differed by three percentage points in phoneme isolation and six percentage points in passage reading (see Table M.1 in Annex M).
- Figures 7 and 8 are an illustration of change in students' scores across the three time points of the study. In general, the figures suggest that both groups of students have similar trajectories from G1 to G2 (see Annex O for figures on other outcomes).

TABLE 14. ENDLINE READING PERFORMANCE SCORES FOR GRADE 1 STUDENTS

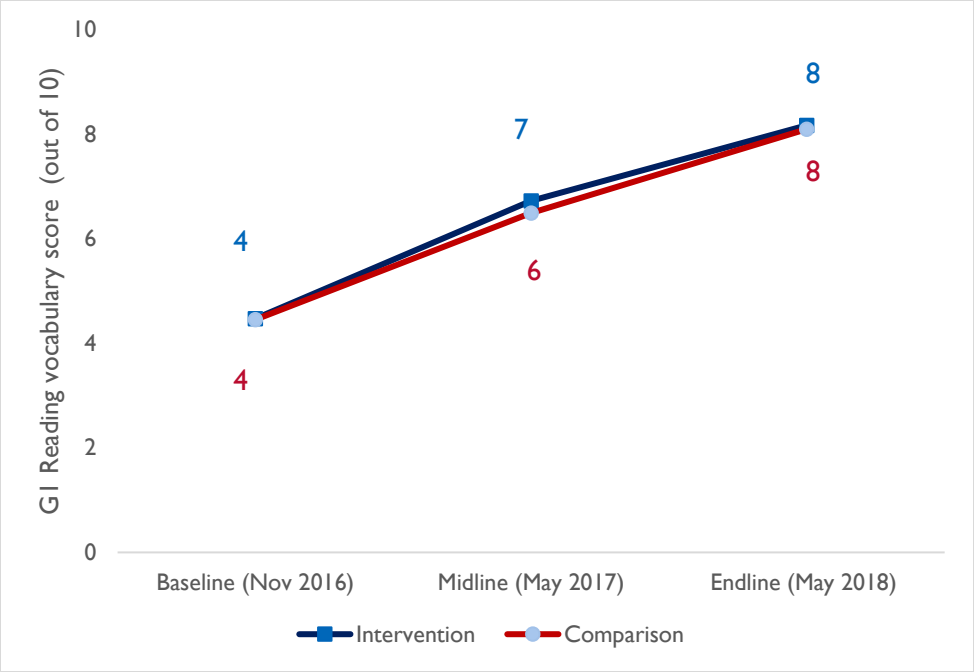
Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C) *	*p*-value	Effect Size	Number of Students
Phoneme Isolation (out of 10)	4.6	4.9	-0.3	0.19	-0.1	1931
Syllable Segmentation (out of 10)	6.8	6.6	0.3	0.34	0.1	1931
Letter Sound Knowledge (out of 100, prorated)	30.5	34.3	**-3.8***	0.008	-0.2	1931
Reading Vocabulary (out of 10)	8.2	8.1	0.0	0.79	0.0	1931
Passage Reading (out of 41, prorated score)	14.6	14.8	-0.2	0.82	0.0	1931
Reading Comprehension (out of 6)	1.5	1.7	-0.2	0.16	-0.1	1931
Number of Schools	**117**	**120**				

Source: RAMP Impact Study - Endline Data 2018 Student Assessments

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Students' home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes are calculated as Hedge's *g*. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated to the midline scale.

*Difference in group means is statistically significant at the .05 level.

FIGURE 7. READING VOCABULARY OVER TIME, FOR STUDENTS WHO WERE IN G1 AT BASELINE



Source: *RAMP Impact Study - Endline Data 2018 Student Assessments.* The figure shows raw baseline scores, and midline and endline scores equated into the baseline scale.

The MoE set the target that only 13 percent of G2 and G3 students were to obtain a score of zero in **reading comprehension** by 2019. RAMP did not have a statistically significant impact on this outcome. In both the intervention and comparison groups, about 40 percent of students who were in G2 at endline[12] were still unable to answer correctly a single **reading comprehension** question (see Table M.1 in Annex M). Supplementary descriptive analysis showed that G1 students who read with comprehension (80 percent correct in the reading comprehension subtask) read 40 correct words per minute on average, while students who did not read with comprehension read only 17 correct words per minute. This level of fluency is far below the benchmark of 46 words per minute set by the MoE in 2014 (Brombacher et al., 2015).

12 These students were in G1 at baseline.

FIGURE 8. GRADE 1 STUDENT READING RESULTS: INTERVENTION (RAMP) AND COMPARISON SCHOOLS

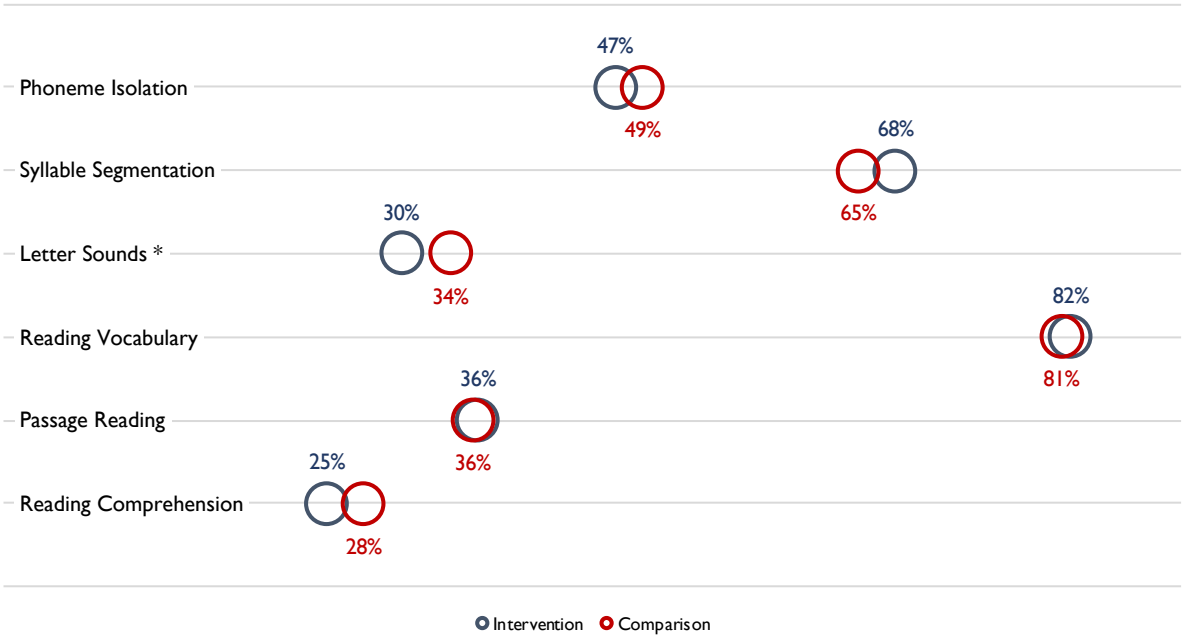
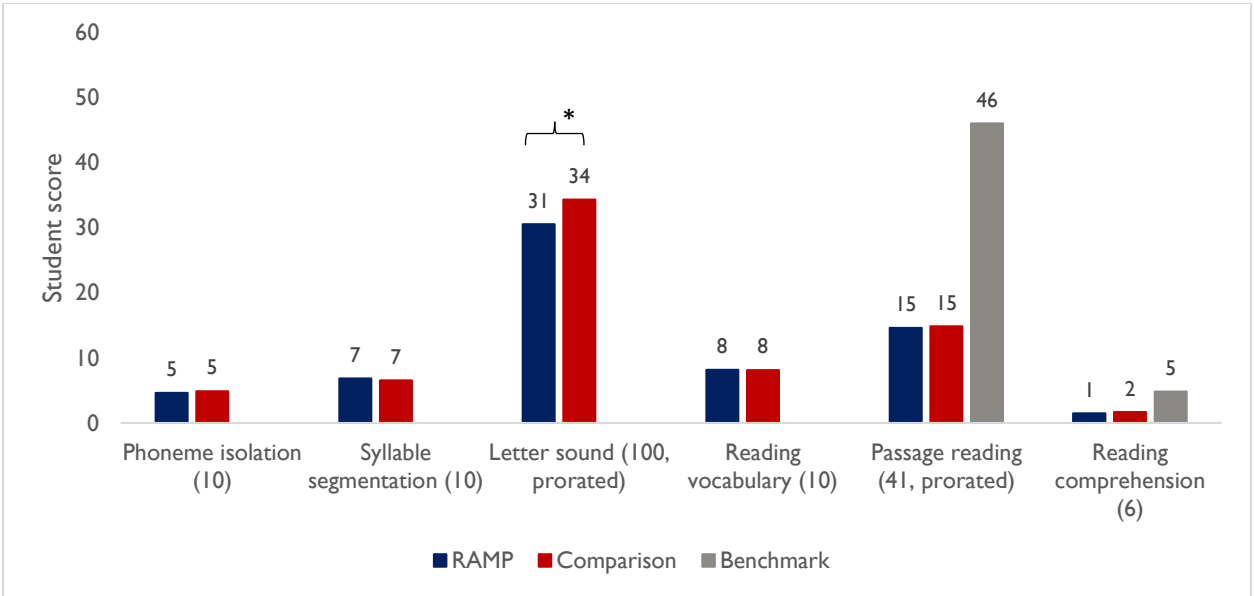


Figure 9 shows that a comparison of RAMP and comparison Grade 1 reading scores features only one competence area in which there is a statistically significant difference between the two groups, and this difference shows higher scores for students in the comparison schools than those in the RAMP schools.

FIGURE 9. GRADE 1 STUDENT SCORES BY READING SUBTASK, INTERVENTION (RAMP) AND COMPARISON SCHOOLS



Source: *RAMP Impact Study - Endline Data 2018 Student Assessments.* The figure shows endline scores equated into the baseline scale. The number of items administered at baseline are shown in parenthesis.

* Difference in group means is statistically significant at the 0.05 level.

Grade 2 Reading Results

This section summarizes impacts of two years of exposure to RAMP, versus no exposure, for students who were in G2 at baseline and G3 at endline. Results showed that two years of exposure to RAMP had a positive impact on students' ability to segment words into syllables. However, as with G1 students, it had a negative impact on students' knowledge of letter-sound correspondence.

- G2 intervention students (who were in G3 at endline) were able to correctly segment nearly one more word into syllables (Figure 10) on average than comparison students (a difference equivalent to 0.2 standard deviations).
- However, G2 intervention students identified three fewer letter sounds than comparison students (a difference of -0.1 standard deviations; see Table 15).

TABLE 15. ENDLINE READING PERFORMANCE SCORES FOR GRADE 2 STUDENTS

Variable (Total # of items)	Intervention (T)	Comparison (C)	Impact (T-C) *	*p*-value	Effect Size	Number of Students
Syllable Segmentation (out of 10)	6.0	5.2	0.8*	0.007	0.2	1931
Letter Sound Knowledge (out of 100, prorated)	33.6	36.7	-3.2*	0.025	-0.1	1931
Non-Word Decoding (out of 50, prorated)	12.6	11.8	0.8	0.27	0.1	1931
Reading Vocabulary (out of 10)	9.1	9.2	0.0	0.78	0.0	1931
Passage Reading (out of 41, prorated score)	22.5	23.1	-0.6	0.49	0.0	1931
Reading Comprehension (out of 6)	2.7	2.8	0.0	0.64	0.0	1931
Number of Schools	118	119				

Source: RAMP Impact Study - Endline Data 2018 Student Assessments

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Students' home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes are calculated as Hedge's g. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated to the midline scale.

*Difference in group means is statistically significant at the .05 level.

RAMP did not have a significant impact on the percentage of students who obtained a **score equal to zero** (i.e., students unable to answer a single question correctly) in each of the six subtasks.

In both groups, fewer than 20 percent of students obtained a zero-score in all subtasks, except for **non-word decoding** and **reading comprehension**. About a third of students in both groups were unable to decode a single invented word correctly and a little over 20 percent were unable to answer a single reading comprehension question correctly (see Table M.2 in Annex M).

FIGURE 10. GRADE 2 STUDENT READING RESULTS: INTERVENTION (RAMP) AND COMPARISON SCHOOLS

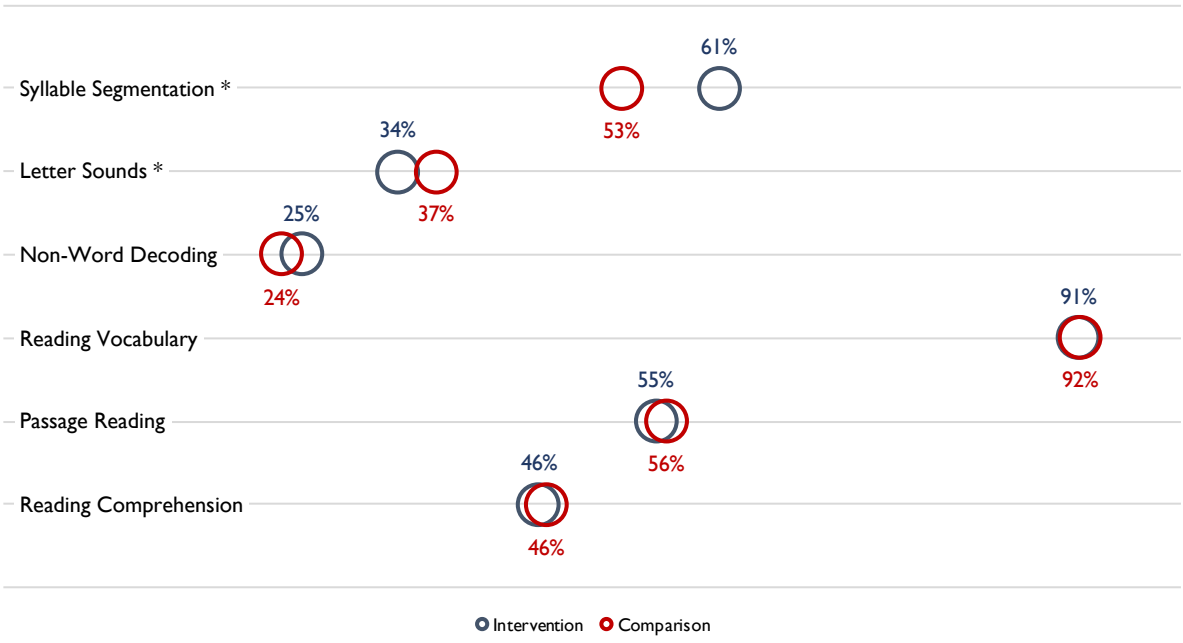
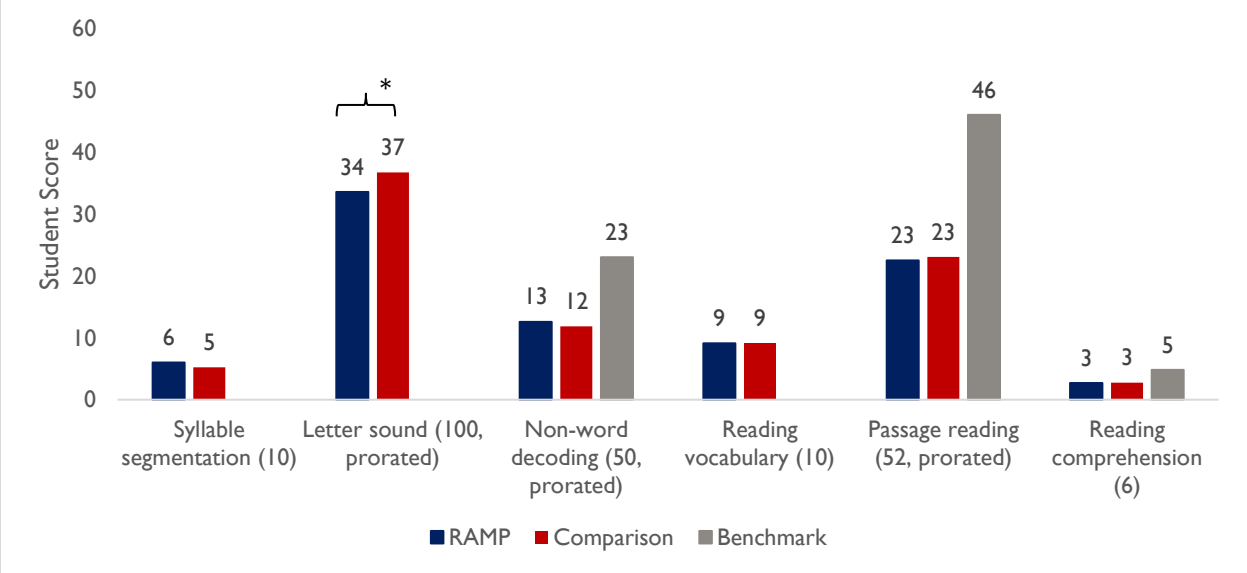


FIGURE 11. GRADE 2 STUDENT SCORES BY READING SUBTASK, FOR INTERVENTION (RAMP) AND COMPARISON SCHOOLS



Source: *RAMP Impact Study - Endline Data 2018 Student Assessments.* The figure shows endline scores equated into the baseline scale. The number of items administered at baseline is shown in parenthesis.

* Difference in group means is statistically significant at the 0.05 level.

READING RESULTS AND GENDER

The results of the gender subgroup analysis revealed no statistically significant differences in RAMP impacts between boys and girls on either average reading scores (see the rightmost column in Tables 16 and 17) or on the percentage of students who obtained a score equal to zero. Exceptions were the percentage of boys and girls who obtained zero-scores on the **knowledge of letter sounds, invented word reading,** and **reading comprehension** subtasks. However, not all of these differences were statistically significant (Annex M).

- G1 girls in the intervention group were seven percentage points more likely to obtain a zero-score in the letter-sound correspondence subtask than girls in the comparison group (a statistically significant difference), whereas the percentage of boys who scored zero in this task was not significantly different between the two groups.
- G2 girls in the intervention group were seven percentage points less likely to obtain a zero-score in the invented word reading subtask than girls in the comparison group, whereas boys in the intervention group were four percentage points more likely to obtain a zero score than comparison boys. Neither of these differences was statistically significant.
- G2 girls in the intervention group were four percentage points less likely to obtain a zero-score in reading comprehension than girls in the comparison group, but this difference was not statistically significant, whereas boys in the intervention group were seven percentage points more likely to obtain a zero score than comparison boys, a difference that was statistically significant.

TABLE 16. IMPACT ON READING PERFORMANCE SCORES FOR GRADE 1 STUDENTS AT ENDLINE, BY GENDER.

Variable (Total # of Items)	Impact for Girls (A)	*p*-value	Impact for Boys (B)	*p*-value	P-value for the difference in RAMP impacts by gender
Phoneme Isolation (out of 10)	-0.3	0.21	-0.2	0.43	0.82
Syllable Segmentation (out of 10)	0.3	0.31	0.2	0.63	0.80
Letter Sound Knowledge (out of 100, prorated)	**-4.8***	0.014	-2.6	0.16	0.40
Reading Vocabulary (out of 10)	0.0	0.84	0.0	0.84	0.97
Passage Reading (out of 41, prorated score)	0.1	0.95	-0.6	0.59	0.64
Reading Comprehension (out of 6)	-0.2	0.27	-0.2	0.21	0.91
Number of Students	**988**		**943**		
Number of Schools	**193**		**192**		

Source: *RAMP Impact Study - Endline Data 2018 Student Assessments.*

Note: Columns A and B present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well as students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated to the midline scale. There are fewer schools than in analyses for the overall sample because single-gender schools only contribute to estimating the impacts on girls or boys. There are 36 female-only schools and 48 male-only schools in the sample.

TABLE 17. IMPACT ON READING PERFORMANCE SCORES FOR GRADE 2 STUDENTS AT ENDLINE, BY GENDER

Variable (Total # of Items)	Impact for Girls (A)	*p*-value	Impact for Boys (B)	*p*-value	P-value for the difference in RAMP impacts by gender
Syllable Segmentation (out of 10)	0.7*	0.023	0.9*	0.022	0.64
Letter Sound Knowledge (out of 100, prorated)	-3.2	0.08	-3.1	0.11	1.00
Non-Word Decoding (out of 50, prorated)	0.8	0.36	0.7	0.44	0.88
Reading Vocabulary (out of 10)	-0.1	0.44	0.0	0.81	0.51
Passage Reading (out of 41, prorated score)	-0.6	0.55	-0.5	0.66	0.94
Reading Comprehension (out of 6)	0.1	0.68	-0.2	0.20	0.25
Number of Students	994		937		
Number of Schools	191		178		

Source: *RAMP Impact Study - Endline Data 2018 Student Assessments.*

Note: Columns A and B present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well as students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated to the midline scale. There are fewer schools than in analyses for the overall sample because single-gender schools only contribute to estimating the impacts on girls or boys. There are 36 female-only schools and 48 male-only schools in the sample.

*Difference in group means is statistically significant at the .05 level.

READING RESULTS AND NUMBER OF SHIFTS

The results revealed a tendency for RAMP to have some negative impacts in single-shift schools and some positive impacts in double-shift schools (see Tables 18 and 19). However, most differences between intervention and comparison students in each type of school were not statistically significant.

For G1 students in single- versus double-shift schools, RAMP had statistically significantly different impacts on **reading vocabulary, passage reading,** and **reading comprehension**. RAMP tended to benefit students in double-shift but not single-shift schools.

- G1 intervention students in single-shift schools obtained similar scores in reading vocabulary to comparison students. In double-shift schools, in contrast, intervention students scored higher than comparison students (a statistically significant difference of nearly one more question answered correctly).
- G1 intervention students in single-shift schools scored slightly lower than comparison students in oral passage reading, whereas the opposite was true in double-shift schools. Neither of these differences, however, was statistically significant.
- G1 intervention students in single-shift schools obtained lower scores in reading comprehension than comparison students (a statistically significant difference of nearly 20 percent of the comparison's group mean). In double-shift schools, intervention students scored higher on average than comparison students, but this difference was not statistically significant.

It is important to bear in mind that the lack of statistically significant differences between double-shift intervention and comparison schools may be due to limited statistical power; there were only 38 double-shift schools in the evaluation sample, compared to 199 single-shift schools.

TABLE 18. IMPACT ON READING PERFORMANCE FOR GRADE 1 STUDENTS AT ENDLINE, BY NUMBER OF SCHOOL SHIFTS

Variable (Total # of Items)	Impact for Single Shift(A)	*p*-value	Impact for Double Shift (B)	*p*-value	P-value for the difference in RAMP impacts by number of shifts
Phoneme Isolation (out of 10)	-0.3	0.17	0.0	0.91	0.49
Syllable Segmentation (out of 10)	0.1	0.65	0.6	0.28	0.46
Letter Sound Knowledge (out of 100, prorated)	**-3.3***	0.032	-5.3	0.13	0.59
Reading Vocabulary (out of 10)	-0.2	0.24	**0.8***	0.000	**0.000***
Passage Reading (out of 41, prorated score)	-1.0	0.33	3.0	0.10	**0.047***
Reading Comprehension (out of 6)	**-0.3***	0.050	0.2	0.28	**0.041***
Number of Students	**1643**		**288**		
Number of Schools	**199**		**38**		

Source: *RAMP Impact Study - Endline Data 2018 Student Assessments.*

Note: Columns A and B present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well as students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated to the midline scale.

*Difference in group means is statistically significant at the .05 level.

In G2, there were no statistically significantly different impacts of RAMP by the number of school shifts (see rightmost column in Table 19), even though the pattern of results resembled that for G1.

TABLE 19. IMPACT ON READING PERFORMANCE SCORES FOR GRADE 2 STUDENTS AT ENDLINE, BY NUMBER OF SCHOOL SHIFTS

Variable (Total # of Items)	Impact for Single Shift(A)	p-value	Impact for Double Shift (B)	p-value	P-value for the difference in RAMP impacts by number of shifts
Syllable Segmentation (out of 10)	0.6	0.07	1.6*	0.001	0.08
Letter Sound Knowledge (out of 100, prorated)	-3.8*	0.015	-0.5	0.89	0.34
Non-Word Decoding (out of 50, prorated)	0.6	0.45	1.4	0.29	0.60
Reading Vocabulary (out of 10)	-0.1	0.51	0.1	0.80	0.59
Passage Reading (out of 41, prorated score)	-0.9	0.30	0.3	0.89	0.56
Reading Comprehension (out of 6)	-0.1	0.36	0.0	0.83	0.58
Number of Students	1636		295		
Number of Schools	**198**		**39**		

Source: *RAMP Impact Study - Endline Data 2018 Student Assessments*

Note: Columns A and B present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well as students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated to the midline scale.

*Difference in group means is statistically significant at the .05 level.

EGMA RESULTS: MATH BY GRADE, GENDER, AND NUMBER OF SCHOOL SHIFTS

Grade 1 Math Results

Students who were in G1 at baseline, and G2 at endline, were assessed on six foundational mathematics skills: the ability to (1) count numbers and (2) objects, (3) identify numerals, (4) compare numerical magnitudes (also known as number discrimination), (5) detect missing number patterns, and (6) solve one-digit addition problems with accuracy and speed.

Exposure to two years of RAMP, versus one year for comparison students, had a negative impact on **addition** and no impacts on other mathematics subtasks (see Table 20 and Figure 10).

- G1 intervention students solved one less addition problem on average per minute than comparison students did. This corresponds to a disadvantage of about 20 percent of the comparison group's mean for this outcome. G1 intervention students were also three percentage points more likely to obtain a score equal to zero in this task, a difference that was statistically significant (see Table M.3 in Annex M).
- RAMP did not have significant impacts on other G1 students' math skills (Figure 10 and 11). Yet the average student in both groups answered correctly nearly seven out of 10 questions on the missing number subtask (see Table 20), which corresponds to the benchmark set by the MoE.
- The target set by MoE was for 58 percent of students to answer 70 percent of questions correctly in the missing numbers task by 2019. This target was achieved in both groups: 70

percent of students in the intervention group and 58 percent in the comparison group answered correctly at least 70 percent of the questions[13] (see Histograms in Annex O).

- However, the percentage of students who scored zero in the missing number subtask (six to seven percent) was above the MoE's target of 3.1 percent by 2019. RAMP did not have a significant impact on this or other zero-score outcomes (see Table M.3 in Annex M).
- Figure 12 displays how students' math scores changed from baseline to endline. In general, the data show that both groups of students had similar growth trajectories from G2 to G3 (see Annex O for figures on other outcomes).

TABLE 20. ENDLINE MATH PERFORMANCE SCORES FOR GRADE 1 STUDENTS

Variable (Total # of items)	Intervention (T)	Comparison (C)	Impact (T-C) *	*p*-value	Effect Size	Number of Students
Counting Numbers (out of 40)	36.3	36.6	-0.3	0.53	0.0	1931
Enumerating Quantities (out of 10)	9.8	9.9	0.0	0.37	-0.1	1931
Number Identification (out of 20, prorated)	32.3	32.6	-0.2	0.79	0.0	1931
Number Discrimination (out of 10)	9.0	9.0	0.0	0.95	0.0	1931
Missing Numbers (out of 10)	6.8	6.9	-0.1	0.49	0.0	1931
Addition Facts - L1 (out of 20, prorated)	5.1	6.1	**-1.0***	0.013	-0.2	1931
Number of Schools	**117**	**120**				

Source: Ramp Impact Study - Endline Data 2018 Student Assessments.

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes are calculated as Hedge's g. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated to the midline scale.

*Difference in group means is statistically significant at the .05 level.

13 These percentages are based on non-weighted frequencies instead of weighted, adjusted regressions.

FIGURE 12. ENDLINE MATH PERFORMANCE FOR GRADE 1 STUDENTS

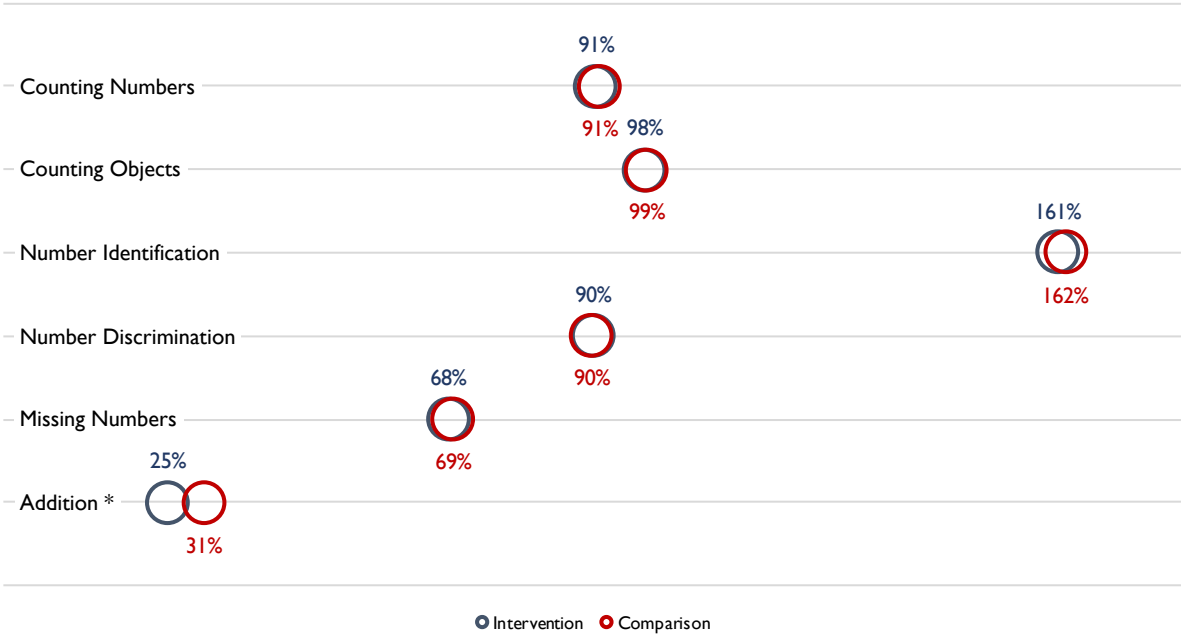
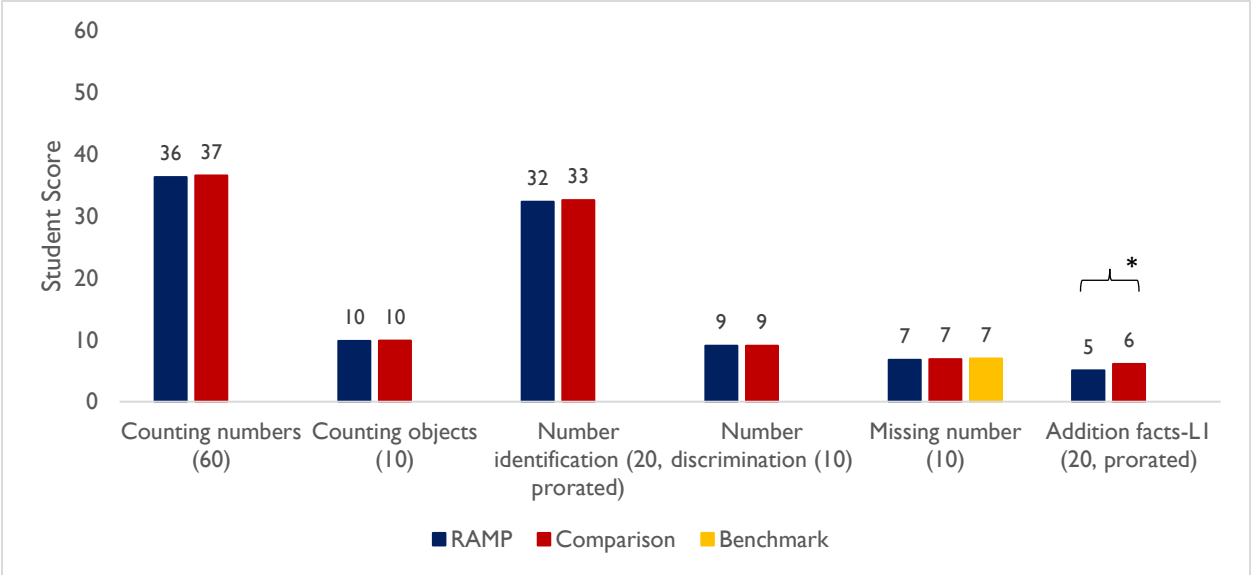


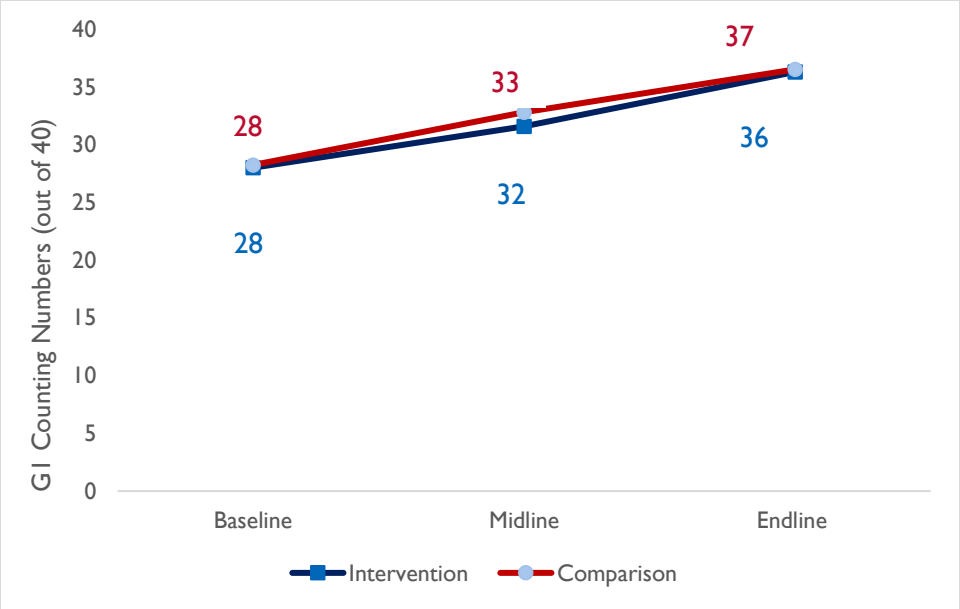
FIGURE 13. GRADE 1 STUDENTS' SCORES BY MATH SUBTASK, FOR INTERVENTION (RAMP) AND COMPARISON SCHOOLS



Source: *RAMP Impact Study - Endline Data 2018 Student Assessments.* The figure shows endline scores equated into the baseline scale. The number of items administered at baseline is shown in parenthesis.

* Difference in group means is statistically significant at the 0.05 level.

FIGURE 14. COUNTING NUMBERS OVER TIME, FOR BASELINE G1 STUDENTS



Source: *RAMP Impact Study - Endline Data 2018 Student Assessments.* The figure shows raw baseline scores, and midline and endline scores equated into the baseline scale. The number of items administered at baseline is shown in parentheses in the vertical axis label.

GRADE 2 MATH RESULTS

Students who were in G2 at baseline, and G3 at endline, were assessed on four of the same mathematics skills as G1 students: (1) identifying numerals, (2) comparing numerical magnitudes (also known as number discrimination), (3) detecting missing numbers, and (4) solving one-digit addition problems. G2 students were also assessed on their ability to solve (5) advanced addition problems and (6) one-digit and (7) advanced subtraction problems.

Results revealed that after two years, RAMP had no detectable impacts on G2 students' mathematics skills (see Table 21 and Figure 15), including the proportion of students obtaining a score equal to zero (see Table M.4 in Annex M).

- Yet the average student in both groups was able to: correctly identify all numbers presented in a minute or less, and correctly answer nine out of 10 number discrimination questions and seven out of 10 percent of missing number questions. The latter corresponds to the benchmark set by the MoE. This general improvement could be the result of the new curriculum booklet and training introduced by the MoE in the 2016-2017 academic year.
- However, the average student in both groups scored below the benchmark set by the MoE of 80 percent correct (Figure 16) or four correct answers in the Level-2 addition and subtraction subtasks (see Table M.4 in Annex M).

TABLE 21. ENDLINE MATH PERFORMANCE SCORES FOR GRADE 2 STUDENTS

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C) *	*p*-value	Effect Size	Number of Students
Number Identification (out of 20, prorated)	35.3	35.0	0.4	0.59	0.0	1931
Number Discrimination (out of 10)	8.6	8.6	0.0	0.95	0.0	1931
Missing Numbers (out of 10)	8.3	8.4	0.0	0.72	0.0	1931
Addition Facts - L1 (out of 20, prorated) timed	9.1	9.4	-0.3	0.52	0.0	1931
Addition Facts - L2 (out of 5)	2.9	2.8	0.0	0.97	0.0	1931
Subtraction Facts - L1 (out of 20, prorated)	7.0	7.2	-0.3	0.50	0.0	1931
Subtraction Facts - L2 (out of 5)	2.2	2.2	0.0	0.76	0.0	1931
Number of Schools	**118**	**119**				

Source: *RAMP Impact Study - Endline Data 2018 Student Assessments.*

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes are calculated as Hedge's g. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated into the midline scale.

*Difference in group means is statistically significant at the .05 level.

FIGURE 15. ENDLINE MATH PERFORMANCE FOR GRADE 2 STUDENTS: INTERVENTION (RAMP) AND COMPARISON SCHOOLS

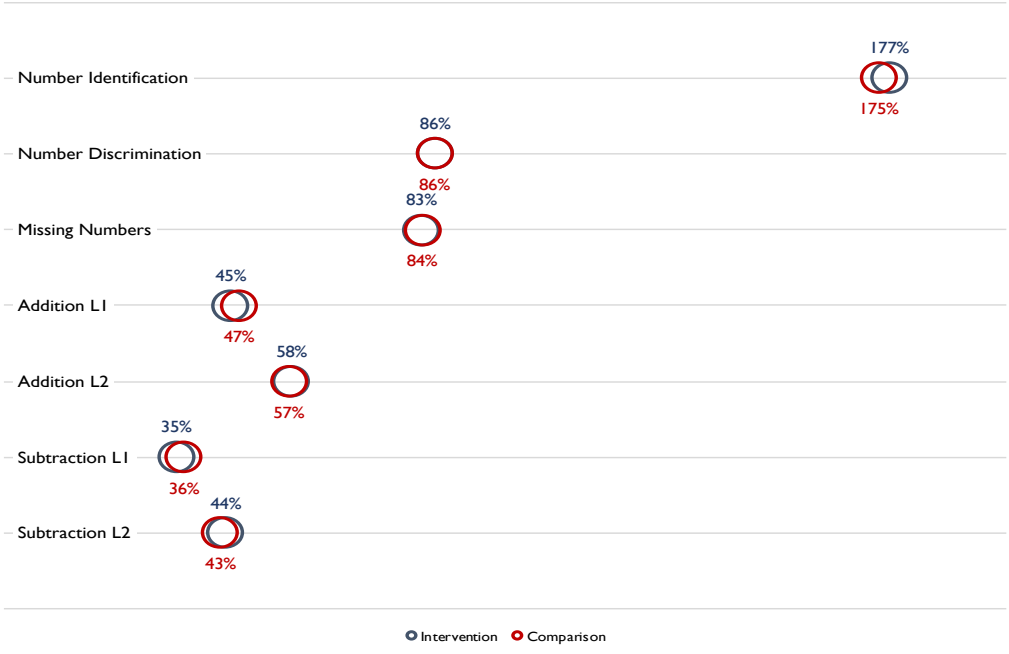
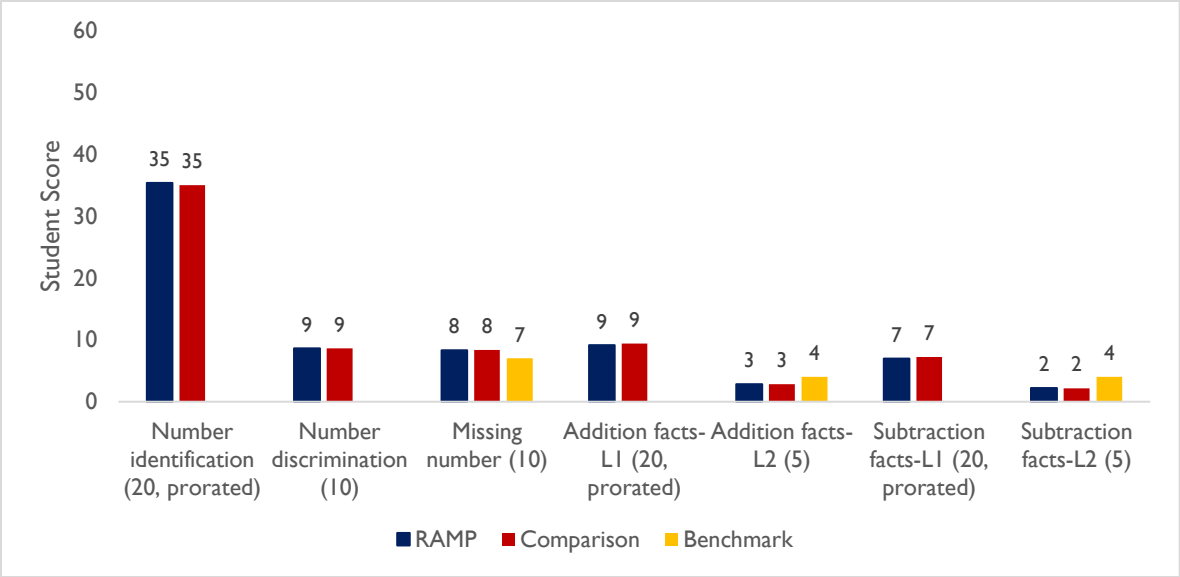


FIGURE 16. GRADE 2 STUDENT SCORES BY MATH SUBTASK, FOR INTERVENTION (RAMP) AND COMPARISON SCHOOLS



Source: *RAMP Impact Study - Endline Data 2018 Student Assessments.* The figure shows endline scores equated into the baseline scale. The number of items administered at baseline is shown in parenthesis.

MATH RESULTS AND GENDER

The results of the gender subgroup analysis revealed no statistically significant differences in RAMP impacts between boys and girls on average mathematics scores (see the rightmost column in Tables 22 and 23) and few differences in the percentage of students who obtained a score equal to zero in each of the subtasks (there are differences in the percentage of zero scores on letter sound knowledge for G1 and non-word decoding and reading comprehension for G2; see Tables M5 and M6 in Annex M).

TABLE 22. IMPACT ON MATH PERFORMANCE SCORES FOR GRADE 1 STUDENTS AT ENDLINE, BY GENDER

Variable (Total # of Items)	Impact for Girls (A)	*p*-value	Impact for Boys (B)	*p*-value	P-value for the difference in RAMP impacts by gender
Counting Numbers (out of 40)	-0.2	0.72	-0.4	0.54	0.84
Enumerating Quantities (out of 10)	0.0	0.74	-0.1	0.37	0.43
Number Identification (out of 20, prorated)	-0.2	0.89	-0.3	0.75	0.89
Number Discrimination (out of 10)	-0.1	0.53	0.1	0.44	0.28
Missing Numbers (out of 10)	-0.2	0.31	0.0	0.95	0.51
Addition Facts - L1 (out of 20, prorated)	-0.6	0.19	**-1.6***	0.010	0.11
Number of Students	**988**		**943**		
Number of Schools	**193**		**192**		

Source: Endline Data 2018 Student Assessments

Note: Columns A and B present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well as students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated into the midline scales. There are fewer schools than in analyses for the overall sample because single-gender schools only contribute to estimating the impacts on girls or boys. There are 36 female-only schools and 48 male-only schools in the sample.

*Difference in group means is statistically significant at the .05 level.

TABLE 23. IMPACT ON MATH PERFORMANCE SCORES FOR GRADE 2 STUDENTS AT ENDLINE, BY GENDER.

Variable (Total # of Items)	Impact for Girls (A)	*p*-value	Impact for Boys (B)	*p*-value	P-value for the difference in RAMP impacts by gender
Number Identification (out of 20, prorated)	0.2	0.78	0.5	0.59	0.82
Number Discrimination (out of 10)	0.1	0.63	-0.1	0.69	0.45
Missing Numbers (out of 10)	0.1	0.35	-0.2	0.22	0.11
Addition Facts - L1 (out of 20, prorated)	-0.1	0.84	-0.5	0.36	0.53
Addition Facts - L2 (out of 5)	0.1	0.36	-0.2	0.34	0.13
Subtraction Facts - L1 (out of 20, prorated)	-0.1	0.89	-0.5	0.30	0.38
Subtraction Facts - L2 (out of 5)	0.1	0.30	-0.1	0.59	0.20
Number of Students	**994**		**937**		
Number of Schools	**191**		**178**		

Source: Endline Data 2018 Student Assessments.

Note: Columns A and B present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well as students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated into the midline scale. There are fewer schools than in analyses for the overall sample because single-gender schools only contribute to estimating the impacts on girls or boys.

*Difference in group means is statistically significant at the .05 level.

MATH RESULTS AND NUMBER OF SHIFTS

As with reading, RAMP tended to have negative impacts on students in single-shift schools and positive impacts on students in double-shift schools. None of the differences in the impacts of RAMP was statistically significant in G1 (see Table 24). RAMP had a different impact on the proportion of G1 students who obtained a zero-score in the **addition** subtask (see Table M11 in Annex M).

- In single-shift schools, G1 intervention students were six percentage points more likely to obtain a zero-score in addition than comparison students were, a difference that was statistically significant. In double-shift schools, there was a difference of 4 percentage points favoring intervention students, but it was not statistically significant (see Table M11 in Annex M).

TABLE 24. IMPACT ON MATH PERFORMANCE SCORES FOR GRADE 1 STUDENTS AT ENDLINE, BY NUMBER OF SCHOOL SHIFTS

Variable (Total # of Items)	Impact for Single Shift(A)	*p*-value	Impact for Double Shift (B)	*p*-value	P-value for the difference in RAMP impacts by number of shifts
Counting Numbers (out of 40)	-0.8	0.11	1.4	0.24	0.09
Enumerating Quantities (out of 10)	0.0	0.40	0.0	0.47	0.98
Number Identification (out of 20, prorated)	-0.8	0.38	1.3	0.43	0.26
Number Discrimination (out of 10)	-0.1	0.56	0.3	0.35	0.26
Missing Numbers (out of 10)	-0.2	0.35	0.0	0.88	0.56
Addition Facts - L1 (out of 20, prorated)	**-1.2***	0.008	-0.6	0.45	0.46
Number of Students	**1643**		**288**		
Number of Schools	**199**		**38**		

Source: Endline Data 2018 Student Assessments

Note: Columns A and B present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well as students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated into the midline scales. There are fewer schools than in analyses for the overall sample because single-gender schools only contribute to estimating the impacts on girls or boys.

*Difference in group means is statistically significant at the .05 level.

In G2, there were differences in RAMP impacts on students' average scores in the **missing number** and **L1- and L2-subtraction** subtasks (see Table 25), as well as on the percentage of G2 students who obtained a score of zero in the two subtraction subtasks (see Table M12 in Annex M).

- G2 intervention students in single-shift schools obtained slightly lower scores than comparison students in the missing number and L1- and L2-subtraction tasks, but the differences were not statistically significant. In double-shift schools, intervention students obtained slightly higher scores than comparison students, but the difference was statistically significant only for the L2-subtraction subtask, where G2 intervention students solved about one more problem correctly than comparison students.
- G2 intervention students in single-shift schools were three and four percentage points more likely to score zero in the L1- and L2-subtraction subtasks, respectively, but these differences were not statistically significant. In double-shift schools, in contrast, intervention students were 12 percentage points less likely to score zero in the L1-subtraction tasks and 23 percentage points less likely to obtain a zero score in L2-subtraction. These differences were statistically significant (see Table M12 in Annex M).

TABLE 25. IMPACT ON MATH PERFORMANCE SCORES FOR GRADE 2 STUDENTS AT ENDLINE, BY NUMBER OF SCHOOL SHIFTS

Variable (Total # of Items)	Impact for Single Shift(A)	*p*-value	Impact for Double Shift (B)	*p*-value	P-value for the difference in RAMP impacts by number of shifts
Number Identification (out of 20, prorated)	-0.1	0.93	1.4	0.39	0.40
Number Discrimination (out of 10)	-0.1	0.67	0.1	0.44	0.39
Missing Numbers (out of 10)	-0.1	0.23	0.3	0.12	**0.050***
Addition Facts - L1 (out of 20, prorated)	-0.5	0.27	0.7	0.38	0.20
Addition Facts - L2 (out of 5)	-0.1	0.54	0.3	0.12	0.09
Subtraction Facts - L1 (out of 20, prorated)	-0.7	0.10	1.4	0.07	**0.016***
Subtraction Facts - L2 (out of 5)	-0.2	0.13	**1.0***	0.002	**0.001***
Number of Students	**1636**		**295**		
Number of Schools	**198**		**39**		

Note: Columns A and B present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well as children's gender, whether the child had a meal before getting to school, liked to read, attended preschool, spoke Arabic as primary language, had books to read other than textbooks, was read to at home, and did math problems at home. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated to the midline scales. There are fewer schools than in analyses for the overall sample because single-gender schools only contribute to estimating the impacts on girls or boys. There are 36 female-only schools and 48 male-only schools in the sample.

*Difference in group means is statistically significant at the .05 level.

UNDERSTANDING THE RESULTS

This section provides additional explanatory information complementary to the results presented in this impact evaluation report. The qualitative data collected over the course of the evaluation is the source of data to support the section presented below. Where relevant, the survey data (collected during the evaluation) is also referenced. Given the qualitative nature of these findings, the evaluation team is not able to determine the extent to which each factor affected the results to date. It is also important to note that most of these factors are inter-related, dynamic, and part of a complex system. Perhaps most importantly, these factors should also be viewed as potential opportunity areas to enhance, in subsequent implementation, the impact of the RAMP intervention.

The following is a summary of the findings presented in the qualitative report[14] on the fidelity of implementation of the RAMP intervention specific to the training and mentoring components (information reflects the views and perspectives of teachers, mentors, coaches, RAMP partners, and IP staff as well as findings from the observation data):

[14] The qualitative report presents the perspectives and views of principals, teachers, mentors, RAMP partners, and IP staff members (based on 87 key informant interviews); the team also collected observations on training and mentoring sessions.

TABLE 26. SUMMARY OF QUALITATIVE FINDINGS AND CONCLUSIONS

Research Question	Intervention Areas	Key Findings and Conclusions
1: ADHERENCE ***Was the planned intervention for teacher training and mentoring implemented to the RAMP design specifications?***	Teacher Training	Teacher training was implemented as per the design of 15 days, but dosage on math varied. Length of training (intervention) was too short, according to responses from teachers and observation of training modules.
	Teachers' Application of RAMP	Teachers found the implementation of routines difficult. RAMP increased workload on teachers without providing additional teaching time. No incentives were provided to teachers to carry out RAMP and increase workload. Limited resources for teaching and schools to implement RAMP. Limited time to implement RAMP due to class crowding. RAMP was not officially included in the curricula as a teaching method by the MOE.
	Mentoring	Based on observation and KII data, mentoring was not implemented as per the RTI plan. Teachers received much lower dosages compared to the design of 12 visits per year (RTI had designed their intervention so that teachers were to receive 6 mentor visits per semester for a total of 12 per year). More mentors were needed than anticipated and additional staff were added causing variance in dosage of mentor visits. Expectations around mentoring process and impact were designed and changed half way through the implementation of RAMP by RTI (the IP). From the implementation process, the IP identified a much larger need for mentors, and increased their number (which included both MOE and CADER staff). This was done because of the high load per mentor, making it difficult to achieve the 6 visits per semester per teacher, required by the RAMP design. Per KIIs and observations, Teachers/Mentors/Principals were not clear about the mentoring process as per the new 3 phased approach designed by the IP. Coaches and mentors encountered barriers to mentoring visits.
	Community Participation	Almost no parental or community participation for RAMP was identified by teachers and principals but highly desired by these two groups. Principal involvement is key to community participation according to participants. The IP has conducted more efforts in this component for RAMP, but as of the evaluation team's data collection these outcomes had largely not been included in RAMP's formal performance reporting.
	Monitoring of RAMP	Tracking and monitoring processes for RAMP were not robust and the feedback loop in the design was not fully implemented. However, the IP has made significant revisions to meet the scale of the activity. The IP has created a more comprehensive and complex M&E system to reflect the needs of the intervention, with protocols for field data collection. This study found limited use of monitoring data (e.g., sign in sheets for training, mentoring reports by coaches and mentors) even though interviews with partners suggested the IP collects many data streams. Although internal analysis may exist of these streams, this evaluation team did not review these.

Research Question	Intervention Areas	Key Findings and Conclusions
2: EXPOSURE/ DOSAGE **What, if any, were barriers in the full implementation of these two training elements (teacher training and mentoring) that could potentially affect, dilute, or diminish the effectiveness of RAMP on students?**	Teacher Training	Teachers received the training dosage of three weeks per the RAMP design. The actual dosage of math instruction, as a RAMP routine, was found to vary in trainings. Not all teachers received a full dosage of math. At the onset of RAMP there was likely an underlying assumption about teacher math skills being similar; data from this study suggest otherwise. Incentives for teachers to implement RAMP are not aligned with MOE curricula and therefore some teachers may be implementing RAMP less than others.[15] Limited access to resources and materials was constraining teachers' ability to implement RAMP in the classroom – this likely led to variations in RAMP dosage to students. Interview data suggested use of course and fine grain tools.
	Mentoring	Planned dosage of 12 visits per year was not fully implemented – rather, mentoring dosage was reduced. On average, the data suggests teachers received 2-4 mentor visits per year. Some cohorts may have received more mentoring exposure than others. Mentoring visit quality varies among teachers – likely affecting level of support and therefore mentoring dosage. Implementation level of the mentoring stages is unclear, and therefore there is limited data on mentoring effectiveness.

Research Question	Intervention Areas	Key Findings and Conclusions
3: PARTICIPANT RESPONSIVENESS **What were the perceptions by stakeholders of these two training elements?**	Teacher Training	Overall, teachers view the RAMP training very positively and seem to want in-service training to continue. However, for RAMP to continue, teachers stated a need for additional support in materials, teacher aids, and other in-class resources. At the start of RAMP, incentives for teachers to implement RAMP were not aligned with MOE curricula. However, efforts by the IP with the MOE have been made to formally integrate RAMP into the teaching requirements. Interview data showed some use of the course and fine grain tool, albeit with great variance across the schools. The IP may wish to further study this variance which could highlight further the diminished effects of RAMP at this stage, as of Spring of 2018. Principals have similar positive views to the teachers about RAMP. At the onset of RAMP, there was likely an underlying assumption about teacher math skills being similar; data from this study suggest otherwise.

[15] At the start of the study, the IP was working with the MOE to incorporate RAMP officially into the curriculum. Per the interviews with the IP and other stakeholders, this process can be lengthy and complex. However, by the end of this study, the evaluation team was informed RAMP would be officially incorporated in the teaching curricula in the coming year, 2019.

Research Question	Intervention Areas	Key Findings and Conclusions
	Mentoring	Perceptions by teachers on the mentors varied by mentor type. MOE mentors were viewed with more authority and RAMP coaches were viewed as supportive colleagues. Implementation level of the mentoring stages is unclear and therefore there is limited data on mentoring effectiveness. There was an overall positive view of mentorship by teachers. Mentors identified barriers to mentoring teachers, such as mentee load, travel/logistics problems, and teacher mobility. The Supervisors stated to have more resource barriers than the RAMP coaches, who are supported directly by the IP for mentoring work.

Based on the qualitative work conducted across the IE (including informal observations, conversations with stakeholders, and discussions with the teams), the study highlights the following issues which emerge as key factors in RAMP outcomes/impacts:

- National scale-up: scaling up a successful pilot project nationally.
- Variations in the definition of RAMP: differences in the understanding and approaches to applying the RAMP methods.
- Causal model and assumptions: RAMP's overall theory of change.
- RAMP Teacher Mentoring Component: implementation and understanding of RAMP's coaching/mentoring model.
- Challenges at the teacher level.
- Materials and resources.
- Monitoring and evaluation: process and approach to tracking changes in RAMP over time.

NATIONAL SCALE-UP: SCALING UP A SUCCESSFUL PILOT PROJECT NATIONALLY

RAMP intervention was modeled after a relatively successful pilot activity. It is important to note however that the pilot activity was quite deliberately "implemented under the best possible conditions."

> "*The classes for the intervention schools were purposefully selected to represent, as far as possible, those classes in which the intervention conditions could be as ideal as possible. This was to ensure that the endline survey measured what could be achieved if the intervention were implemented under the best possible conditions*."
>
> RTI Impact Analysis Report 2014, page 33.

The extent of the success of the pilot intervention may be in part due to the conditions under which it was implemented – ideal controlled conditions within a small selected number of schools. Scaling up such an intervention nationally is bound to confront variation in the speed at which: components can be rolled out, staff can be prepped and launched into the field, and M&E systems adapt to field conditions. As a project scales-up nationally, the implementation of it stretches to meet the requirements but can lead to variation in dosage, whether overtime or in spatial terms. These may account for the limited, or no impacts, observed to date.

As shown in the qualitative report, dosage varied across multiple factors (i.e., RAMP application by teachers, teaching workshop times for math routines, mentoring dosages) was seen in RAMP. This variation in dosage has been decreasing over time as the IP revises implementation processes to

standardizing RAMP components across all schools. An example of a scale-up issue concerned limited materials for teachers to apply RAMP in the classroom. During interviews with teachers they stated they could not conduct RAMP routines as they had to copy by hand materials for the students due to limited photocopying capacity within schools. The IP did learn of this issue and resolved the problem for Cohorts 2 and 3. Class size was also another challenge in the implementation of RAMP, likely not encountered in the pilot. As RAMP went to a national scale, the implementation of routines for larger classes was problematic due to time limitations; teachers are required to cover specific material per the Ministry's set curricula for each grade.

VARIATIONS IN THE DEFINITION OF RAMP: DIFFERENCES IN UNDERSTANDING AND APPROACHES TO APPLYING THE RAMP METHODS

The qualitative data collected over the course of the evaluation pointed to evidence of the differences in key stakeholders (teachers, coaches, mentors, principals) understanding of RAMP methods/routines, and how best to apply them. This may be due to variation in dosage or because the RAMP approach is not highly prescriptive when it comes to when and how to implement the approach. During the RAMP trainings teachers are instructed to apply RAMP in one of three ways:

- The teacher assigns work to most of the class while working on the foundational skills of a small group of students, using RAMP methods;
- The teacher uses one of the extra lessons in the week to work on student foundational skills using RAMP routines; or
- The teacher spends 10-15 minutes at the start of the lesson applying one or more of the RAMP routines.

These findings were also substantiated by the survey data collected from teachers who were observed. When asked which approach they used to apply RAMP in classrooms, of those who responded to the question, around 40 percent reported spending 10-15 minutes at the start of the lesson doing one or more RAMP routines. Close to 18 percent of the teachers reported using one of the extra lessons in the week to apply RAMP routines and only nine percent mentioned the approach of assigning work to most of the class while working on the foundational skills of a small group of students, using RAMP methods. Around 24 percent of the teachers opted for the "other" option. Teacher interviews (see RAMP qualitative study), also support the survey data. Teachers went to so far as to state they would only do RAMP routines when they knew observers were in their classroom, suggesting inconsistent RAMP dosage to children. The above approaches are qualitatively different and could account for variable impact based on the type and frequency of approach applied.

RAMP'S OVERALL THEORY OF CHANGE

As an intervention, RAMP is primarily focused on the teachers, starting with a ten-day training that takes place during the summer break. This training is followed by coaches and mentors observing teachers during the school term. Teachers are offered another five-day training between semesters. Given this focus on the teachers, it is to be expected that the initial impact of the intervention will be on the teachers as well. This is consistent with the results of the teacher observation impact study, which has identified some areas in which the RAMP teachers are applying more reflective methods compared to teachers who have not been exposed to RAMP. The findings to date also point to managing expectations around the pace at which RAMP methods are being adopted by teachers. This also has implications for expectations around student-level impact and the amount of time it will take for such impact to take place, assuming implementation fidelity. Again, these findings suggest a possible lag-effect, supporting the recommendation to re-measure RAMP impacts in subsequent years.

It is also important to consider the overall dosage of RAMP. As mentioned above, all together RAMP training is 15 days of teacher training. So, while it is supported by other key elements such as coaching and mentoring support to teachers, the training itself is relatively short. Successfully applying what has been learned during the training requires several elements to also work effectively.

In addition to the above, while some elements of the theory of change are supported by existing literature (for example, coaching and mentoring), there is a need to consider the evidence base around the combination of inputs that constitute RAMP to understand the necessary conditions required for impact. The qualitative study presents the barriers encountered by teachers to fully implement RAMP, and therefore identifies some obstacles to the existing assumptions about the RAMP model.

RAMP TEACHER MENTORING COMPONENT: IMPLEMENTATION AND UNDERSTANDING OF RAMP'S COACHING/MENTORING MODEL

The coaching/mentoring component of the RAMP intervention evolved into a three-phase approach according to the RTI team. A shift from one phase to the next was not based on satisfaction of any specific criteria, but rather based on maturity of the cohort (i.e. the passage of time). Below is a summary of the evolved phased approach:

- Phase 1: Coach/mentor visits and frequency of visits were the most important aspect of this phase to prompt the teachers to implement RAMP methods. This phase was not evaluative, as it did not aim to evaluate how effectively the teachers were applying the RAMP methods.
- Phase 2: During this phase the coaches could provide feedback to the teachers. As of spring 2018, this phase had primarily started in Cohort 1 schools.
- Phase 3: During this phase schools will be receiving differentiated support based on student performance. RAMP has been able to classify schools based on student performance using their Lot Quality Sampling (LQS) study. The classification includes A (teacher level support), B (community of practice level support), and C (cluster-level support) categories.

This approach was different from the original intent of having coaches/mentors provide feedback to teachers during the visits following the teacher training (phase 1). This is significant, as it is typically most useful to provide feedback in the timeframe immediately after the training. Not evaluating how teachers were implementing RAMP in phase 1 was not only a missed opportunity to ensure implementation fidelity, it may also have resulted in some teachers applying the RAMP methods inaccurately or less effectively than was originally intended.

Based on our qualitative data, it appears that there was no awareness of these phases among mentors, coaches or teachers. As per some of our qualitative respondents, there were also instances where the coaches and mentors provided inconsistent and at times even contradictory feedback to the teachers, then what they had learned during the training.

Lastly, the overall workload for coaches and mentors varied significantly with a much larger workload for the RAMP coaches. This could also have a bearing on variation in quality of their work. Mentors and coaches provided various obstacles for RAMP mentoring:

- The number of teachers per mentor/coach is very high, limiting their ability to reach every teacher 6 times a semester. From the 16 mentors interviewed, RAMP coaches conducted 943 mentoring visits, compared to 474 by supervisors (MoE mentors)[16].

16 For those who participated in the qualitative survey (16 mentors)

- Supervisors have more limited time to visit teachers then RAMP coaches; MoE staff have other responsibilities and duties limiting their time for mentorship.
- There are logistical issues in reaching all teachers as some are located far from the mentor and require extra time for travel.
- Sometimes the mentors found the teacher absent when they traveled to the school and were unable to complete the mentoring visit.

CHALLENGES AT THE TEACHER LEVEL

Our qualitative data collection also pointed to several factors at the teacher level that may have made it difficult to consistently implement RAMP at the individual teacher level as well as across the education sector. These included teacher workload, teacher turnover and delays in teacher onboarding.

When asked about the potential challenges related to implementing RAMP methods in classrooms, several teachers reported that their existing heavy workload makes it difficult for them to integrate the RAMP routines. Assigning and following up on RAMP worksheets given to individual students, for instance, requires additional time. The challenges around the workload may be addressed by more explicitly integrating the RAMP routines into the Jordanian curriculum, so applying these routines is not viewed as an effort in addition to the curriculum, but rather as part of implementing the curriculum.

The RTI team also reported unusually high rates of teacher turnover within the early grades. Practically, this has resulted in the RTI team having to train many more teachers than they had originally envisaged. This was also evidenced by the attrition within the teacher observation sample of the impact evaluation. Within one year, around one third of the teachers observed during the baseline data collection were no longer teaching the same classes. In addition to stretching activity resources, since more teachers need to be trained, teacher turnover can also result in lack of pedagogical consistency, including how RAMP methods may be integrated into the classrooms.

In addition to the above, the RTI team also noted a key challenge related to the delays in onboarding of teachers. As per the RTI team, close to one-third of the teachers are not on board at the beginning of the semester. This also means that they are not able to attend and participate in the summer trainings. To address this gap, the RTI team has had to develop another round of training offered during the semester. While RTI efforts in this regard deserve to be acknowledged, the delays in teacher onboarding can significantly impact the introduction of RAMP methodologies from the beginning of the semester.

MATERIAL AND RESOURCES

When asked about the challenges associated with implementing RAMP routines, several teachers pointed to material constraints such as not having enough RAMP worksheets for every student in the classroom. In some schools, absence of a photocopier was a related issue raised in this regard. These issues were particularly highlighted during the first round of qualitative data collection undertaken in the fall of 2016. Subsequently, the RAMP team has provided additional supporting material to the MoE and as a result some of these issues may no longer be as relevant.

MONITORING AND EVALUATION

When it comes to effectively using M&E to inform and improve activity implementation, the evaluation team found a few missed opportunities.

COACHING AND MENTORING

While coaching and mentoring is a critical element of the overall theory of change of RAMP, the RTI team was not able to effectively integrate the learning from the vast amount of M&E data they were collecting for this component. RTI was able to establish a somewhat robust M&E system that included the coaches and mentors' observation data with over 100,000 records. However, there were several issues with RTI's approach to conducting teacher observation.

- Coaches and mentors were not effectively trained on the observation methodology as there was no measurement of inter-rater reliability among observers, nor was there a codebook that could encourage observers to apply the observation tool consistently. The lack of inter-rater reliability can also translate into coaches and mentors potentially providing contradictory feedback to the teachers.
- The observation records were not reviewed in a timely manner to allow for learning to inform the subsequent rounds of teacher and/or coaches/mentor training. For example, the observation data could have identified specific routines or elements of the general RAMP approach that teachers were struggling to implement effectively.
- The evolved phased approach for coaching/mentoring, as described by the RTI team, was not reflected in their M&E system.

TEACHER'S APPROACHES

On the M&E side, more could have been done to better understand teachers' approach to applying the RAMP methods in their classrooms. For example, there was no effort made to assess, even for a small sample, whether and how teachers were using the "coarse grain" tool. This could have been helpful for understanding whether teachers were using this tool and if so, if it was being used as had been intended.

CONCLUSIONS

The RAMP intervention aimed to develop teachers' instructional skills, so they regularly assess students' learning needs in reading and math and apply differentiated teaching methodologies to improve student learning outcomes (see Figure 1). These improved instructional practices, over time, were intended to improve students' reading comprehension and math performance, helping students meet grade level standards.

The RAMP impact evaluation was designed to (1) assess whether RAMP training and mentoring support affected teachers' instructional practices, (2) estimate RAMP impacts on students' proficiency in reading and math, and (3) determine whether RAMP impacts varied by select student and school characteristics. Findings are summarized below, followed by recommendations for the future.

TEACHERS

The impact study of teachers' practices showed G3 teachers reported receiving training and coaching and were observed using RAMP strategies and learning materials during math and reading lessons. There was no evidence, however, that any of these activities translated into broad-based changes in teachers' use of instructional time, student engagement, or classroom management. Endline results showed RAMP had few impacts on lesson content. RAMP reduced the time spent on identifying, reading, and writing characters and numbers; RAMP reduced the number of reading lessons that included writing activities. RAMP increased the proportion of lesson time spent building students' vocabulary. There were no impacts

on any of the other five content areas measured, either in terms of overall coverage or the percent of lesson time spent on each area. Of the 17 student engagement and classroom management practices examined, RAMP only influenced the quality of teacher feedback during math lessons and increased the proportion of reading lessons where the teacher used whole class instruction.

These endline results were consistent with those from the descriptive study of G1 and G2 teacher practices, where few statistically significant differences were detected between the two groups and a clear pattern of results could not be discerned. Findings from classroom observations were also consistent with teachers' self-reports of their own instructional practices, including the use of assessments strategies and methods to support underachieving students and students with behavioral difficulties (see Annex Q).

RAMP did not have a fully developed theory of change that prioritized the most salient program inputs and linked them to short, medium, and long-term outcomes. It may be that the design did not adequately account for and aim to influence the root causes of poor student performance. From the baseline teacher observation, the data showed that instructional quality prior to RAMP implementation was already modest with teachers spending most of the lesson time on-task. They also practiced appropriate lesson pacing, covered content such as phonemic awareness, vocabulary, and reading comprehension skills, and had coaching and mentoring support when needed (see Annexes J and K). Thus, it may be that RAMP aimed to influence practices that teachers adequately executed already. Or it may be that implementing at scale reduced the potency of the intervention, for example, by limiting the quality or frequency of coaching and mentoring. Further, teachers may have been overwhelmed and unable to implement RAMP strategies in their already hectic schedules. They must already implement the MoE curriculum, which was revised during the 2016-2017 school year and focused on both content and instruction. Building on the RAMP experience, stakeholders will be better positioned to design an evidence-based approach with clearly articulated components that can be more effective in yielding positive impacts.

Again, it is possible that the intervention spilled over from trained teachers to untrained comparison teachers, which in turn reduced impacts if G3 comparison teachers also improved their practices. For example, it may be that G1 and G2 teachers encouraged G3 teachers to implement RAMP strategies before they were trained. While the evaluation did not intentionally collect data on this, it is possible. A non-trivial number of G3 teachers in the comparison group reported having participated in RAMP training, however most of these teachers received a scaled back version with several hours of training versus several weeks of the RAMP intervention.

STUDENTS

The lack of RAMP impacts on teachers' practices support evaluation findings on student scores, with little evidence of positive (or negative) impacts on students' reading and math outcomes. After approximately two school years, RAMP only had a positive impact on G2 students' ability to segment words into syllables and no positive impacts on other math or reading outcomes in either grade. These results were largely consistent with those from midline analysis, where a positive effect was detected on syllable segmentation, but not on other reading or math outcomes. At midline, RAMP also had a positive impact on G1 students' passage reading, but that result did not hold at endline.

Contrary to expectations, RAMP had negative impacts on G1 and G2 students' knowledge of letter sounds, G1 students' addition, and on the percentage of G1 students who obtained a zero-score in the phoneme isolation and oral passage reading subtasks. The negative impacts on students' knowledge of letter sounds and ability to isolate phonemes may be related to a decreased focus on character reading and identification relative to the comparison group, as evidenced by the impact study of G3 teachers' practices. We cannot fully explain this change but sharing these findings with teachers and asking about

their interpretation might yield more insights. These negative impacts were not detected at midline, but a negative impact was detected on G1 students' ability to count numbers.

STUDENT SUBGROUPS

Results from the analysis of RAMP impacts by students' gender were consistent from midline to endline but results from the analysis of impacts by number of school shifts were not. Endline results indicated that, overall, RAMP did not have a different impact on boys versus girls, and these findings were consistent with midline results from this evaluation. At endline, there were differential impacts on the proportion of students who obtained zero-scores in three reading subtasks (knowledge of letter sounds, invented word reading, and reading comprehension), but neither gender had a consistent advantage. This pattern of results differed, however, from results from the pilot study of RAMP, which found that girls appeared to benefit more than boys did from the intervention (Brombacher, 2015). Note, however, that the pilot study was not an impact evaluation and it is unclear whether the results were validated. Further, the difference in results may be due to several factors, including the use of different study designs, different populations, or a deliberate effort by the implementer to reduce differential impacts of RAMP by gender.

Unlike midline results, endline findings revealed a tendency for RAMP to have a negative impact on single-shift schools and a positive impact on double-shift schools. However, differences between the intervention and comparison groups were often non-significant and conclusions were limited by the relatively small number of double-shift schools in the evaluation sample. Further the sample did not contain enough schools that reported receiving infrastructure support to assess impacts in the analysis.

RECOMMENDATIONS

The team proposes recommendations in five main categories based on the evaluation: program design, program implementation, monitoring and evaluation, reporting and dissemination, and evaluation timeline.

PROGRAM DESIGN

The RAMP program suffered from several design flaws that may have undermined the program's success. For example, in human development interventions -specifically in the education sector- projects develop and use a theory of change reflective of the assumptions/expectations to produce the desired outcomes. The development hypothesis underlying the RAMP program did not have enough detail to guide implementation effectively. The hypothesis lacked a clear articulation of the salient programmatic components and a prioritization for how these components would be implemented and contribute to teachers improved instructional practices and student outcomes over time. While program designers may have understood some of the root causes of student performance, it is not clear that the intervention prioritized the key constraints or that it fully understood the weaknesses in instructional practices. We recommend that USAID continue to build on efforts to establish a shared culture of monitoring, learning, and evaluation imbedded early on in the program design phase. By including external evaluators early in the design phase, there can be collaborative efforts aimed at avoiding design flaws prior to program scale up.

RECOMMENDATIONS

- Implementers should develop a theory of change, whereby all stakeholders validate the framework and its associated logic model. The stakeholders should provide feedback on whether it can be implemented with fidelity given the context, timeline, and resources.
- The theory of change and logic model should include short, medium, and longer-term indicators which are specific, measurable, appropriate, realistic, and time bound (SMART) so that the implementer and evaluator can track progress linked to key inputs and implementation.
- Once the theory of change is widely accepted, the implementer should ensure that programmatic components align with project goals or modify goals to reflect the situation on the ground. While the research team found some evidence that programmatic components were updated to reflect the Jordanian context, examples were also found where the program should have been modified to account for persisting barriers to implementation. One example is teachers reporting that it was difficult to implement both the RAMP routines and the Ministry curriculum. While this finding was shared with the implementer, there was no evidence by endline that the program was modified to meet this challenge.
- Additionally, such a partnership would be expected to enable an external evaluation team to consider and propose implementation plans that would support a randomized control trial with data collection intervals matched to implementation.
- To support the implementer-evaluation partnership, USAID is advised to include in future program designs a full recognition of planned evaluation efforts, commencing both program implementation and evaluation design at the same time.

IMPLEMENTATION

Despite the urgency to quickly improve the quality of education throughout Jordan, it is recommended that stakeholders extend the implementation timeframe to ensure quality rollout of a well-designed

program. Effective implementation requires a realistic and achievable plan for scaling up countrywide and it generally requires a longer timeline. The RAMP program has been evolving and responding to unanticipated challenges, which is appropriate. However, realistically, a country-wide effort is too large in scale to implement with fidelity, when the intervention is not fully designed, adapting to challenges, and not externally tested. Better-quality implementation likely requires smaller-scale implementation. Only after reaching success at a small scale, can an intervention be scaled with fidelity and sustainability. While we acknowledge that the implementer and local stakeholders believed the pilot was successful, we believe an evaluability assessment, conducted by an external evaluation agency, would have highlighted key weaknesses in the implementation plan, theory of change, and the underpinning study design which formed the basis of the decision to move to scale.

RECOMMENDATIONS

- In future programming, it is recommended stakeholders consider conducting an evaluability assessment (EA). This tool has been used to improve program design and implementation in health and human development programs (Davies 2013, Miller et. al 2016). An EA should be conducted prior to large scale program implementation to ensure the intervention has the potential to achieve program goals. EAs involve a series of steps or components and require stakeholders to work together to fully clarify the program design, explore the program reality at various stages of pilot activities, conduct external assessments of the program outcomes, enlist stakeholder input into design and implementation corrections, and make a collaborative decision about whether to scale up the intervention or conduct an impact evaluation (Davies 2013).
- Another way to improve RAMP implementation is to conduct frequent analysis of program monitoring data collected during teacher coaching and mentoring visits. This data could have informed stakeholders of the implementation gaps in real time throughout the school year. Unfortunately, this data was not systematically analyzed or used, and there were issues with how the data was collected reducing its value. Components around implementation of RAMP could have been made more effective earlier, had the monitoring data been mined for recommendations to improve further the program.
- Another suggestion is for implementers to work closely with evaluation partners from the project inception to collaboratively develop M&E plans. This type of partnership would ensure that all necessary information is collected, analyzed, and utilized to improve program implementation without repetition or wasted resources. This includes transparently discussing the learning potential and rigor of various study designs and analytical procedures. This partnership, launched at program inception, would also enable an external evaluator to help think through implementation plans that would support a randomized controlled trial with data collection intervals matched to implementation.

MONITORING AND EVALUATION

To achieve a successful evidence-based program, it is important to ensure there is a realistic and manageable monitoring and evaluation plan designed to yield high quality, informative data. Implementers, like evaluators, should avoid over collecting data that is not used and work with evaluation partners to ensure relevant indicators/measures are monitored and assessed along an appropriate and realistic timeline.

RECOMMENDATIONS

- Implementers should align their M&E plan, staffing time, technical capacity for analysis and interpretation and utilization of results for programmatic improvements.
- Implementers must also provide ongoing training to staff for data analysis as needed to ensure there are no barriers to data utilization.
- Another option is to outsource the data analytic tasks to ensure they are completed. Implementers, with help from evaluators, should identify, prioritize, and use key indicators to measure the quality of implementation as well as identify areas for improvement. These teams should develop analytic plans prior to data collection and confirm the plans are realistic given limited resources. Overall, these teams should ensure methods are appropriate so that data does not provide a false sense of program outcomes. For example, research plans must guard against over sampling from stronger regions or schools and instrumentation must be appropriately targeted to students at each grade level to measure grade appropriate outcomes. Implementers should beware of using inadequate study designs to measure impact or change over time and seek an external review of methods to confirm alignment between the study design, evaluation questions, analytic plan, and interpretation of data.

REPORTING AND DISSEMINATION

Again, we recommend that USAID continue to build on efforts to establish an environment and a shared culture that is conducive to external evaluation. This includes adequate communication as well as information and document sharing. For example, the evaluation team recommends implementers regularly share all key components of program design, implementation, and M&E efforts—in the local language and English—so that all stakeholders can access materials. In this case, the evaluation team could have provided feedback on the teacher training in more detail earlier on had materials been made accessible and if revised materials were available.

RECOMMENDATIONS

- Provide all materials in local and English languages, to all stakeholders including evaluation teams.
- Evaluators should have access to the instrumentation used by implementers to align data collection tools as well as understand other evidence that is informing stakeholders. For example, the implementer's Lot Quality Assurance Sampling (LQAS) study was shared with government stakeholders and the media, setting high expectations for student performance. Findings from the LQAS study conflict with the evaluation findings. The evaluation team could have benefited from additional information about how these internal metrics were acquired.
- Information on the study design, instrumentation, training procedures, interrater reliability scores, and analytic plans would all be helpful so that stakeholders, working towards a common goal, have the necessary information to provide helpful inputs and problem solve.

EVALUATION TIMELINE

Funders and stakeholders desire quick results; however, evaluations need to be designed to allow a longer exposure period. As mentioned, teachers may need more time to fully adopt and implement RAMP instructional strategies and students may require more years of RAMP trained teachers to demonstrate improved performance. If funders, implementers, evaluators and other stakeholders work in partnership to carefully determine the length of exposure needed to meet short, medium, and long-term program goals, then the evaluation should be designed with a long enough timeline to measure these outcomes.

Ongoing program monitoring data, if adequately collected, processed, and shared, can provide stakeholders with confidence about the quality of the program until the evaluation can capture desired outcomes.

RECOMMENDATION:

- Implement a second endline to test RAMP impact after a longer time frame to potentially capture lag effects.

REFERENCES

Brombacher, A. "National Intervention Research Activity for Early Grade Mathematics in Jordan." International Commission on Mathematical Instruction, Study 23 Conference Proceedings. Macau, China, 2015.

Brombacher, A. and Marissa Gargano. "Early Grade Reading and Mathematics Initiative (RAMP) 2017 Midline Survey Report." RTI International, Research Triangle Park, 2017.

Brombacher, A., P. Collins, C. Cummiskey, E. Kochetkova, and A. Mulcahy-Dunn. "Student Performance in Reading and Mathematics, Pedagogic Practice, and School Management in Jordan." Research Triangle Park, NC: RTI International, August 2012.

Brombacher, A., Stern, L., Nordstrum, L., Cummiskey, C., & Mulcahy-Dunn, A. "Education Data for Decision Making (EdData II): National Early Grade Literacy and Numeracy Survey–Jordan." Intervention Impact Analysis Report. Retrieved from Research Triangle Park, NC, 2014.

Brombacher, A., Stern, L., Nordstrum, L., Cummiskey, C., & Mulcahy-Dunn, A. "Education Data for Decision Making (EdData II): National Early Grade Literacy and Numeracy Survey–Jordan. Intervention Impact Analysis Report." Research Triangle Park, NC: RTI International, November 2015.

Davies, R. "Planning Evaluability Assessments: A Synthesis of the Literature with Recommendations." Working Paper 40 Vol. London: UK Department for International Development, 2013. Web.

Golinelli, D., Ridgeway, G., Rhoades, H., Tucker, J., & Wenzel, S. (2012). Bias and variance trade-offs when combining propensity score weighting and regression: with an application to HIV status and homeless men. *Health Services and Outcomes Research Methodology, 12*(2-3), 104-118.

Hamre, B.K., Pianta, R.C., Downer, J.T., DeCoster, J., Mashburn, A.J., Jones, S.M., Brown, J.L., Cappella, E., Atkins, M., Rivers, S.E. "Teaching through interactions." The Elementary School Journal, vol. 113 no. 1, 2013, pp. 461-487.

Kim, Young-Suk Grace, Helen N. Boyle, Stephanie Simmons Zuilkowski, and Pooja Nakamura. "Landscape report on early grade literacy." Washington, DC: USAID, 2016.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods, 9*(4), 403.

McEwan, P. J. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." Review of Educational Research vol. 85, no. 3, September 1, 2015, pp. 353–94.

Miller, C., K. Place, J. Tyler, G. Kafkas, and K. Boller. "Evaluability Assessment: Fursa kwa Watoto (FKW)." Cambridge, MA: Mathematica Policy Research, 2016.

RTI International. "USAID Early Grade Reading and Math Project (RAMP) Volume I Revised Technical Application." Amman, Jordan: RTI International, December 2014.

RTI International. RAMP Early Grade Reading and Mathematics Initiative Accomplishments. Amman, Jordan. January 2017.

Schochet, Peter Z. "The Late Pretest Problem in Randomized Control Trials of Education Interventions (NCEE 2009-4033)." Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, 2008.

United Nations Educational, Scientific and Cultural Organization. "UNESCO 2014." Washington, DC: 2015. http://unesdoc.unesco.org/images/0023/002324/232432e.pdf

USAID. "School Rationalization Baseline Study Report." Amman, Jordan: USAID, 2011.

What Works Clearinghouse. "Designing quasi-experiments: Meeting what works clearinghouse standards without random assignment." Webinar. March 3, 2015. https://ies.ed.gov/ncee/wwc/Multimedia/23.

RAMP IMPACT EVALUATION FINAL REPORT

SUPPORTING ANNEXES

SEPTEMBER 2019

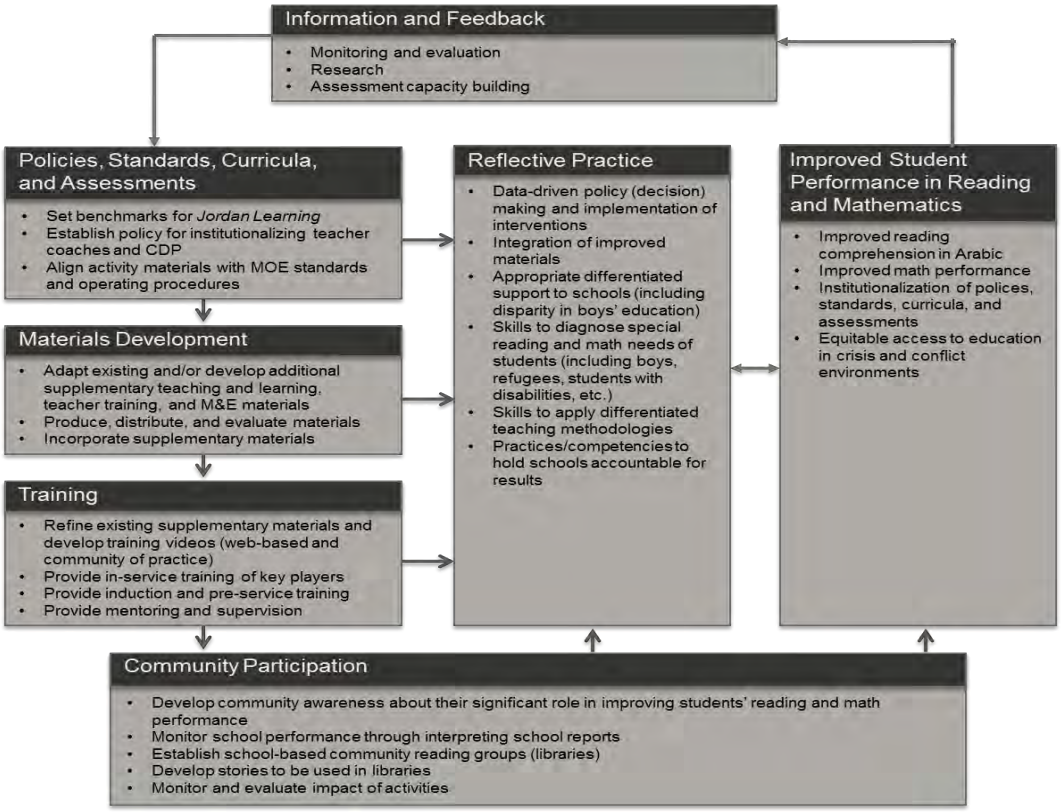
This publication was produced for review by the United States Agency for International Development. It was prepared by Management Systems International (MSI), A Tetra Tech Company.

CONTENTS

ANNEX A. USAID RAMP DEVELOPMENT HYPOTHESIS, IMPLEMENTATION SCHEDULE, AND TIMELINE

According to RTI, the Development Hypothesis, presented in Figure A.1, incorporates data-driven decision making and learning into the cycle to improve learning outcomes in reading and math education. In the RAMP activity, data generation was intended to serve both as feedback on activity performance and as an input for formulating new decisions, policies, and standards.

FIGURE A.1: USAID RAMP DEVELOPMENT HYPOTHESIS.



As shown in Table A.1, RAMP was implemented in 1,087 schools (Cohort 2) in the 2016–2017 academic year and in 749 additional schools in the 2017–2018 academic year (Cohort 3). RTI

assigned governorates (and their schools) into cohorts based on operational ease; therefore, a randomized control trial was not feasible. Instead, the impact studies capitalized on the staggered implementation approach to measure impacts on students and teachers in Cohort 2 relative to students and teachers in matched schools in Cohort 3. For the evaluation, the intervention group consisted of a sample of Cohort 2 schools and the comparison group consists of a matched sample of Cohort 3 schools that received the intervention a year later. Cohort 1 was not included in the RAMP impact evaluation because implementation had begun before the impact evaluation was designed.

TABLE A.1. RAMP IMPLEMENTATION SCHEDULE

RAMP Cohort	Implementation Year	Begin in School Year	Grade Level Implemented in	Number of Schools	Governorates
Cohort 1	Year 1	2015–2016	KG2, G1, G2	623	Ajloun, Jerash, Karak, Zarqa
	Year 2		G3		
Cohort 2	Year 1	2016–2017	KG2, G1, G2	1,087	Amman, Aqaba, Irbid, Tafilah
	Year 2		G3		
Cohort 3	Year 1	2017–2018	KG2, G1, G2	749	Balqa, Maan, Madaba, Mafraq
	Year 2		G3		

TABLE A.2. TIMELINE OF RAMP IMPLEMENTATION AND EVALUATION ACTIVITIES

Year	2016		2017				2018		
Quarter	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**
RAMP Implementation									
G1 and G2									
Cohort 2	RAMP Training	Teacher Coaching Visits (6)							
Cohort 3					RAMP Training	Teacher Coaching Visits (6)			
G3									
Cohort 2					RAMP Training	Teacher Coaching Visits (6)			
Cohort 3									RAMP Training
Data Collection		**Baseline**		**Midline**				**Endline**	
Students									
G1		EGRA/ EGMA		EGRA/ EGMA	End of school year; students advance one grade.				
G2		EGRA/ EGMA		EGRA/ EGMA				EGRA/ EGMA	
G3								EGRA/ EGMA	

Year	2016		2017				2018		
Quarter	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**	**Q4**	**Q1**	**Q2**	**Q3**
Teachers/Principals/Education Leaders									
G1 and G2		-Classroom Observations (Descriptive Study) -Qual. Interviews -Teacher/ Principal Surveys		- Classroom Observations. (Descriptive Study) -Teacher/ Principal Surveys				-Teacher/ Principal Surveys	
G3				Teacher Observation (Impact Study Baseline)				Teacher Obs. (Impact Study Endline)	
				Qual. Interviews					

ANNEX B. SCHOOL SAMPLING AND MATCHING PROCEDURES TO INCREASE SIMILARITY BETWEEN THE INTERVENTION AND COMPARISON GROUPS

A rigorous impact evaluation requires that a group of students exposed to the intervention be compared to a group of similar students who have not been exposed. The comparison group serves as the counterfactual, allowing the researchers to estimate what would have happened if the intervention had not been implemented. It is essential that the two groups differ only in their exposure to the intervention and are otherwise as similar as possible. This similarity allows evaluators to attribute any post-intervention differences between the groups to the intervention, instead of preexisting differences. Randomized controlled trials (RCT) are considered the gold standard of impact evaluation because they have a higher probability than other designs to produce groups that are similar, on average, on both observable and unobservable characteristics.

RTI assigned governorates (and their schools) to one of three cohorts and staggered the implementation of RAMP in the cohorts over three consecutive years. An RCT was not feasible because RTI assigned the governorates based on operational ease, rather than using random assignment. Thus, the impact evaluation used a quasi-experimental design that took advantage of the phased implementation of RAMP by cohorts. Specifically, given that Cohort 3 schools (last group to receive the treatment) were to start implementing RAMP one year after Cohort 2 schools, Cohort 3 schools could serve as a comparison group. The evaluation team used a matching procedure to find a group of schools as similar as possible to schools in the intervention group among Cohort 3 schools, as described below. (Cohort 1 schools were not included because they had started implementing RAMP a year before the evaluation started.)

PROPENSITY SCORE MATCHING TO CREATE SIMILAR COMPARISON AND INTERVENTION GROUPS

To select a representative sample of intervention schools and identify similar comparison schools, a two-stage process was implemented. In the first stage, a preliminary sample of schools was identified for which additional school-level data were collected. In the second stage, the final sample of matched intervention and comparison schools was selected for the evaluation. Each of these stages is explained below.

STAGE 1: FIRST ROUND OF SCHOOL SAMPLING AND MATCHING

The sampling frame for the evaluation was Jordan's Education Management Information System (EMIS) database, which is the most comprehensive list of schools in the country. Cohort 1 schools were dropped from the evaluation sampling frame (n = 666) because they had started implementing RAMP a year earlier and could not be included in the evaluation. Three hundred (300) schools that did not include all RAMP target grades were also excluded from consideration, leaving 1,028 schools in Cohort 2 (intervention group) and 679 schools in Cohort 3 (potential comparison group).

To select a sample of Cohort 3 schools that were as similar as possible to the intervention schools in Cohort 2, the evaluation team estimated a propensity score (PS) model for the likelihood of schools being in the RAMP intervention group, using sixteen independent school-level variables from the EMIS database (see full list in Table B.1). Propensity score matching with nearest neighbor matches (with replacement) were used to identify six comparison schools[1] that most closely resembled each intervention school based on the characteristics included in the propensity score model. All available intervention (n=1,028) and potential comparison schools (n=679) were included in the propensity score estimation process, although not all schools were selected for the evaluation sample.

The EMIS data on Cohort 2 and Cohort 3 schools were also used to create six strata based on shifting status and gender of the school's students (that is, single-shift all girls, single-shift all boys, and single-shift mixed-gender schools, and double-shift all girls, double-shift all boys, and double-shift mixed-gender schools). Then, a random sample of 160 schools was selected from the intervention group, proportionally to the size of each stratum. Then, using the propensity scores, a sample of 416 potential comparison schools that were similar to the intervention group was selected. The evaluation team had concerns about the quality of matches produced using EMIS data alone; therefore, school-level data were collected from a sample that was larger than the intended evaluation sample of 120 intervention and 120 comparison schools, to guard from school loses due to potential errors in the sampling frame. These data are referred to as the "verification data" in this report.

The evaluation aimed to select a final sample of 200 schools based on power calculations indicating that, in the best-case scenario[2], impacts on oral reading fluency and addition equivalent to at least 0.26 to 0.27 standard deviations could be detected. In a less favorable scenario[3], a sample of 200 schools afforded power to detect effect sizes as small as 0.37 to 0.39 standard deviations (for a full description of power calculations refer to the Evaluation Design Report, [MESP, 2016]). A sample of 240 schools (120 intervention and 120 comparison schools) was ultimately selected, to safeguard from school-level attrition.

TABLE B.1. SCHOOL CHARACTERISTICS USED IN THE FIRST-STAGE PROPENSITY SCORE MODEL

School Characteristics
1. Number of shifts in each school (single, double)
2. Whether the school was urban or rural
3. Gender of each school's students (all girls, all boys, mixed)
4. Highest grade in each school

1 More than one comparison school per intervention school was selected to increase the probability that all intervention schools had at least one suitable match once the data were verified. The evaluation team used propensity score matching with replacement, which means that a single comparison school could serve as a match for several intervention schools. For this reason, 416 comparison schools were identified as suitable matches for 160 intervention schools.

2 In the best-case scenario, R2 = .40; persistent factors explain 30 percent of the variation in students' outcomes, the ICC is .25, and attrition is 25 percent at the student level and 5 percent at the school level.

3 In a less favorable scenario, R2 = .20, persistent factors explain 20 percent of the variation in students' outcomes, the ICC is .44, and attrition is 50 percent at the student level and 10 percent at the school level.

School Characteristics
5. Lowest grade in each school
6. Whether or not the school had kindergarten
7. Number of sections in G1
8. Number of G1 students
9. Number of sections in G2
10. Number of G2 students
11. Number of sections in G3
12. Number of G3 students
13. Number of classrooms in each school
14. Percent of school staff who were teachers or administrative staff
15. Percent of the female school's teachers
16. School's student-teacher ratio.[4]

Verification data were not collected from 54 sampled schools (approximately 9 percent) because they had closed, did not include primary grades, served special needs or refugee camps (and were therefore outside the scope of the study), were unable to be reached for verification, or had merged with another sampled school. Of the 534 schools where verification data were collected (Figure B.1), 44 schools (eight percent) were excluded from the next stage because they were a duplicate shift for another school (under either the same or a different EMIS unique school identifier) or because they did not have teachers, students, and/or sections in one of the RAMP target grades, according to field reports.

This process left 490 unique schools eligible for the second stage of the matching and sampling process: 150 schools in the intervention group and 340 in the comparison group.

STAGE 2: SECOND ROUND OF SCHOOL MATCHING AND SAMPLING

Probability sampling was used to select a final intended sample of 120 intervention schools out of the 150 eligible schools. Schools were sampled at random, proportionally to the strata used in the first stage. Then, the evaluation team used a combination of EMIS and verification data to conduct 1:1 propensity score matching to select the final sample of 120 comparison schools out of the 340 potential comparison schools.

A list of key school characteristics used for second-stage matching is available in Table B.2. Updated measures of the characteristics used in the first round were used, as well as other characteristics that were considered important for the evaluation but that were not available in EMIS. Indicators for school strata, gender, number of shifts, and region were also included.

Propensity scores were estimated using the R package TWANG (Ridgeway et al, 2016). TWANG estimates the propensity score using boosting, a machine learning algorithm, instead of the

4 Before estimating the propensity score model, the research team top-coded student-teacher ratios in 24 schools at their 99th percentile to reduce the effect of large outliers.

traditional logistic regression. Boosted propensity score estimation (McCaffrey et al, 2004) uses a flexible estimation technique that adaptively captures the functional form of the relationship between the schools' characteristics and the intervention indicator with less bias than traditional approaches, such as logistic regression. In other words, the need for model selection is eliminated, while the need for variable selection is limited to the selection of the list of potential confounders/school characteristics. Boosting is an iterative process and at each iteration it selects the next term to be included in the model. The terms included at each iteration are functions of the schools' characteristics included in the model and their interactions terms, up to three-way interactions.

The boosting algorithm stops at the iteration that optimizes the balance of the covariates included in the model between the treated group and the comparison group. More specifically, TWANG allows for several criteria to measure balance between the two groups. In this application, the maximum Kolmogorov-Smirnov statistic was chosen as criterion, which means that the boosting algorithm stopped at that iteration that minimizes the largest Kolmogorov-Smirnov statistic.

The propensity score boosted logistic regression was fitted using the set of potential confounders/school characteristics listed in Table B.2. Among this set of potential confounders, the boosting algorithm found that strata indicators from the first stage, the average percent of overcrowding, whether the school offers free meal, the number of G3 students, and the percent of students who are poor mattered the most in achieving good balance between the schools in the intervention and comparison groups.

The second-stage matching process produced 120 comparison schools that were the best possible matches to the 120 sampled intervention schools, based on 40 school characteristics that were available in the EMIS and verification datasets. After implementing the matching procedure, there were no school characteristics for which the standardized bias percentage between the groups was greater than 25 percent.[5] Differences greater than 25 percent would have indicated that the groups were not similar enough to support unbiased inferences about the impact of RAMP (What Works Clearinghouse, 2014). For 28 out of the 40 measured characteristics, standardized bias was between 5 percent and 25 percent, indicating that statistical adjustments could be used to satisfy the equivalence requirement (see Table B.2).

TABLE B.2. MEANS AND STANDARDIZED BIAS AFTER MATCHING.

School Characteristic	Intervention Group Mean	Comparison Group Mean	Absolute % Bias After Matching	Interpretation
Number of Non-Classrooms in School	7.5	7.6	1.1	

5 Standardized bias percentage is calculated by dividing the difference in means between the two groups by the standard deviation of the entire sample and multiplying it by 100. The evaluation team used the standards established by The What Works Clearinghouse for interpretation. See http://ies.ed.gov/ncee/wwc/default.aspx. The What Works Clearinghouse is an initiative of the U.S. Institute of Education Sciences to evaluate studies on the effectiveness of programs, policies and practices. The What Works Clearinghouse Standards Briefs lay out rules to assess the quality of studies and are highly regarded in the field of program evaluations.

School Characteristic	Intervention Group Mean	Comparison Group Mean	Absolute % Bias After Matching	Interpretation
Percent of Staff who are Administrators	13.5%	13.4%	1.7	Satisfies equivalency requirement
Total Early Grade Teachers	6.9	7.0	1.9	
G3 Average Language Score in Previous Year	79.4	79.6	2.1	
Percent of Schools Receiving Donor Intervention (Not Training or Infrastructure-Related)	8.5%	9.2%	2.2	
Free Meal Provided	26.5%	27.5%	2.2	
G3 Math Average Score	79.2	79.0	2.3	
Sections of G3	1.9	1.9	3.5	
Percent of Schools with Kindergarten	31.6%	30%	3.5	
Index for School Repairs	15.9	16.2	4.3	
Year School Constructed	1987.5	1988.3	4.9	
Number of G3 Students	55.3	53.2	4.9	
Number of G1 Students	52.8	50.5	5	Requires statistical adjustment
Sections of G2	1.8	1.9	5.7	
Percent of Early Grade Students who are Female	51.3%	49.4%	6.3	
Sections of G3	1.9	1.9	6.5	
Number of G2 Students	54.7	51.6	6.6	
Number of Classrooms in School	13.1	12.7	6.7	
Maximum Early Grade Teacher Experience (In Years)	88.9%	83.3%	7.2	
Early Grade Student-Teacher Ratio	22.7	22.0	7.3	
Index of Water and Sanitation Non-Availability in School	12%	14.1%	8.1	
Percent of Schools Where Grade Students Use Library	52.1%	46.7%	10.9	
Percent Change in Enrollment from Beginning of Previous School Year	3.5%	1%	11.2	
Average Level of Overcrowding	114.6%	112.1%	11.8	
Percent of Students who are Syrian	9.4%	5.8%	12.1	
Maximum Grade in School	7.3	7.7	12.9	
G3 Minimum Language Score in Previous Year	54.4	53.1	13.7	
Percent of Schools Receiving Donor-Funded Training Intervention	13.7%	9.2%	14.2	
Percent of Schools that are Urban	53%	45.8%	14.3	
Percent of Students who are Poor	33.7%	37.4%	14.4	
Percent of Schools where Students Have Received Their Textbooks for the Year	94%	90%	14.9	
Percent of Early Grade Teachers with BA	95.1%	93.1%	16.9	
Percent of Schools Receiving Donor-Funded Infrastructure Intervention	46.2%	37.5%	17.5	

School Characteristic	Intervention Group Mean	Comparison Group Mean	Absolute % Bias After Matching	Interpretation
Percent of Schools Renovated in Previous 5 Years	29.9%	21.7%	18.9	
Percent of Schools with Donor-Funded Intervention (Any Type)	56.4%	46.7%	19.5	
Minimum Early Grade Teacher Experience (In Years)	4.1	4.9	19.6	
Percent Change in Enrollment Between Beginning and End of Previous Year	1.4%	2.2%	19.8	
Percent of Early Grade Teachers who are Women	78.9%	70.3%	20.1	
Length of School Day	5.5	5.6	20.3	
G3 Minimum Mathematics Score In Previous Year	55.3	53.2	21.7	
None	-	-	-	Does not satisfy equivalency requirement

ANNEX C. STUDENT SAMPLING AND SCHOOL AND STUDENT ATTRITION

SAMPLING OF STUDENTS FOR DATA COLLECTION

At baseline, one classroom per school was randomly selected from each grade, G1 and G2, for a total of two classrooms sampled per school. On average, 10 students per classroom (20 students per school) were randomly selected to take part in the reading and math assessments. The same students were assessed at midline, when they were about to finalize G1 and G2, and at endline, when they were at the end of G2 and G3, respectively. Students who had moved to another school were not followed or replaced. At endline, the number of students assessed per grade ranged from one to eleven due to varying classroom sizes.[6] On average, nine students were assessed at endline in both G2 and G3.

SAMPLE ATTRITION

Students were grouped within schools in this evaluation's design. Therefore, bias in the impact estimates could arise if schools dropped from the study, but also if students dropped from the study.

SCHOOL ATTRITION

At Endline, the research team was unable to collect data from two intervention schools that had shut down. This level of overall school attrition (1 percent) and differential attrition (2 percent) is considered low and is unlikely to lead to bias in the impact estimates greater than .05 standard deviations on the outcomes (What Works Clearinghouse, 2014).

TABLE C.1. SCHOOL SAMPLE SIZE AND ATTRITION RATES

Number of Schools	All Schools	Intervention	Comparison
Schools Sampled at Baseline	240	120	120
Schools Assessed at Endline	238	118	120
Attrition Rate at Endline (Percent)		2	0

Source: Endline Data 2018

Note: The attrition rate is estimated as the difference between the number of schools that were sampled at baseline and those who were assessed at follow-up, divided by the number schools sampled at baseline. Schools where data were collected from at least one grade are included in the number of schools assessed at endline.

6 According to field reports, some schools had fewer than 10 students available, so enumerators tested fewer than 10 students. In some schools, where the total number of students in the classroom was greater than 10 but less than 15, enumerators tested all the children rather than excluding only a few students.

STUDENT ATTRITION

At the student-level, the attrition rate at endline was below 20 percent in both groups and grades. More specifically, the number of students assessed at baseline was 4,681 (2,324 in the intervention group and 2,357 in the comparison group). The number of students assessed at endline was 3,893 (1,939 and 1,954 students in the intervention and comparison groups, respectively). Together, overall attrition (17 percent) and differential attrition (< 1 percent) represent a tolerable threat of bias under both optimistic and cautious assumptions, unlikely to lead to bias in the impact estimates greater than .05 standard deviations on the outcomes (What Works Clearinghouse, 2014). In other words, by strict standards, student attrition was not an important bias in this study and was lower than the anticipated attrition rate. The analysis sample used in this report consists of students who were assessed at both baseline and endline.

The student-level sample size power calculations completed at the evaluation-design stage assumed 25 percent student attrition and 5 percent attrition at the school level. Actual attrition at the student- and school-levels is below the levels of attrition anticipated in the statistical power calculations.

TABLE C.2. STUDENT SAMPLE SIZE AND ATTRITION RATES

	Grade 2 (Grade 1 at Baseline)			Grade 3 (Grade 2 at Baseline)		
Number of Students	**All Students**	**Intervention**	**Comparison**	**All Students**	**Intervention**	**Comparison**
Study Sample at Baseline[7]	2351	1162	1189	2330	1162	1168
Completed Assessments at Endline	1952	973	979	1941	966	975
Attrition Rate at Endline (Percent)	16.97	16.27	17.66	16.70	16.87	16.52

Source: Endline Data 2018

Note: The attrition rate is estimated as the difference between the number of children who had baseline data and those who were assessed at endline, divided by the number children sampled at baseline. Following guidelines by the What Works Clearinghouse (2014, pp. 14), attrition at the child level is based on the schools remaining in the sample. That is, the denominator only includes sampled children in the schools where data were collected at endline. There are three G3 students in the comparison group who were assessed at endline only and are not included in this table.

7 Excludes students from 2 schools that closed at endline.

ANNEX D. RE-ASSESSING THE SIMILARITY OF THE MATCHED INTERVENTION AND COMPARISON GROUPS USING STUDENT-LEVEL BASELINE DATA

Even though the matching procedure described in Annex B produced similar comparison and intervention groups with respect to key school-level characteristics, the evaluation team reassessed the similarity between the groups using baseline student-level data collected for the impact evaluation.

Baseline data analysis revealed that students in intervention schools had statistically significantly better performance than students in comparison schools in most reading and mathematics tasks. The magnitude of the differences, when they were statistically significant, ranged from 0.20 to 0.49 standard deviations, indicating that the groups did not satisfy baseline equivalence requirements (see the Baseline Report: RAMP Impact Evaluation [MESP, 2017]). The internal validity of evaluations that use quasi-experimental matching designs relies on finding a comparison group that is as similar as possible to the intervention group, except for exposure to the intervention. Therefore, the baseline differences found in this study would have prevented the attribution of post-intervention differences solely to RAMP. To address this issue, the evaluation team estimated student-level propensity score weights to increase the similarity between the students in the intervention and comparison schools.

STUDENT PROPENSITY SCORE WEIGHTS TO INCREASE SIMILARITY BETWEEN THE INTERVENTION AND COMPARISON GROUPS

Propensity score methods are increasingly being used to reduce the effects of baseline differences between groups when studies do not use a randomized-controlled design (Austin, 2011). The evaluation team estimated a propensity score (PS) model for the likelihood of students being in the intervention group using student baseline data. The model was fit using a boosted propensity score estimation approach (McCaffrey et al, 2004), which can predict the likelihood of being in the intervention group from a large number of covariates, while allowing for non-linear relationships between the covariates and the propensity score. Other methods require restricting the number of variables and therefore risk introducing bias due to the omission of covariates that may be important to modeling selection into the intervention group, or the misspecification of functional form of the relationship between covariates and selection into the intervention group. This approach has been shown to produce less bias than traditional approaches such as logistic regression (McCaffrey et al, 2004). Table D.1 lists all student characteristics included in the model.

The evaluation team used the propensity scores to reweight students in the comparison schools so that the weighted sample in the comparison group is more similar to the students in the intervention schools. The rightmost column in Table D.1 shows that the use of these weights effectively produced two groups of students that can be considered statistically equivalent based on strict equivalency requirements set by the What Works Clearinghouse (WWC) I. All remaining differences between the groups were relatively small. Consider, for example, that after weighting,

the largest difference among all characteristics equals 0.17 standard deviations (see line 2 for G2 students). This amounts to a difference of 8 percentage points in the number of girls in the intervention and comparison groups. The evaluation's impact estimation models adjust for characteristics where the effect size is greater than 0.05, as per WWC's guidelines.

TABLE D.1. COMPARISON OF STUDENT CHARACTERISTICS BETWEEN INTERVENTION AND COMPARISON GROUPS AT BASELINE, BEFORE AND AFTER WEIGHTING

		Before Weighting			After Weighting		
		Means		Difference Between Means (Effect Size)	Means		Difference Between Means (Effect Size)
		Intervention Group (T)	Comparison Group (C)		Intervention Group (T)	Comparison Group (C)	
	G1 Students						
G1 Student Characteristics							
1.	Age	6.2	6.2	0.06	6.2	6.2	0.00
2.	Female	55.4	50.9	0.09	58.3	50.8	0.15
3.	Attended Preschool	85.1	83.9	0.03	86.4	87.7	-0.04
4.	Attended the Same School the Year Before	17.3	23.3	-0.15	18.8	20.1	-0.03
5.	Missed One Day of School the Prior Week	51.2	53.7	-0.05	49.8	53.2	-0.07
6.	Had a Meal Before Getting to School	80.8	88.5	-0.22	81.1	83.5	-0.06
7.	Arabic is Primary Language Spoken at Home	98.7	94.1	0.25	98.7	98.0	0.06
8.	Parents Knew Last Time Student Received A Good Grade	98.1	96.9	0.08	98.0	97.3	0.04
G1 Reading Habits and Home Environment							
9.	Likes to Read	95.9	96.6	-0.04	95.7	96.4	-0.03
10.	Has Books to Read Other than Textbooks	60.3	56.7	0.08	60.0	60.0	0.00
11.	Reads with Other Kids or Parents at Home	84.2	82.6	0.05	85.0	84.9	0.00
12.	Often Reads Aloud to Another Person at Home	65.9	66.5	-0.01	65.1	65.7	-0.01
13.	Is Read to at Home	61.0	62.0	-0.02	61.1	61.2	0.00
14.	Does Math Problems at Home	96.0	96.2	-0.01	95.8	96.4	-0.03
15.	Gets Help with Homework at Home	88.7	88.8	0.00	89.3	89.1	0.01
16.	Takes Private After-School Lessons in Reading or Math	16.2	18.6	-0.06	17.2	16.7	0.01
G1 Student Reading at Baseline							
17.	Orientation to Print Raw Score (Out of 5)	2.4	2.1	0.25	2.5	2.4	0.07
18.	Phoneme Isolation Raw Score (Out of 10)	3.9	3.4	0.25	3.9	3.8	0.05
19.	Syllable Sound Raw Score (Out of 10)	4.8	3.1	0.40	4.8	4.4	0.09

		Before Weighting			After Weighting		
		Means		Difference Between Means (Effect Size)	Means		Difference Between Means (Effect Size)
		Intervention Group (T)	Comparison Group (C)		Intervention Group (T)	Comparison Group (C)	
20.	Letter Sound Raw Score (Out of 100)	22.0	19.6	0.17	22.8	22.4	0.02
21.	Reading Vocabulary Raw Score (Out of 10)	4.5	3.4	0.35	4.5	4.4	0.05
G1 Student Math Performance At Baseline							
22.	Count Numbers Raw Score (Out of 40)	28.0	25.3	0.25	28.6	28.0	0.06
23.	Enumerating Quantities Raw Score (Out of 10)	9.0	8.4	0.29	9.0	8.9	0.04
24.	Number Identification Raw Score (Out of 20)	28.3	25.0	0.24	28.8	28.2	0.04
25.	Number Discrimination Raw Score (Out of 10)	8.0	7.2	0.31	8.1	8.0	0.05
Number of Students		1,183	1,189	--	974	979	--
	G2 Students						
G2 Student Characteristics							
1.	Age	7.3	7.4	-0.15	7.32	7.32	-0.01
2.	Female	56.3	50.0	0.13	61.79	53.41	0.17
3.	Attended Preschool	82.9	77.6	0.14	83.49	85.41	-0.05
4.	Attended the Same School the Year Before	78.5	79.6	-0.03	79.93	79.53	0.01
5.	Missed One Day of School the Prior Week	44.7	48.7	-0.08	44.75	46.96	-0.04
6.	Had a Meal Before Getting to School	78.6	83.8	-0.14	78.22	80.62	-0.06
7.	Arabic is Primary Language Spoken at Home	98.8	97.0	0.12	98.74	98.76	0.00
8.	Parents Knew of Last Time Student Received a Good Grade	98.3	97.6	0.05	98.41	97.93	0.04
G2 Reading Habits and Home Environment							
9.	Likes to Read	96.6	96.4	0.01	97.1	98.7	-0.11
10.	Has Books to Read other than Textbooks	70.6	63.1	0.16	69.3	69.5	-0.01
11.	Reads with Other Kids or Parents at Home	85.3	85.4	0.00	85.3	85.4	0.00
12.	Often Reads Aloud to Another Person at Home	63.3	61.2	0.05	61.8	63.5	-0.04
13.	Is Read to at Home	57.0	66.1	-0.19	56.0	56.4	-0.01
14.	Does Math Problems at Home	98.6	96.4	0.14	98.4	98.4	0.00

		Before Weighting			After Weighting		
		Means		Difference Between Means (Effect Size)	Means		Difference Between Means (Effect Size)
		Intervention Group (T)	Comparison Group (C)		Intervention Group (T)	Comparison Group (C)	
15.	Gets Help With Homework at Home	89.8	90.0	-0.01	89.3	90.2	-0.03
16.	Takes Private After-School Lessons in Reading or Math	19.8	20.1	-0.01	19.7	17.9	0.05
G2 Student Reading at Baseline							
17.	Syllable Sound Raw Score (Out of 10)	5.2	3.5	0.46	5.3	4.8	0.13
18.	Letter Sound Raw Score (Out of 100)	28.3	26.8	0.08	29.3	30.0	-0.04
19.	Invented Word Raw Score (Out of 50)	7.9	5.6	0.33	7.8	7.7	0.02
20.	Reading Vocabulary Raw Score (Out of 10)	6.9	5.8	0.36	6.9	6.8	0.04
21.	ORF Passage Raw Score (Out of 41)	12.1	8.4	0.30	12.0	11.8	0.02
22.	Reading Comprehension Raw Score (Out of 6)	1.4	1.2	0.11	1.4	1.5	-0.09
G2 Student Math at Baseline							
23.	Number Identification Raw Score (Out of 20)	23.8	20.6	0.26	23.6	23.6	0.00
24.	Number Discrimination Raw Score (Out of 10)	7.7	7.0	0.27	7.6	7.6	0.01
25.	Missing Number Raw Score (Out of 10)	5.8	5.2	0.20	5.8	5.9	-0.03
26.	Addition L1 Raw Score (Out of 20)	8.9	7.4	0.29	9.0	8.7	0.06
27.	Addition L2 Raw Score (Out of 5)	2.3	1.9	0.23	2.4	2.2	0.08
28.	Subtraction L1 Raw Score (Out of 20)	6.2	5.0	0.29	6.3	6.0	0.06
29.	Subtraction L2 Raw Score (Out of 5)	2.2	1.8	0.20	2.3	2.1	0.06
	Number of Students	1,181	1,168	--	965	978	

Source: RAMP Impact Study – Baseline 2016 Student Assessments

Note: The table presents ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. The last three columns on the right-side show results from models including student propensity score weights for the analysis sample (students who were assessed in both baseline and endline). Effect sizes are the difference between the mean outcome of the intervention group and the mean outcome of the comparison group divided by the pooled within-group standard deviation of that outcome measure (WWC, 2017). Effect sizes > .05 (absolute value) are shaded. All comparisons before weighting include 240 schools. Comparisons after weighting include 238 schools. The table shows the maximum number of students included in the analyses. Sample sizes vary by outcome due to item non-response.

The evaluation team also tested whether there were differences in how boys and girls, and students in single- and double-shift schools, performed across the intervention and comparison groups at baseline. There were no statistically significantly differences in the performance of these subgroups, by study group. Students in the intervention and comparison groups were also statistically comparable in terms of the percentage who obtained zero-scores in the reading and mathematics subtasks (full results available upon request).

ANNEX E. MEASUREMENT INSTRUMENTS FOR THE IMPACT STUDY OF STUDENT LEARNING

Students' reading and math skills were measured using the Early Grade Reading Assessment (EGRA) and the Early Grade Math Assessment (EGMA). To contextualize the information gathered from student academic assessments, the team also collected student, teacher, and principal surveys. Next, a description is provided on how each of the surveys was developed, followed by a description of the EGRA/EGMA and assessors' training procedures.

STUDENT SURVEY

The research team developed a brief student survey to orally administer to students. The tool (Annex E.4) included 22 questions about students' educational background (for example, whether the student attended preschool, attended the same school the prior school year, and missed one day of school last week), home environment (for example the language commonly spoken at home, whether the student has someone to help with homework, and whether parents are aware of and encourage students' good academic performance), perception of emotional and instructional support from teachers (for example, whether the teacher praises students' good performance, how the teacher responds when the student cannot answer a question correctly, and whether the teacher reads stories in class), and reading and mathematics habits, including the literacy environment in the home (for example, whether the student likes to read, has books to read at home, reads with someone at home, and does math problems at home).

TEACHER SURVEY

The research team developed a 44-item teacher survey (Annex E.3) which included questions about teachers' background (including their gender, number of years teaching, ranking position, education level, and grade(s) they teach); professional development opportunities (for example, whether they had received pre-service or in-service training to teach reading and math); their classroom size and composition (specifically the number and gender of students), and the availability of resources for teaching (for example; whether they have a class library and enough reading books and textbooks for their students). The survey also included items on teaching practices (for example whether records are kept on students' attendance, if student performance is assessed, and what strategies are used to support students with difficulties and to manage disobedience). Additionally, the survey asked what support teachers receive from peers, administrators, and parents (for example, whether they cooperate with colleagues to prepare lesson plans, whom they consult with if they need help, and whether parents follow up on their children's homework), and the level of supervision from school administrators and other supervisors (for example, the number of times the level supervisor visited the school since the beginning of the school year, and the number of times the principal observed their classroom). The questionnaire took 30-60 minutes to administer, on average.

PRINCIPAL SURVEY

The research team developed a 51-item questionnaire for school principals (Annex E.2) that included questions about their background (for example their gender, education level, whether they received training in school administration) and administrative practices (including whether they keep records of students' attendance and monitor students' academic progress, what they do to manage underperforming teachers), as well as about the school's composition (for example, the school gender, grades taught, the number of periods/shifts, number of students per grade and by gender) resources (for example, whether the school has open spaces, shaded areas, and labs, whether it shares the building with another school, benefits from financial support, and received enough textbooks at the beginning of the academic year, and whether students feel safe in the school building), support from parents (for example, the existence of an active parents' committee and whether support from the committee is satisfactory), and level of supervision (for example, the number of times the supervisor visited the school for support or inspection in the last year). The questionnaire took 30-60 minutes to complete, on average.

INSTRUMENT DEVELOPMENT: EGRA AND EGMA

The EGRA and EGMA consist of a series of individual subtasks that measure foundational skills that contribute to the development of reading and mathematics. At endline, six EGRA tasks and six EGMA tasks were administered to second graders, and six EGRA tasks and seven EGMA tasks were administered to third graders. All tasks were administered orally in Standard Arabic and some were timed. Administration was 15-30 minutes on average. A brief description of each task, for each time point, is presented in Table E.1.

Subtasks were selected and adapted in consultation with 25 reading and Arabic language and math specialists, including 18 experts from the Ministry of Education (MoE).

INSTRUMENT DEVELOPMENT

Multiple instruments were used during RAMP evaluation. All instruments used with students, teachers and principals to answer Research Question 2 were developed in collaboration with the MoE from 21 to 25 August 2016. These include student tests for beginning Grade 1, beginning Grade 2/end of Grade 1, end of Grade 2 and end of Grade 3. For the RAMP evaluation, beginning Grade 1 and Grade 2 student tests were used at baseline in November 2016. End of Grade 1 and end of Grade 2 student tests were used at midline in May 2017. End of Grade 2 and end of Grade 3 student tests were used at endline in May 2018.

Student tests and questionnaires were field tested in 20 public schools in Jarash and Zarqa in October 2016. A mobile data application was developed by the Client Innovations team at MSI and was used by the data collection teams for the completion of the EGRA and EGMA and questionnaires. The instrument development process is described below.

EGRA/MA TOOLS AND QUESTIONNAIRES

The EGRA/MA tools were developed through a multi-step process that included:

- Selecting the EGRA and EGMA subtasks and developing content (items) appropriate for beginning of Grade 1, end of Grade 1/beginning of Grade 2, end of Grade 2 and end of Grade 3 students;
- Field testing (piloting) the tools in public schools;
- Analyzing the field test data to establish the validity and reliability of the tools and guide improvements to the content; and
- Reviewing and finalizing the tools.

The content of the EGRA/MA subtasks was drafted during a five-day tool development workshop led by MESP in collaboration with the MoE in August 2016. The item writers were introduced to the EGRA/MA methodology and examined the proposed EGRA/ME subtasks against the national curriculum to determine appropriate content levels needed for the RAMP evaluation at baseline, midline and endline. All subtasks were drafted at grade level, e.g. beginning or end of Grade 2. The item writers also agreed on questions to be used for the student, teacher, and principal questionnaires. In all, 25 reading and Arabic language specialists and math specialists participated in the tool development (August 2016) and tools revision and finalization (October 2016) workshops. Eighteen of the workshop participants were from the MoE.

In October 2016, the tools were field tested in 20 public schools in Jarash and Zarqa, which include Cohort 1 schools and are thus out of sample for this study. Enumerators and quality control officers were trained on EGRA/MA and questionnaire administration methods. The six-day training focused on understanding and practicing the content of the tools, including the introduction to the student, obtaining consent, the instructions for each subtask, and the use of the tablets and the data collection application.

Following the field test, the MESP team conducted data cleaning and analysis. The results were used during the tools revision and finalization workshop held in October 2016 with the MoE. Participants agreed on the subtasks to be retained and revised the items for the beginning Grade 1 and the beginning Grade 2 EGRA/MA tests in order to improve the validity and reliability of the subtasks and tools. The subtasks administered at each time point are shown in Table E.1.

MOBILE DATA COLLECTION APPLICATION

MSI specialists developed an electronic data collection application specifically for EGRA/MA in Jordan. The application was designed to: recognize Arabic language script, work online or offline, randomize the order of the tests (i.e. some students started with EGRA while others started with EGMA), synchronize data to a cloud database as soon as an internet connection was established, and feed into a virtual dashboard that summarized in real-time the collected data for monitoring purposes.

This application was loaded onto tablets purchased in Jordan for this study. The use of tablets facilitates the standardization of the test administration since it automates many features of the assessment such as the use of the timer for the timed subtasks and the auto-stops/discontinuation (triggered when a student incorrectly answers a certain number of items). The tablets also provide an additional layer of security by preventing leakage of the data and the content of the tools. The data collection application, tools, and questionnaires were field tested in Jordanian schools in October 2016 and were later finalized based on the field test experience.

TABLE E.1. MEASUREMENT INSTRUMENTS FOR THE IMPACT STUDY OF STUDENTS' LEARNING

Subtask	Subtask Description	Administered at Baseline		Administered at Midline		Administered at Endline		No. Items at Baseline	No. Items at Midline	No. Items at Endline	Task is Timed	Discontinuation / Administration Rule
		G1	G2	G1	G2	G2	G3					
Reading												
Orientation To Print	Knowledge of early print concepts such as directionality.	Yes	No	Yes	No	No	No	5	5	n/a	No	n/a
Phoneme Isolation	Ability to segment a word into individual phonemes.	Yes	No	Yes	No	Yes	No	10	10	10	No	Stop if student gets first 5 words incorrect
Syllable Segmentation	Ability to segment a word into individual syllables.	Yes	Yes	Yes	Yes	Yes	Yes	10	10	10	No	Stop if student gets first 5 words incorrect
Letter Sound Knowledge	Knowledge of letter-sound correspondence. Letters are presented in random order in both upper and lower cases.	Yes	Yes	Yes	Yes	Yes	Yes	100	100	100	Yes	Stop if student gets all items in the first line incorrect
Non-Word Decoding	Ability to decode individual non-words following common orthographic structure from grade-level text.	No	Yes	No	Yes	No	Yes	50	50	50	Yes	Stop if student gets all words in the first line incorrect

Subtask	Subtask Description	Administered at Baseline		Administered at Midline		Administered at Endline		No. Items at Baseline	No. Items at Midline	No. Items at Endline	Task is Timed	Discontinuation / Administration Rule
		G1	G2	G1	G2	G2	G3					
Reading Vocabulary	Receptive language skills of individual words and phrases related to body parts, common objects, and spatial relationships.	Yes	Yes	Yes	Yes	Yes	Yes	10	10	10	No	Stop if student gets first 5 images incorrect
Passage Reading	Ability to read a grade-level passage of approximately 60 words. It is scored for accuracy and rate.	No	Yes	Yes	Yes	Yes	Yes	41	52	52 in G2; 60 in G3	Yes	Stop if student cannot read any words correctly in the first sentence
Reading Comprehension	Ability to answer questions about the grade-level passage. Question types include explicit and inferential.	No	Yes	Yes	Yes	Yes	Yes	6	6	6	No	Not administered to children with auto-stop in passage reading. Children are not asked questions related to the passage reading sections they were unable to read.
Mathematics												
		G1	G2	G1	G2	G2	G3					
Counting Numbers	Ability to count out loud.	Yes	No	Yes	No	Yes	No	40	60	60	At endline,	Stop if student gives two

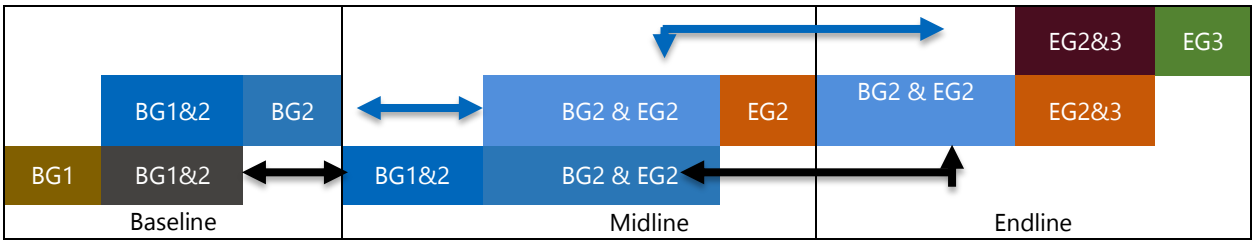
Subtask	Subtask Description	Administered at Baseline		Administered at Midline		Administered at Endline		No. Items at Baseline	No. Items at Midline	No. Items at Endline	Task is Timed	Discontinuation / Administration Rule
		G1	G2	G1	G2	G2	G3					
											students were given 90 seconds	incorrect answers. Stop if student stopped for more than 3 seconds for the second time
Counting Objects (Also Known As Enumerating Quantities)	Ability to count objects.	Yes	No	Yes	No	Yes	No	10	10	10	No	Stop if student gives incorrect answers for the first 5 items
Number Identification	Ability to identify numerals.	Yes	Yes	Yes	Yes	Yes	Yes	20	20	20	Yes	Stop if student gives incorrect answers for the first 5 items
Number Discrimination	Ability to compare numerical magnitudes.	Yes	Yes	Yes	Yes	Yes	Yes	10	10	10	No	Stop if student gives incorrect answers for the first 4 items
Missing Numbers	Ability to detect number patterns.	No	Yes	Yes	Yes	Yes	Yes	10	10	10	No	Stop if student gives incorrect answers for the first 4 items
Addition Facts – L1	Fluency (accuracy and speed) in solving one-digit addition problems.	No	Yes	Yes	Yes	Yes	Yes	20	20	20	Yes	Stop if student gives incorrect answers for the first 5 items
Addition Facts – L2	Ability to solve basic addition problems.	No	Yes	No	Yes	No	Yes	5	5	5	No	Not administered if L1 was auto-stopped. Stop if student gives incorrect answers

Subtask	Subtask Description	Administered at Baseline		Administered at Midline		Administered at Endline		No. Items at Baseline	No. Items at Midline	No. Items at Endline	Task is Timed	Discontinuation / Administration Rule
		G1	G2	G1	G2	G2	G3					
												for the first 4 items
Subtraction Facts – L1	Fluency (accuracy and speed) in solving one-digit subtraction problems.	No	Yes	No	Yes	No	Yes	20	20	20	Yes	Stop if student gives incorrect answers for the first 5 items
Subtraction Facts – L2	Ability to solve basic subtraction problems.	No	Yes	No	Yes	No	Yes	5	5	5	Yes	Not administered if L1 was auto-stopped. Stop if student gives incorrect answers for the first 4 items

ANNEX F. INSTRUMENT EQUATING PROCESS

In equating EGRA and EGMA tools from baseline to endline a Fixed Common Item Parameter (FCIP, Kim 2006) was used for phonemic isolation, reading vocabulary, syllable segmentation, and non-words subtasks in EGRA, and enumerating quantities, number identification, number discrimination, missing number, addition, and subtraction subtasks in EGMA. Thus, common or anchor items from the baseline and midline administration were identified during the endline calibrations, and their Item Response Theory (IRT) parameters were fixed to the baseline and midline values. This method results in all person and item parameters at endline being on the same scale as the baseline and midline. Below is a description of the assessment design used for test equating across the baseline, midline, and endline.

FIGURE F.1. ASSESSMENT DESIGN FOR VERTICAL TEST EQUATING



At baseline, there was a set of anchor items (BG1&2) between beginning of Grades 1 (BG1 in Gold) and 2 (BG2 in Blue) that helped putting the beginning of Grades 1 and 2 onto the same measurement scale. At midline, the same beginning of Grade 2 (BG2 in Blue) test from the baseline was used for end of Grade 1 and a new test (EG2 in Orange) was developed for end of Grade 2 but there was a set of anchor items (BG2&EG2) between beginning and end of Grade 2 that helped bringing the BG2 and EG2 onto the same measurement scale. Thus, the measurements from beginning of Grade 1 at baseline to end of Grade 2 at midline were on the same scale. At endline, the same end of Grade 2 (EG2 in Orange) test from the midline was used for end of Grade 2 and a new test (EG3 in Green) was developed for end of Grade 3 but there was a set of anchor items (EG2&3) between end of Grades 2 and 3 that helped putting the EG2 and EG3 onto the same measurement scale. As a result of the equating for the adjacent grade levels we constructed a vertically equated scale for beginning of Grade 1 to end of Grade 3. In other words, the measurements at baseline, midline, and endline were onto the same reporting scale.

DETERMINING THE SETS OF COMMON OR ANCHOR ITEMS

During the development stage of endline EGRA and EGMA tools, items that were also administered at baseline and midline were identified as potential anchor or common items. These items were designated based on the following criteria:

1. The average difficulty of the anchor items is about the same as the average difficulty of those items in the baseline and midline EGRA and EGMA tools.

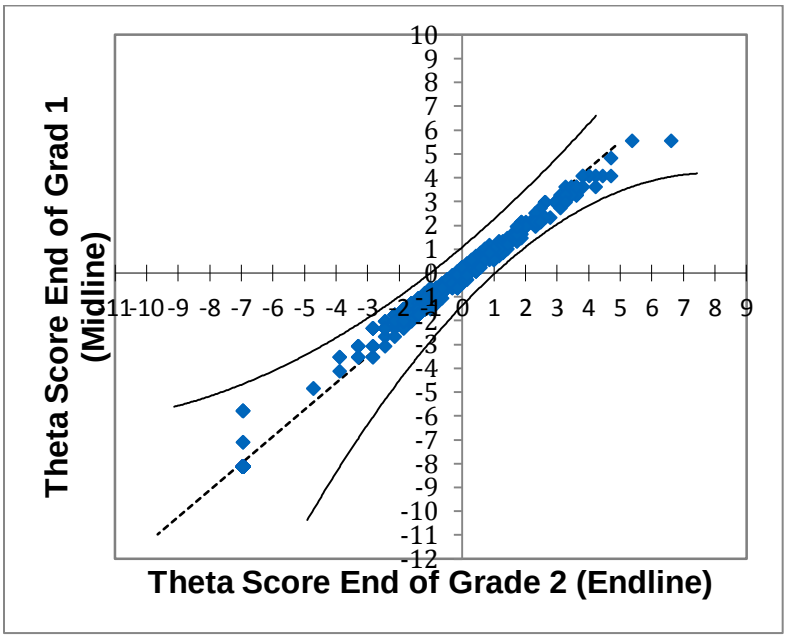
2. The total points from the anchor items are about equivalent to 30% of the total points on the test.
3. The position of each item at endline EGRA and EGMA tools is about the same as its position in the baseline and midline EGRA and EGMA tools.
4. There should not be any change in the item content from the baseline and midline to the endline EGRA and EGMA tools.

To determine the final set of anchor items for each grade level, a differential item functioning (DIF) approach using the delta method was applied. The difficulty levels (p-values) of each anchor item were transformed to the delta metric. Each item has two p-values, one for the baseline/midline and the other for the endline. The delta scale is an inverse normal transformation of percentage correct to a liner scale with a mean of 13 and standard deviation of 4 (Holland & Wainer, 1993). A high delta value indicates a difficult item. The delta values for the potential anchor items were computed for each subtask in each grade level. The delta values of the anchor items are then plotted: the endline values are on the X-axis and the baseline/midline values are on the Y-axis. A regression line is drawn and then the perpendicular distance of each item to the regression line are calculated. The items with less than two standard deviations away from the regression line are considered as operational anchor items and counted towards student score. The items with more than two standard deviations away from the regression line are not considered as operational anchor items. The difficulty levels of those items are not fixed for equating but are counted toward student score.

EQUATING OF ORAL READING PASSAGES AND COMPREHENSION ITEMS BETWEEN BASELINE, MIDLINE AND ENDLINE

In equating EGRA tools from baseline/midline to endline a common examinee method was used for oral reading passages (ORP) and comprehension subtasks. The ORPs and comprehension items used at the baseline/midline and the endline were administered to a representative sample of about 1,000 examinees at the endline to examine relative difficulty levels of the two passages and their corresponding comprehension items. Afterwards, separate IRT analyses were performed on the two ORPs to estimate examinees' ability estimates (also called theta scores). Then, two sets of theta scores along with the 95 percent confident band based upon the standard errors (Bond & Fox, 2001) are plotted in a scatter plot as shown below. The 95 percent confidence band provides a means to evaluate the extent to which the two ORPs were measuring the same construct within a reasonable degree of measurement error.

FIGURE F.1. RELATIVE DIFFICULTY LEVELS OF MIDLINE AND ENDLINE ORAL READING PASSAGES



If any examinees' paired theta scores were located outside the confidence band then those examinees should be removed from the analysis. Afterwards, the average theta score of each ORP was computed and the difference between the two ORP averages was obtained from subtracting one from the other. If the difference was zero, then no adjustment in the two item-examinee maps were needed. If the difference was not zero, some adjustment was necessary to make two ORPs commensurable. The above figure shows an example that the average theta of endline was smaller than that at the baseline by 0.11. Therefore, the item-examinee maps of the endline ORP was shifted to the right. After the adjustment, the mean difference theta score was 0 and thus the two ORPs were equivalent.

CREATING CONVERSION TABLES FOR BASELINE, MIDLINE, AND ENDLINE

After test equating was carried out based on either anchor items or common examinee method, the endline scores were converted into equivalent midline and baseline scores. The vertically equated endline scores of the examinees were then used to compare their performance at baseline and midline. An example of the conversion table for Grade 3 syllable segmentation is presented below.

TABLE F.1. EQUATED ENDLINE SCORES FOR SYLLABLE SEGMENTATION

Endline: Grade 3	Midline: Grade 2	Baseline: Beginning of Grade 2
0	0	0
1	2	3
2	3	4
3	4	5

4	5	6
5	6	7
6	7	8
7	8	8
8	9	9
9	9	9
10	10	10

The examinees who obtained a score of 3 out of 10 at endline has the same level of knowledge, skills, and abilities (KSA) in reading who had a score of 4 out of 10 at midline and 5 out of 10 at baseline. However, the examinees who obtained a score of 9 out 10 at endline would also obtain 9 out of 10 at both midline and baseline. After endline scores were converted into equivalent baseline and midline scores, the distribution of scores (by examinees) at endline, midline, and baseline were compared to assess change in reading outcomes over time due to the RAMP intervention.

ANNEX G. ANALYTIC APPROACH

STUDENT ASSESSMENT: IMPACT ESTIMATES

To take advantage of the longitudinal design of the evaluation and improve the precision of the estimates, the evaluation team fitted weighted ordinary least squares (OLS) regressions on students reading and math scores and adjusted for students' baseline scores. The analysis sample only included students who were assessed at both baseline and endline (see Annex C for details about the analysis sample). The weights were calculated at baseline and were the product of the school sampling weights, student sampling weights, and propensity score weights constructed to reduce imbalances between the intervention and comparison students found at baseline (the computation of school and student sampling weights is described next.) The regression model used to estimate endline impacts of RAMP on student outcomes is the following:

$$(1)\ Y_{is1} = \alpha + \beta_1 T_s + \beta_2 S1_{is0} + e_{is1}$$

where Y_{is1} is the outcome of interest (such as the score on a reading task) for student i in school s at endline. α is a constant term, T_j is an indicator equal to one for schools in Cohort 2 (the intervention group) and zero for those in Cohort 3 (the matched comparison group). Accordingly, the parameter β_1 is the coefficient of interest, which indicates the difference in the outcome between the intervention and the comparison groups. $S1_{is0}$ is a vector that represents the baseline (time 0) characteristics of student i in school s, including age, gender, and whether the student had a meal before school, attended preschool, and likes reading, as well as baseline reading and math scores. e_{is1} is a child-level residual. The regression uses robust cluster standard errors (RCSE) to account for the nesting of students in schools.

The evaluation team computed regression-adjusted weighted means of the relevant measures for each study group (intervention and comparison) and tested whether the differences in those adjusted means were statistically different from zero. P-values from the tests of differences in group means were used to assess statistical significance of the differences in means. The evaluation team defines the difference in means between treatment and control group as statistically significant whenever the p-value of the test is lower than 0.05.

In reporting and interpreting impact estimates, both the statistical significance and substantive significance are important. Statistical significance is the probability that the observed difference between the groups is due to chance, but it does not say anything about the magnitude of such difference. For this reason, standardized mean differences between the groups were calculated using Hedge's g, one of the most commonly used effect size indexes, defined as the difference between the mean outcome for the intervention group and the mean outcome for the comparison group, divided by the pooled within-group standard deviation of the outcome measure. As a rule of thumb, effect sizes of 0.25 standard deviations or larger are considered to be substantively important. The WWC recommends that effects sizes of such magnitude are interpreted as qualified positive or negative effects, regardless of whether they reach statistical significance (WWC, 2014).

The evaluation team also examined the program effects for subgroups defined by gender and the number of shifts in the school, to explore whether the impacts for girls and boys and for students in single- and double-shift schools differed. The regression model used to estimate impacts for subgroups is the following:

$$(2)\ Y_{is1} = a + b\theta_{is} + \beta_1 T_s + \beta_2 S1_{is0} + \mu\ \theta_{is} * T_s + e_{is1}$$

where Y_{is} is the outcome of interest (such as the score on a reading task) for student i in school s at endline. The variable θ_{is} is an indicator that takes the value of one for female students and zero for male students (or one for double-shift schools and zero for single-shift schools). The variable $\theta_{is} * T_s$ is the interaction of the gender (or number of shifts) variable and the intervention indicator. The other variables are the same as in the model for equation 1. The parameters of interest for the subgroup analysis are β_1—the impact of the intervention on males or single-shift schools— and $\beta_1 + \mu$ —the impact of the intervention on females or double-shift schools. This regression also uses RCSE and students' propensity score weights.

SAMPLING WEIGHTS SCHOOL SAMPLING AND MATCHING

As described in Annex B, sample selection took place in two stages. At each stage, intervention schools were randomly sampled, and comparison schools were selected via propensity score matching based on their similarity to intervention schools. Therefore, the computation of school sampling weights pertains to the intervention schools only, and all of our results would generalize to all Cohort 2 schools (intervention group).

The probability that a school i was selected in the second stage of sampling is given by:

$$(3)\ P(S2_i) = P(S1_i)\ P(S2_i\ |S1_i) = (\frac{n1}{N})(\frac{n2}{n1})$$

where $S1_i$ indicates that school i was selected in the first stage and $S2_i$ indicates that school i was selected in the second stage. N is the size of the sampling frame for the intervention schools. $n1$ is the sample size of the sample in stage 1 and $n2$ is the sample size of the sample in stage 2. Also, note that there were six different sampling probabilities, one for each of the sampling strata.

In theory, with a clean sampling frame (that is, no ineligible or out of scope schools) the sampling weights would simply be the inverse of the product of the sampling probabilities in the two stages of sampling.

In practice, however, the frame was not completely clean, and we found ineligible or out of scope schools at each stage (see the sample selection section above for more details). Therefore, in equation (3), $n1$ actually represents the number of eligible schools sampled. Similarly, N represents the number of eligible schools in the frame. We estimated that the number of eligible schools in the sampling frame was equal to the product of N times the stratum eligibility rate estimated in the sample.

The weights computed for the intervention schools were applied to their control matched schools. As we mentioned before, a comparison school was only selected because it matched a sampled

intervention school, therefore, each comparison school received the same weight as the intervention school it was matched to, so that the weighted results are generalizable to all Cohort 2 schools.

The unconditional weights for each student i in school j were computed using the following formula:

$$(4)\ Weight_{ij} = \frac{Total\ number\ of\ first\ (or\ second)\ grade\ students\ per\ school}{Number\ of\ first\ (or\ second) grade\ sampled\ in\ schools}$$

Conditional student weights were computed at baseline by dividing student weights by the school sampling weights previously described. The same weights were used at baseline and endline.

TESTING WHETHER RESULTS CHANGE WHEN BASELINE ADJUSTMENTS ARE NOT INCLUDED

In addition to the main analyses, for the impact study on students' learning the evaluation team estimated separate models that did not adjust for baseline math and reading outcomes. It was important to test whether results changed when baseline scores were excluded because baseline data collection was carried out approximately 6 to 8 weeks after the 2016-2017 school year began. Therefore, if baseline scores reflected early effects of RAMP, adjusting for baseline data could potentially introduce a downward bias in the impact estimates, making the differences between the groups look smaller than they really were. The analysis adjusts for baseline scores for three reasons: (1) Adjusting for baseline scores is required of quasi-experimental designs where pre-intervention differences between the groups would invalidate any conclusions about intervention impacts, (2) including baseline information can substantially increase the precision of impact estimates (Shochet, 2010), and (3) there was no persuasive evidence to suggest that RAMP had an impact on students' outcomes after only 6 to 8 weeks of implementation. Results from models that exclude baseline adjustments are presented in Annex P.

Also, the evaluation team originally proposed difference-in-differences (DiD) as the main framework to estimate intervention impacts. The main idea behind DiD is to estimate how outcomes change for students exposed to RAMP and compare this to how outcomes change for students in the comparison group. To produce valid impact estimates, a key assumption of DiD is that the outcomes were changing in a similar manner in both groups before the intervention. In other words, it is important that there are no preexisting differences in the trends of those outcomes. Unfortunately, it was not possible to test this assumption for the RAMP evaluation, because the Jordanian educational system does not use a standardized measure of student achievement in the early grades.

As mentioned above, the evaluation team opted to use OLS regressions that take advantage of the longitudinal design of the evaluation and to improve the precision of the impact estimates. This approach is an improvement upon a DiD analysis because it does not require that the groups have similar pre-intervention trends in the outcomes to produce unbiased impact estimates. Propensity score weights were calculated to ensure that impact inferences are based on groups

of students that are statistically equivalent at baseline, according to criteria set by the WWC (for more details about baseline equivalence see Annex D).

Finally, the impact's estimation model used RCSE to adjust for the nesting of students in schools. This approach differs from the multi-level modeling approach that was used to assess baseline equivalence for the RAMP Baseline Report. A recent study concludes that multi-level models and linear OLS models with RCSE produce similar policy conclusions (Kautz, Schochet & Tilley, 2017). However, OLS models were preferred for the estimation of midline and endline impacts to maximize the balance between groups that was achieved with the use of propensity score weights.

Overall findings are consistent, regardless of the specification used (See Annex O).

TEACHER OBSERVATION: DESCRIPTIVE ANALYSIS AND IMPACT ESTIMATES

The analysis of teacher outcomes in the descriptive and impact studies of teacher practice uses a slightly modified model and overall approach as the student analysis, which accounts for differences between the studies' samples and objectives.

THE DESCRIPTIVE STUDY OF TEACHER PRACTICES

The descriptive study, which was carried out at baseline and midline only, aimed to provide insights into teachers' practices in intervention versus comparison schools. The subsample of schools included in this study was revised between baseline and midline to improve its representativeness of varying levels of baseline student achievement, governorates, and strata. Therefore, regressions did not include the term for baseline characteristics. Regressions for outcomes based on the teacher observations were clustered at the teacher/classroom level to account for the nesting of multiple observation segments (between 1 and 3 per teacher and subject being taught) within classrooms; regressions for self-reported teacher characteristics were clustered at the school level to account for nesting of G1 and G2 teachers within schools. The models did not include weights because the subsample of schools was selected purposively based on the average levels of baseline student achievement in each school; schools were drawn equally from high, medium, and low-achieving schools. Thus, descriptive results from the teacher descriptive cannot be generalized to the overall population of Cohort 2 schools.

THE IMPACT STUDY OF TEACHER PRACTICES

For the impact study, regressions for outcomes based on teacher observations were clustered at the classroom level to account for the nesting of multiple observation segments within classrooms; standard errors for self-reported teacher outcomes were not clustered because the sample only included one teacher per school. The analysis incorporates school sampling weights, adjusted from the impact study of students' learning, to account for the stratified random subsample of 100 schools.

At endline, some teachers assessed in 2017 had left the school or were teaching another grade. To use data from all teachers observed in 2017 and 2018, the evaluation team estimated the

impacts of the intervention on teacher practices using a difference-in-differences (DID) framework, which allows for replacement teachers to contribute data at different time points. The model can be represented as follows

$$(5)\ y_{jt} = \propto + Post_t + \pi Interv_j + \rho Interv_j * Post_t + \vartheta Z_{jt} + \mu_j:$$

Where y_{jt} is the outcome of interest (for example, the average student participation score) for a teacher in school *j* in time *t*; $Post_t$ is a dummy variable where one represents the post-intervention period and zero otherwise; $Interv_j$ is a binary variable equal to one if the school was assigned to receive the intervention before the evaluation period and zero otherwise; Z_j is a vector of baseline school- and teacher-level characteristics that can be related to the outcome of interest but that are not expected to change as a result of the intervention (for example, school location or governorate); and μ_j is a school-specific random error term.

The parameter of interest in Equation (4) is ρ, the DiD estimate, which represents the average impact of the intervention on teacher practices, adjusting for other factors. This is an intent-to-treat estimate because not all teachers took advantage of the program (for example, they may have participated in RAMP training but used RAMP resources partially or not at all). Therefore, it can be interpreted as the impact of teaching in a school assigned to receive RAMP, regardless of actual take-up and implementation of RAMP.

Adopting a DiD framework allowed for the inclusion of all teachers observed at baseline and all teachers observed at endline, even if the same teacher was not observed at both time points. This model estimates the change in teaching practices (between baseline and endline) in the RAMP schools and then compares it to the change in practices in the comparison group. Therefore, the impact is estimated as the difference between treatment and comparison groups on the changes in teaching practices that occurred during the year. As noted above, this model does rely on the assumptions that the outcomes were changing in a similar manner in both groups before the intervention, which unfortunately could not be tested using available data. However, the baseline analysis confirmed that teacher characteristics and instructional practices were broadly similar across both groups; moreover, any changes in teacher characteristics between baseline and endline were also similar across groups (see Annex J). These two trends allow more confident attribution of differences at endline to the RAMP intervention.

ANNEX H. OBSERVER TRAINING AND OUTCOME MEASURE DEVELOPMENT FOR THE DESCRIPTIVE AND IMPACT STUDIES OF TEACHERS' PRACTICES

TRAINING, PILOTING, AND DATA COLLECTION

For the impact study of teacher practices and the April 2017 data collection of the descriptive study, the evaluation team led a six-day training of enumerators. The data collection team reviewed the COTI tool, practiced completing the tool using RAMP training videos, and spent four days conducting pilot observations in G3 classrooms in Amman and Zarqa[8], followed by debrief sessions.

Inter-rater reliability (IRR) was calculated among the three, four, or five observers who visited a given classroom on each day and results were discussed with the team. The evaluation team clarified the questions, responses, and definitions related to items that had low IRR in order to improve overall IRR across enumerators. A supplementary note-taking tool was introduced and used in schools after two days of piloting using COTI alone. Across the four days of piloting, the evaluation team achieved an average IRR of 84 percent. Data collection was conducted in April and May 2017, and April 2018.

OUTCOME DEVELOPMENT

The COTI tool is designed to capture a wide range of specific instructional practices and classroom characteristics. Estimating differences between groups on each of the more than 200 measured practices would be difficult to interpret. Therefore, the evaluation team selected a small number of indicators from the COTI tool to report on directly and used principal component analysis (PCA) to identify a smaller set of outcomes that could be reliably measured across the rounds of data collection. PCA is a statistical technique that extracts the most important information from a set of variables and creates a set of linear composite measures, up to the number of variables included. The measures, known as components, are statistically unrelated to one another.

The outcomes were developed using an iterative, multi-step process, based on the baseline data collected for the impact evaluation in April 2017.

The evaluation team developed weighted sums of related items from the COTI to produce summary measures of the frequency and variety of specific teacher practices such as assessing

8 To avoid interfering with data collection for the impact study of students' learning, the evaluation team ensured that none of the pilot schools overlapped with the evaluation sample: the evaluation team avoided piloting in sampled schools in Amman and Zarqa in Cohort 1, and these were therefore not part of the impact evaluation sample.

students, using materials during the lesson, and providing feedback to students, as well as classroom dynamics such as positive and negative student-teacher interactions.

Variables that were conceptually related to one another were grouped together. Then, the team standardized the variables, and reverse-coded (i.e. multiplied by -1) those that reflected a negative outcome, or absence of another measured behavior in the same construct.

Correlations were calculated among the variables in these constructs. Principal component analysis was performed on the pooled sample of math and reading observations using those conceptually related variables.

The evaluation team interpreted the results of the correlations and PCA and adjusted the variable groupings accordingly. The team focused on the first and second component of each construct, that is, the variables that were behaving in similar ways within larger groups. The team retained these variables and dropped those that split off into sub-groupings as well as those that were not correlated with any other variables. Groups where all variables showed low correlations were dropped as constructs, since they could not be reliably measured. Some of these variables were then tested as part of other constructs or added to the list of variables that would be analyzed separately.

Steps 3 and 4 were repeated until sets of variables were identified that were conceptually related, positively correlated, and where the first measured component explained at least 60 percent of the total variance in the scores.

The evaluation team used the final components to predict the scores for each segment observed and rescaled the resulting scores so that they had a minimum of zero and a maximum of 100. The team predicted the scores for the descriptive study using the same components that were generated for the impact study. The team did this so that the scores would be based on a large sample size and directly comparable with the impact evaluation results, but not be affected by RAMP implementation. One additional construct (RAMP implementation) was generated based don the descriptive study sample only.

At endline, outcomes were calculated and scaled to ensure direct comparability between baseline and endline results. The team accomplished this by standardizing input variables for PCA using the baseline mean and standard deviations and applying the original component weights to these standardized values. The evaluation team rescaled all resulting scores (baseline and endline together) so that they maintained a minimum of zero and a maximum of 100.

This process produced the following outcomes:

TABLE H.1. OUTCOME MEASURES FOR THE DESCRIPTIVE AND IMPACT STUDIES OF TEACHERS' PRACTICES

Outcome	Outcome Description
Instructional Practices	
RAMP Implementation (Descriptive Sample Only)	Teachers and/or students used RAMP worksheets, and the teacher demonstrated RAMP practices and integrated RAMP and MoE instructional practices
Time on Task (Not Computed with PCA)	The approximate percent of the segment time that was not spent on transitions, disruptions, or other non-instructional tasks.
Quality Instructional Practices	The extent to which the teacher: demonstrated adequate pacing that allowed students to learn, used scaffolding techniques to build on students' understanding of concepts, and clearly organized the lesson content.
Teacher Engagement Practices	The extent to which the teacher asked students questions and used speech, visual aids, and body language to emphasize the lesson content.
Teacher Encouragement	The extent to which the teacher used positive reinforcement of students' behavior and encouraged them to respond to his/her feedback
Teacher Feedback	The extent to which the teacher provided constructive feedback to students when they responded to questions or completed written work.
Use of Materials to Support Instruction	The extent to which the teacher used and let students use materials such as drawing materials, manipulatives, and the manipulatives supported the lesson.
Differential Instruction Practices	The extent to which the teacher used tailored instructional practices to students' different abilities and needs.
Student Engagement	The extent to which students were engaged in the lesson and with each other by clapping and writing on the chalk or white board.
Student Participation	The extent to which students collaborated on assignments, activities and answering the teacher's questions.
Positive Classroom Climate	The extent to which the teacher contributed to a negative learning environment by ignoring students, leaving them unsupervised, providing negative verbal and nonverbal reinforcement, and being sarcastic or verbally harsh.
Content Areas	Content areas were coded as a 1 if they included at least one of the following topics or activities, or 0 otherwise. In addition, the endline data allows estimation of the percent of time spent on each content area, and student participation in each.
Reading	
Character Reading and Identification	Identifying characters/symbols; pronouncing words with unreadable characters; paired pictures with character/sound/word work
Phonemic Awareness	Distinguishing/isolating sounds; phonemes with a high or low diacritic; phonemes with a short or long diacritic; distinguishing consonants; distinguishing short vowels; distinguishing long sounds; distinguishing difference between words ending with double vibrio and words ending with "N"; distinguishing / pronouncing words with double characters, short vowels; blending sounds and syllables; manipulating sounds; identifying words with same rhythm
Vocabulary	Word meaning; word families-replace some syllables to form new words; synonyms/antonyms; elaborating adjectives
Writing	Writing characters; writing words; writing phrases; writing sentences; writing multiple sentences; writing paragraphs; writing about a topic, characters and/or ideas in a book/text; using a book or other text as a model for writing; talked about grammar or spelling; functional writing; creative writing
Reading Comprehension	Discussed book title, author/illustrator, publication date, table of contents; predicted the story title based on an illustration; predicted the story title based on text; summarized the story or the information presented in the text; talked about the story subject and/or theme

Outcome	Outcome Description
	of the story; talked about key features of the story; talked about story structure (parts of a story/text); talked about characters in the text, who they are, their motivation; talked about the text
Mathematics	
Number Identification and Writing	Identifying numbers; writing numbers on board or paper; comparing numbers
Counting	Counting in ones; placing numbers on a number line; counting in rhymes and songs; counting in steps; counting small sets of counters in ones; counting out small groups of counters; estimating and counting larger sets of counters in ones; counting in groups; counting large sets of counters in groups
Basic Arithmetic	Single digit arithmetic; arithmetic with multiples of ten, hundreds, thousands; completing tens (hundreds and thousands), adding & subtracting multiples of ten; bridging tens (hundreds and thousands); doubling and halving; addition/subtraction problems (change, combine, and compare problems)
Advanced Arithmetic	Problems that support skills in division (sharing and grouping); multiplication problems (repeated addition, use of a grid or array type structure); problems for skills in fractions, ratio, rate, proportions; math thinking or calculating strategies
Classroom Structure	Classroom structure was coded as a 1 if the following was observed, and 0 otherwise
Whole Class	Teacher provided instruction or direction to the whole class
Large Group	Teacher provided instruction or direction to a group comprised of more than half the class, but less than the whole class, and/or a group of this size interacted with one another.
Small Group	Teacher provided instruction or direction to a group comprised of more than two students, but less than half the class, and/or a group of this size interacted with one another.
Partners/Pairs	Teacher provided instruction or direction to a group comprised of two students, and/or a group of this size interacted with one another.
Individual	Teacher provided instruction or direction to an individual student

Although they measure different instructional practices, the constructs produced by this process were closely correlated with one other, which supports that the COTI tool can be used to capture a hypothesized, underlying measure of teaching quality. Teachers that were higher ranked in one of the measures tended to be ranked higher, on average, on most other measures. Negative classroom climate, as expected, was negatively correlated with the other measures (Figure H.1). For example, the lowest 25 percent of teachers in terms of engagement practices across both subjects spent an average of 92 percent of the segment time on task, and, on average, scored in the lowest 28 percent for quality instructional practices and the lowest 19 percent for the use of materials. Meanwhile, the top 25 percent of teachers in terms of teacher engagement spent 98 percent of the segment time on task, scored higher than 67 percent of teachers for quality instructional practices, and higher than 65 percent of all teachers for the use of materials.

ANNEX I. ADDITIONAL TEACHER AND LESSON CHARACTERISTICS FROM THE DESCRIPTIVE STUDY OF TEACHERS' PRACTICES

At midline, around six months after teachers in the intervention group were trained by RAMP, 79 G1 and G2 teachers were observed in order to address research question 1a, "What practices do trained and non-trained G1 and G2 teachers implement with regard to teaching reading, writing, language and math?"

The following tables provide demographic information on the teachers who were observed, information on the RAMP training they received, and the results of the observations. This study was discontinued at endline because the completed impact study provides more rigorous evidence of RAMP implementation in the classroom. Because the evaluation team was unable to observe teachers before training to establish baseline equivalence, observation results should not be interpreted as evidence of RAMP's impacts, only suggestive of some of the influences RAMP may have had on instructional practices.

TEACHER CHARACTERISTICS

Overall, G1 and G2 teachers in both study groups and grades had similar background characteristics in terms of gender, years of teaching experience, and type of position (substitute versus permanent). Overall, the intervention and comparison groups were not significantly different from each other, although teachers in the intervention group were less likely to have a bachelor's degree and more likely to hold a higher diploma or graduate degree.

TABLE I.1. TEACHER CHARACTERISTICS IN THE DESCRIPTIVE SAMPLE AT MIDLINE

Variable	Intervention Group (T)	Comparison Group (C)	T minus C *(p-value)	Sample Size
Female (Percent)	80.0	84.6	-4.6 (0.71)	79
Years of Teaching Experience	11.3	11.0	.31 (0.87)	79
Teacher is a Substitute (Percent)	10.0	15.4	-5.4 (0.62)	79
Diploma (Percent)	10.1	10.2	-.13 (0.99)	79
Bachelor's Degree (Percent)	55.0	71.8	-16.9 (0.15)	79
Higher Diploma, Master's or PhD Degree (Percent)	35.0	18.0	17.0 (0.10)	79

Source: RAMP Descriptive Study of Teachers' Instructional Practices - Midline 2017 Data COTI Tool

Note: The table presents unweighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors at the school level. Columns T and C present group means for teachers in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison

groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. Sample size refers to the analytical sample used for difference testing.

*Difference in group means is statistically significant at the .05 level.

RAMP IMPLEMENTATION

The data revealed that 98 percent of intervention and 33 percent of comparison teachers reported participating in RAMP training. It is surprising that so many comparison teachers report RAMP training but based on differences between the average lengths of training reported in each group, it is likely that they were confusing the full training with a shorter RAMP training, which was implemented by the MoE across Jordan in 2016.

TABLE I.2. SELF-REPORTED RAMP TRAINING IN G1 AND G2 AT MIDLINE

Variable	Intervention Group (T) %	Comparison Group (C) %	T minus C * (p-value)	Effect Size	Sample Size
Received Training from RAMP (Percentage)	97.5	33.3	64.2* (0.00)	1.8	79
Of Teachers who were Trained, Percent Reporting:					
1-3 Days of Training	0.3	60.7	-60.5* (0.001)	NA	52
4-5 Days of Training	5.0	31.1	-26.1 (0.06)	-0.84	52
6-7 Days of Training	2.3	8.5	-6.2 (0.43)	-0.32	52
More Than 7 Days of Training	92.4	-0.4	92.8* (0.00)	NA	52

Source: RAMP Descriptive Study of Teachers' Instructional Practices - Midline 2017 Teacher Survey.

Note: The table presents unweighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. Columns T and C present group means for teachers in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the intervention and comparison group means at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. Sample size refers to the analytical sample used for difference testing. NA: Not calculated due to a lack of variation in one or both groups

*Difference in group means is statistically significant at the .05 level.

TABLE I.3. RAMP COACHING MEASURES IN G1 AND G2 AT MIDLINE

Variable	Intervention Group (T)	Comparison Group (C)	T minus C * (p-value)	Effect Size	Sample Size
Teachers Able To Get Mentoring And Coaching Support (Percent)	97.6	69.1	28.5* (.00)	0.81	79
Of Teachers who Attended RAMP Training, Percent Reporting:					
RAMP Coach Observed Classroom This School Year	89.6	25.5	64.1* (0.00)	1.83	51
Coach Observed Lesson 1-3 Times	12.4	26.3	-13.8 (0.34)	-0.37	51

Coach Observed Lesson 4-6 Times	43.7	-0.4	44.1* (0.00)	NA	51
Coach Observed Lesson More Than 6 Times	33.5	-0.4	33.8* (0.00)	NA	51

Source: RAMP Descriptive Study of Teachers' Instructional Practices - Midline 2017 Teacher Survey

Note: The table presents ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. Columns T and C present group means for teachers in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. Sample size refers to the analytical sample used for difference testing.

NA: Not calculated due to a lack of variation in one or both groups.

*Difference in group means is statistically significant at the .05 level.

Note that some comparison teachers reporting attending RAMP training were referring to a shorter introductory course rather than the full RAMP teacher training intervention.

TABLE I.4. OTHER RAMP IMPLEMENTATION MEASURES IN G1 AND G2 AT MIDLINE

Variable	Intervention Group (T)	Comparison Group (C)	T minus C * (p-value)	Effect Size	Sample Size
Ramp Implementation (All Teachers Observed)	25.1	9.85	15.2* (0.00)	0.77	79
Of Teachers Who Were Trained RAMP Training, Percent Reporting:					
Used The RAMP Coarse Grain Assessment Tools To Determine Student Level In The Current Year (Percent)	92.2	61.9	30.3 (0.09)	0.87	52
Used The RAMP Fine Grain Assessment Tool To Determine Student Level In The Current Year (Percent)	87.1	61.9	25.2 (0.16)	0.64	52

Source: RAMP Descriptive Study of Teachers' Instructional Practices - Midline 2017 Teacher Survey and COTI Tool.

Note: The table presents unweighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. Columns T and C present group means for teachers in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. Sample size refers to the analytical sample used for difference testing.

*Difference in group means is statistically significant at the .05 level.

CLASSROOM STRUCTURE

Class structure was similar in both groups. Small groups were more commonly used during math than during reading lessons.

TABLE I.5. CLASSROOM CHARACTERISTICS IN G1 AND G2 AT MIDLINE

Variable	Intervention Group (T)	Comparison Group (C)	T minus C* (p-value)	T math	T read	T (read-math)* (p-value)	C math	C read	C (read-math)* (p-value)	Math x T* (p-value)
Class Size	18.4	19.1	-1.8 (0.37)	11.9	21.8	9.9* (0.00)	13.7	21.5	7.9* (0.00)	2.1 (0.31)
Whole Class (%)	100.0	100.0	NA	100.0	100.0	NA	100.0	100.0	NA	NA
Large Group (%)	0.0	4.0	-7.3 (0.08)	1.0	1.0	NA	8.3	2.6	-5.7 (0.22)	5.7 (.22)
Small Group (%)	33.0	30.0	-6.5 (0.55)	35.4	31.8	-3.6 (.69)	41.9	22.4	-19.5* (0.02)	15.8 (0.19)
Partners/ Pairs (%)	17.6	10.0	5.5 (0.55)	25.5	16.8	-8.7 (0.25)	20	7.8	-12.2 (0.10)	3.5 (0.74)
Individual (%)	100.0	100.0	100.0	NA	100.0	100.0	NA	100.0	100.0	NA

Source: RAMP descriptive study of teachers' instructional practices - Midline 2017 COTI tool

Note: The table presents unweighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. T minus C is the difference in the means between the intervention and comparison groups at midline. T (read-math) and C (read-math) show the differences between math and reading lessons within the same group, and Math by T interaction is the differential effect of math and intervention status combined. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. Sample size refers to the analytical sample used for difference testing. N=191 lessons.
Math and reading observations pooled to allow hypothesis testing and within-classroom comparisons.
NA: Not calculated due to a lack of variation in one or both groups.
*Difference in group means is statistically significant at the .05 level.

LESSON CONTENT

Out of the five lesson content areas that were studied in both grades, there were few significant differences. G2 intervention teachers were nearly 20 percentage points more likely than G2 comparison teachers to include writing activities in the lesson. This was largely

driven by an increased focus on writing characters and words, rather than more complex writing such as sentences and paragraphs. There were also substantively large, but statistically insignificant differences; G1 teachers were less likely to focus on character reading and identification, and G2 teachers were more likely to focus on phonemic awareness.

TABLE I.6. READING LESSON CONTENT IN G1 AND G2 AT MIDLINE

Variable	Intervention Group (T) %	Comparison Group (C) %	T minus C * (p-value)	Effect Size	Sample Size
G1					
Character Reading and Identification	64.0	81.3	-17.3 (0.20)	-0.39	57
Phonemic Awareness	92.0	96.9	-4.9 (0.44)	-0.21	57
Vocabulary	20.0	12.5	7.5 (0.44)	0.20	57
Writing	96.0	96.9	-0.9 (0.86)	-0.05	57
Reading Comprehension	28.0	31.3	-3.2 (0.81)	-0.07	57
G2					
Character Reading and Identification	30.8	37.0	-6.30 (0.68)	-0.13	53
Phonemic Awareness	73.1	51.9	21.2 (0.16)	0.43	53
Vocabulary	30.8	33.3	-2.6 (0.86)	-0.05	53
Writing	96.2	77.8	18.4* (0.03)	0.54	53
Reading Comprehension	34.6	44.4	-9.80 (0.53)	-0.20	53

Source: RAMP Descriptive Study of Teachers' Instructional Practices - Midline 2017 COTI Tool

Note: The table presents unweighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. Columns T and C present group means for observation segments in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. Sample size refers to the analytical sample used for difference testing.

*Difference in group means is statistically significant at the .05 level.

There were no statistically significant differences in the four math lesson content areas in either G1 or G2, but there were some substantively large differences between groups. G2 intervention teachers were slightly more likely to cover foundational skills, such as writing and identifying numbers than comparison teachers. G2 intervention teachers were also more likely to cover advanced arithmetic (skills that support reasoning related to multiplication and division), although the reverse was true in G1.

TABLE I.7. MATH LESSON CONTENT G1 AND G2 AT MIDLINE

Variable	Intervention Group (T) %	Comparison Group (C) %	T minus C * (p-value)	Effect Size	Sample Size
G1					
Number Identification and Writing	85.0	90.5	-5.5 (0.61)	-0.16	41
Counting	65.0	71.4	-6.4 (0.67)	-0.13	41
Basic Arithmetic	100.0	100.0	NA	NA	41
Advanced Arithmetic	35.0	57.1	-22.1 (0.15)	-0.44	41
G2					
Number Identification and Writing	95.0	75.0	20.0 (0.12)	0.56	40
Counting	60.0	55.0	5.0 (0.75)	0.10	40
Basic Arithmetic	100.0	100.0	NA	NA	40
Advanced Arithmetic	95.0	80.0	15.0 (0.14)	0.44	40

Source: RAMP Descriptive Study of Teachers' Instructional Practices - Midline 2017 COTI Tool.

Note: The table presents unweighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. Columns T and C present group means for observation segments in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. Sample size refers to the analytical sample used for difference testing.

NA: Not calculated due to a lack of variation in one or both groups.

*Difference in group means is statistically significant at the .05 level.

INSTRUCTIONAL PRACTICES

Despite RAMP's focus on assessment, comparison teachers were more likely to use assessment practices during the lesson to support instruction than intervention teachers. This was the only finding out of the 12 measured instructional practices that was statistically significantly different between the two groups. However, there were some substantively large differences. Intervention teachers scored higher for student engagement practices and were less likely to demonstrate behaviors that cultivate a negative classroom climate.

TABLE I.8. INSTRUCTIONAL PRACTICES DURING G1 AND G2 READING LESSONS AT MIDLINE

	Intervention Group (T)	Comparison Group (C)	T minus C * (p-value)	Effect Size	Sample Size
Percent of Time on Task	96.4	93.9	2.5 (0.32)	0.23	110
Quality Instructional Practices	70.1	69.2	0.9 (0.82)	0.05	110
Teacher Engagement Practices	36.6	37.2	-.600 (0.85)	-0.04	110
Teacher Encouragement	41.2	33.0	8.20 (0.30)	0.22	110
Teacher Feedback	36.1	38.7	-2.6 (0.56)	-0.12	110
Use of Materials to Support Instruction	18.3	18.1	.2 (0.97)	0.01	110
Use of Assessment to Support Instruction	40.5	50.0	-9.5* (0.05)	-0.44	110
Differential Instruction	25.9	31.5	-5.60 (0.25)	-0.23	110
Student Engagement	83.7	72.7	11 (0.06)	0.38	110
Student Participation	22.8	25.3	-2.5 (0.78)	-0.06	110
Positive Classroom Climate	66.0	67.4	-1.4 (0.75)	-0.07	110
Negative Classroom Climate	14.7	21.4	-6.7 (0.18)	-0.30	110

Source: RAMP Descriptive Study of Teachers' Instructional Practices - Midline 2017 COTI Tool.

Note: The table presents unweighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. Columns T and C present group means for observation segments in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. Sample size refers to the analytical sample used for difference testing.

*Difference in group means is statistically significant at the .05 level.

There were no statistically significant impacts on instructional practices during math lessons. However, teachers in the intervention group spent an additional 4.1 percent of the observation segment time on task (i.e. time on instructional activities as opposed to transitions between activities, dealing with student disruptions, and other non-instructional tasks); this finding had a large effect size and was marginally significant (p=0.08).

TABLE I.9. INSTRUCTIONAL PRACTICES DURING G1 AND G2 MATH LESSONS AT MIDLINE

Variable	Intervention Group (T)	Comparison Group (C)	T minus C * (p-value)	Effect Size	Sample Size
Percent of Time on Task	98.0	93.9	4.1 (0.08)	0.47	79
Quality Instructional Practices	90.9	88.5	2.4 (0.58)	0.12	81
Teacher Engagement Practices	36.4	33.6	2.8 (0.38)	0.20	81
Teacher Encouragement	36.2	36.7	-.5 (0.95)	-0.01	81
Teacher Feedback	47.5	46.1	1.4 (0.79)	0.06	81
Materials to Support Instruction	32.4	25.3	7.10 (0.26)	0.25	81
Assessment to Support Instruction	46.8	45.4	1.4 (0.75)	0.07	81
Differential Instruction	28.1	27.5	.600 (0.92)	0.02	81
Student Engagement	78.8	85.3	-6.5 (0.29)	-0.22	81
Student Participation	32.5	31.6	.8 (0.93)	0.02	81
Positive Classroom Climate	65.9	67.4	-1.5 (0.75)	-0.07	81
Negative Classroom Climate	16.1	20.2	-4.10 (0.50)	-0.15	81

Source: RAMP Descriptive Study of Teachers' Instructional Practices - Midline 2017 COTI Tool

Note: The table presents unweighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. Columns T and C present group means for observation segments in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. Sample size refers to the analytical sample used for difference testing.

*Difference in group means is statistically significant at the .05 level.

ANNEX J. BASELINE EQUIVALENCY AND TEACHER REPLACEMENTS FOR THE IMPACT STUDY OF TEACHERS' PRACTICES IN G3

This annex describes three sets of results: (1) baseline equivalence results for the impact study of teachers' practices, (2) a discussion of teacher attrition and replacements between baseline and endline, and (3) an analysis of change in teacher demographic characteristics between baseline and endline, to test whether impact estimates may change due to changes in the sample.

BASELINE EQUIVALENCE IN G3 TEACHER PRACTICES

Baseline equivalence was assessed because (1) the original analytical plan was OLS-regression-based, which requires that impact inferences be based on groups that are statistically similar before the RAMP's introduction, and (2) evaluating overall instructional practices before RAMP is important context that helps in interpreting the results of the evaluation.

The baseline data showed that G3 intervention and comparison teachers were similar before RAMP was implemented. Teachers in both intervention and comparison classrooms had similar exposure to RAMP and used instructional time to cover similar types of lesson content in reading and mathematics (Tables J.1 and J.2).

The intervention and comparison groups were similar for most measures of classroom management and student engagement. Teachers in both groups spent the vast majority of lesson time on-task, practiced appropriate lesson pacing, and demonstrated few behaviors that cultivate a negative classroom climate. This suggests that the groups were relatively well balanced prior to RAMP (Tables J.3-J.5).

However:

- Intervention teachers had higher scores on quality instructional practices during reading lessons, which measured the pacing and organization of the lesson. This had a moderately large effect size and was statistically significant.
- Comparison teachers scored significantly better on assessment practices than comparison teachers in reading lessons. The difference was statistically significant and substantively large. The difference was not significant for assessment practices in math.
- Intervention teachers were more likely than comparison teachers to use materials to support reading and math instruction. The difference was statistically significant and substantively large.
- The groups were similar in terms of classroom structures. The only statistically significant difference in classroom structure was that 13 percentage points more intervention teachers had students work as pairs than comparison teachers.
- Both intervention and comparison teachers were significantly more likely to have students working in small groups during mathematics than during reading lessons.

TABLE J.1. READING LESSON CONTENT AT BASELINE

Variable	Intervention Group (T) %	Comparison Group (C) %	T minus C * (p-value)	Effect Size
Character Reading and Identification	36.4	39.7	-3.3 (0.59)	-0.07
Phonemic Awareness	59.8	60.0	-.2 (0.97)	0.00
Vocabulary	60.8	56.5	4.3 (0.47)	0.09
Writing	92.7	87.6	5.10 (0.13)	0.17
Reading Comprehension	59.3	53.5	5.80 (0.31)	0.12
Number of Lessons	135	161		
Number of Schools	100	100		

Source: RAMP Impact Study - Midline 2017 COTI Tool

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors to account for the clustering of observations within some classrooms. Columns T and C present group means for observation segments in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation.

TABLE J.2. MATHEMATICS LESSON CONTENT AT BASELINE

Variable	Intervention Group (T) %	Comparison Group (C) %	T minus C * (p-value)	Effect Size
Number Identification and Writing	70.8	63.4	7.4 (0.28)	0.16
Counting	48.3	44.3	4.0 (0.57)	0.08
Basic Arithmetic	44.0	32.4	11.6 (0.10)	0.24
Advanced Arithmetic	70.0	63.3	6.7 (0.31)	0.14
Number of Lessons	108	104		
Number of Schools	100	100		

Source: RAMP Impact Study - Midline 2017 COTI Tool

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors to account for the clustering of observations within some classrooms. Columns T and C present group means for observation segments in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation.

TABLE J.3. BASELINE INSTRUCTIONAL PRACTICES DURING READING LESSONS

Variable	Intervention Group (T)	Comparison Group (C)	T minus C * (p-value)	Effect Size
Percent of Time on Task	96.4	95.3	1.2 (0.33)	0.13
Quality Instructional Practices	86.4	79.8	6.6* (0.04)	0.24
Teacher Engagement Practices	35.8	34.3	1.4 (0.47)	0.09
Teacher Encouragement	48.7	51.4	-2.7 (0.59)	-0.07
Teacher Feedback	38.5	35.0	3.5 (0.26)	0.15
Materials	14.4	8.9	5.5* (0.03)	0.30
Assessment	39.2	48.0	-8.8* (0.002)	-0.41
Differential Instruction	28.8	30.5	-1.6 (0.64)	-0.06
Student Engagement	67.9	68.2	-.3 (0.95)	-0.01
Student Participation	16.4	11.9	4.5 (0.29)	0.14
Positive Classroom Climate	65.8	64.2	1.6 (0.56)	0.08
Negative Classroom Climate	15.8	20.0	-4.2 (0.22)	-0.17
Number of Lessons	135	161		
Number of Schools	100	100		

Source: RAMP Impact Study - Midline 2017 COTI Tool

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors to account for the clustering of observations within some classrooms. Columns T and C present group means for observation segments in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation.

*Difference in group means is statistically significant at the .05 level.

TABLE J.4. BASELINE INSTRUCTIONAL PRACTICES DURING MATH LESSONS

Variable	Intervention Group (T)	Comparison Group (C)	T minus C *	Effect Size
Percent of Time on Task	95.4	94.7	0.8 (0.57)	0.08
Quality Instructional Practices	87.6	87.8	-0.1 (0.97)	-0.01
Teacher Engagement Practices	42.1	44.6	-2.5 (0.29)	-0.15
Teacher Encouragement	49.3	51.7	-2.4 (0.66)	-0.06
Teacher Feedback	45.8	49.4	-3.6 (0.35)	-0.14
Materials	31.0	23.2	7.80* (0.03)	0.30
Assessment	42.0	48.5	-6.5* (0.04)	-0.30
Differential Instruction	29	33.2	-4.2 (0.23)	-0.17
Student Engagement	77.1	81.1	-3.9 (0.39)	-0.12
Student Participation	30.5	23.0	7.5 (0.18)	0.19
Positive Classroom Climate	65.3	66.3	-1.0 (0.73)	-0.05
Negative Classroom Climate	18.0	18.3	-0.4 (0.91)	-0.02
Number of Lessons	108	104		
Number of Schools	100	100		

Source: RAMP Impact Study - Midline 2017 COTI Tool

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors to account for the clustering of observations within some classrooms. Columns T and C present group means for observation segments in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation.

*Difference in group means is statistically significant at the .05 level.

The tables below show the percent of math and reading lessons in which teachers implemented different lesson structures. Analyses aimed to identify differences between the groups, as well as understand differences between the two subjects within the same classrooms prior to RAMP's implementation. While more teachers in both groups used small groups during math versus reading lessons, these patterns did not differ significantly between the intervention and comparison groups.

TABLE J.5. CLASSROOM STRUCTURES USED DURING MATH AND READING LESSONS AT BASELINE

Variable	Intervention Group (T)	Comparison Group (C)	T minus C* (p-value)	T Math	T Read	T (read-math)* (p-value)	C Math	C Read	C (read-math)* (p-value)	Math x T* (p-value)
Class Size	24.7	23.0	2.2 (0.06)	25.1	24.3	-0.7 (0.06)	22.8	23.1	0.3 (0.45)	-1.0 (0.06)
Class Structure (Percent of Lessons)										
Whole Class	98.8	99.3	1.0 (0.32)	100.0	97.8	-2.2 (0.09)	99.0	99.4	0.3 (0.77)	-2.5 (0.14)
Large Group	3.3	0.4	3.7 (0.10)	4.7	2.2	-2.5 (0.25)	1.0	0.0	-1 .0 (0.32)	-1.5 (0.52)
Small Group	29.0	26.8	2.7 (0.68)	37.3	22.5	-14.8* (0.001)	34.6	21.8	-12.8* (0.01)	-2.0 (0.75)
Partners/Pairs	22.1	10.9	12.6* (0.02)	24.0	20.6	-3.4 (0.41)	11.4	10.6	-0.8 (0.81)	-2.5 (0.63)
Individual	99.6	99.6	0.0 (0.98)	99.0	100	1.0 (0.32)	99.0	100.0	1.0 (0.32)	0.0 (0.99)
Number of Lessons	265	243		108	135		104	161		
Number of Schools	100	100		100	100		100	100		

Source: RAMP Impact Study - Midline 2017 COTI Tool

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors to account for the clustering of observations within some classrooms. Columns T and C columns present group means for observation segments in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively, overall and separately for math and reading. T minus C is the difference in the means between the intervention and comparison groups at midline. T (read-math) and C (read-math) show the differences between math and reading lessons within the same group, and Math by T interaction is the differential effect of math and intervention status combined. P-values are from tests of differences between group means and are shown in parentheses. Math and reading observations pooled to allow hypothesis testing and within-classroom comparisons.

*Difference in group means or interaction coefficient is statistically significant at the .05 level.

APPROACH TO SCHOOL-LEVEL ATTRITION BETWEEN BASELINE AND ENDLINE

One teacher each in the 200 sampled schools was observed at baseline. As discussed in Annex C, two sampled schools closed, and one stopped offering G3 between baseline and endline. To minimize the loss of power due to attrition, the evaluation team followed the

three teachers from these schools, who continued to teach G3 classes in other schools (students were not followed to other schools as part of the impact study on students' learning). The teachers were observed at their new schools and their data were analyzed as though they remained at their original posts. Power calculations for this study assumed 5 percent attrition at the school level, but actual attrition was lower than anticipated.

TEACHER-LEVEL REPLACEMENT BETWEEN BASELINE AND ENDLINE

The initial design of the impact study of teachers' practices planned to replace teachers who were no longer teaching in sampled schools. At endline, previously observed teachers in 69 of the 200 schools were unavailable because of changes in teaching assignments, retirement, and similar reasons (Table J.6). As planned, the evaluation team randomly selected another G3 teacher in the same school to observe at endline.

According to guidelines from the WWC (2015), attrition poses a low risk of bias to this study because despite the high overall attrition (35 percent), the differences between attrition rates in the intervention and comparison groups were relatively low (3 percentage points).

TABLE J.6. TEACHER SAMPLE SIZE AND REPLACEMENT RATES

	All Teachers	Intervention	Comparison
Teachers Observed at Baseline	200	100	100
Teachers Observed at Endline	131	67	64
Previously Observed Teacher Observed at Different School	3	2	1
Replacement Teacher Randomly Selected at Sampled School	69	33	36
Teacher Replacement Rate	34.5	33.0	36.0

Source: Endline Data 2018

Note: Replacement rate is estimated as the difference between the number of teachers who were observed at both baseline and endline, divided by the number of teachers observed at baseline.

CHANGES IN TEACHER DEMOGRAPHIC CHARACTERISTICS BETWEEN BASELINE AND MIDLINE

If the replacement of teachers between baseline and endline led to differential change in teacher characteristics that were related to their teaching practices, then the impact estimates may be influenced by these trends in addition to the influence of RAMP. The results in this section showed that teachers observed at both time points were similar in terms of their gender, teaching experience, and level of education (Table J.7). Intervention teachers at endline were more likely to hold a permanent teaching position relative to comparison teachers, but this difference was not statistically significant. Also, impact estimates for teacher practices were robust to including only teachers present at both time points, and to adding regression controls for teacher and lesson characteristics such as permanent teaching status, teachers' experience and education. Results from robustness tests are available upon request.

TABLE J.7. TEACHER CHARACTERISTICS AT BASELINE AND ENDLINE

Variable	T Group			C Group			Differential Change Over Time* (T2-T1)-(C2-C1)	Effect Size
	Baseline Mean (T1)	Endline Mean (T2)	Change Over Time (T2-T1)	Baseline Mean (C1)	Endline Mean (C2)	Change Over Time (C2-C1)		
Teacher Characteristics								
Female (%)	83.7	83.7	0.0	81.8	79.8	-2.0	2.0	0.08
Years Of Teaching Experience	9.4	10.3	0.8	10.1	11.1	0.9	-0.1	-0.03
Permanent Teaching Position (%)	84.4	87.4	3.0	89.1	82.9	-6.2	9.2	0.47
Teacher Education (% of Teachers)								
Diploma	4.9	3.0	-2.0	5.7	3.9	-1.8	-0.2	0.06
Bachelor's Degree	75.2	72.2	-2.9	72.8	75.6	2.8	-5.7	-0.18
Higher Diploma, Master's or PhD Degree	19.9	24.8	4.9	21.5	20.5	-1.0	5.9	0.17
Able to Get Mentoring and Coaching Support	78.3	95.1	16.8	70.2	93.0	22.9	-6.0	0.04
Number of Teachers/Schools	100	100		100	100			

Source: RAMP Impact Study – Midline and Endline Data, 2017/2018 COTI Tool

Note: Columns T1, T2, C1, and C2 present ordinary least squares regression-adjusted group means (or percentages) at each time point, with weights accounting for school sampling probabilities. The differential change over time column reflects the coefficient and significance of the difference-in-differences estimator. Effect sizes for continuous outcomes are Hedges' g. Effect sizes for dichotomous outcomes are the Cox index

ANNEX K. EXPLORATORY ANALYSIS OF RAMP IMPACTS ON SPECIFIC LESSON TOPIC COVERAGE DURING MATH AND READING LESSONS

Principal Components Analysis (PCA) and other data reduction techniques were employed to reduce the number of COTI teacher practices outcomes on which to perform impact analysis. Tables K.1 and K.2 show results from the individual math and reading lesson topics that were included in the COTI tool, and the group percentages at baseline and endline. The final summary outcomes were produced by taking the maximum of each category, such that it was 1 if an overall content area was covered during a lesson, and 0 otherwise. Table K.3 shows items related to teacher feedback that underlie the PCA-based composite of RAMP's positive impact on feedback during math lessons.

RAMP provided training to G3 teachers in the summer of 2017. The study followed teachers who were observed at baseline in spring 2017 continued teaching G3 during endline in spring 2018. If a teacher was no longer teaching G3, they were replaced with another current G3 teacher in 2018. The RAMP impact can be interpreted as difference between the intervention and comparison groups in the change over time in each outcome. This analysis of these individual items is exploratory because analyzing a large number of outcomes makes it likely that some will show impacts purely by chance. However, they provide insights into the specific lesson topics and teacher practices that contributed to impacts on composite outcomes.

TABLE K.1. RAMP IMPACTS ON READING LESSON TOPIC COVERAGE, BY CONTENT AREA

Variable	T Group			C Group			RAMP Impacts* (T2-T1)-(C2-C1)
	Baseline Mean (T1)	Endline Mean (T2)	Change Over Time (T2-T1)	Baseline Mean (C1)	Endline Mean (C2)	Change Over Time (C2-C1)	
Character Reading And Identification (% Of Lessons)							
Identifying Characters/Symbols	26.0	1.9	-24.1	24.4	9.9	-14.6	-9.6
Pronouncing Words With Unreadable Characters	17.3	9.3	-8.0	25.8	16.2	-9.5	1.6
Phonemic Awareness (% Of Lessons)							
Distinguishing/Isolating Sounds	11.0	11.5	0.5	13.6	13.7	0.2	0.4
Phonemes With A High Or Low Diacritic	14.5	12.7	-1.8	8.0	8.3	0.3	-2.1
Phonemes With A Short Or Long Diacritic	6.6	10.9	4.3	11.7	9.3	-2.5	6.8
Consonants	27.4	10.8	-16.6	31.5	12.8	-18.7	2.1
Short Vowels	5.9	0.9	-5.0	8.2	2.6	-5.6	0.6
Long Sounds	8.8	5.6	-3.2	5.5	9.0	3.4	-6.6
Double Vibrio, Words Ending With N	25.1	12.1	-13.0	32.3	15.5	-16.7	3.7
Double Character (Shadda), Short Vowel	15.3	6.5	-8.8	15.4	13.8	-1.7	-7.1
Blending Sounds And Syllables	12.5	8.9	-3.5	10.5	10.2	-0.3	-3.2

Variable	T Group			C Group			RAMP Impacts* (T2-T1)-(C2-C1)
	Baseline Mean (T1)	Endline Mean (T2)	Change Over Time (T2-T1)	Baseline Mean (C1)	Endline Mean (C2)	Change Over Time (C2-C1)	
Manipulating Sounds	10.2	12.3	2.1	10.7	1.8	-8.9	11.0*
Identifying Words With Same Rhythm	2.2	6.1	3.9	3.1	3.1	0.0	3.9
Vocabulary (% Of Lessons)							
Word Meaning	58.9	49.9	-9.0	48.4	49.0	0.7	-9.7
Word Families: Replace Some Syllables To Form New Words	5.3	30.9	25.6	3.1	6.0	2.9	22.7*
Synonyms	21.7	18.9	-2.8	14.9	10.3	-4.6	1.8
Elaborating Adjectives	11.9	6.6	-5.3	9.7	-0.0	-9.7	4.4
Writing Activities (% Of Lessons)							
Writing Characters	12.0	4.7	-7.3	13.0	5.4	-7.6	0.3
Writing Words	66.8	46.2	-20.6	65.4	54.4	-11.0	-9.6
Writing Phrases Or Sentences	43.5	49.8	6.3	30.7	57.4	26.8	-20.5*
Writing Multiple Sentences Or Paragraphs	15.0	10.3	-4.7	14.4	18.5	4.1	-8.8
Writing About A Topic, Characters And/or Ideas In A Book	8.1	3.2	-4.8	1.9	1.9	-0.0	-4.8
Using A Book Or Other Text As A Model For Writing	3.2	0.9	-2.3	3.5	-0.0	-3.5	1.3
Talked About Grammar Or Spelling	74.4	38.7	-35.7	75.6	50.0	-25.6	-10.1
Functional Writing	6.0	3.7	-2.3	9.1	12.5	3.4	-5.7
Creative Writing	-0.0	6.0	6.0	0.6	4.6	4.0	2.0
Reading Comprehension (% Of Lessons)							
Discussed Book Title, Author/Illustrator, Publication	0.7	6.9	6.2	1.4	1.0	-0.5	6.6*
Predicted The Story Title Based On An Illustration	3.6	4.6	1.0	0.6	4.8	4.2	-3.2
Predicted The Story Title Based On Text	12.9	9.4	-3.5	13.2	9.3	-4.0	0.5
Summarized The Story Or Text	25.2	27.6	2.4	28.9	24.5	-4.3	6.7
Talked About The Story Subject And/or Theme Of The Story	46.1	51.3	5.2	40.4	40.7	0.3	4.9
Talked About Story Structure (Parts Of A Story/Text)	24.5	30.4	5.8	18.3	14.7	-3.6	9.4
Talked About Characters In The Text	34.4	18.3	-16.1	34.7	10.1	-24.6	8.5
Talked About The Text	33.4	41.8	8.4	26.2	36.6	10.4	-2.0
Number Of Lessons	135	106		161	109		
Number Of Teachers/ Schools	100	100		100	100		

Source: RAMP Impact Study – Midline and Endline Data 2017/2018 COTI Tool

Note: The primary outcome measures to which each topic belongs are indicated in light blue headings. Columns T1, T2, C1, and C2 present ordinary least squares regression-adjusted group means (or percentages) at each time point, with weights accounting for school sampling probabilities. The RAMP Impacts column reflects the coefficient and significance of the difference-in-differences estimator. Errors are clustered at the classroom level to account for the clustering of observations within some classrooms

*Difference in differences is statistically significant at the .05 level.

TABLE K.2. RAMP IMPACTS ON MATH LESSON TOPIC COVERAGE, BY CONTENT AREA

Variable	T Group			C Group			RAMP Impacts* (T2-T1)-(C2-C1)
	Baseline Mean (T1)	Endline Mean (T2)	Change Over Time (T2-T1)	Baseline Mean (C1)	Endline Mean (C2)	Change Over Time (C2-C1)	
Number Writing and Identification (% of Lessons)							
Identifying Numbers	14.2	9.8	-4.4	8.4	10.8	2.4	-6.8
Writing Numbers on Board or Paper	69.0	40.8	-28.2	62.5	66.9	4.4	-32.7*
Comparing Numbers	5.6	4.0	-1.6	3.9	1.1	-2.9	1.2
Counting (% of Lessons)							
Counting in Ones	34.7	27.6	-7.1	28.1	41.6	13.5	-20.6*
Placing Numbers on a Number Line	2.7	5.3	2.5	3.7	5.3	1.6	1.0
Counting in Rhymes and Songs	2.7	2.9	0.2	2.8	8.0	5.1	-4.9
Counting in Steps	10.8	12.4	1.5	9.4	10.3	1.0	0.6
Counting Small Sets of Counters in Ones	4.6	1.0	-3.6	3.8	4.0	0.2	-3.8
Counting out Small Groups of Counters	0.9	1.0	0.1	0.0	3.0	3.0	-2.9
Estimating and Counting Larger Sets of Counters in Ones	0.9	2.0	1.1	1.0	1.0	0.0	1.1
Counting in Groups	2.7	9.9	7.1	2.9	6.0	3.1	4.1
Counting Large Sets of Counters in Groups	0.9	1.0	0.1	1.0	0.0	-1.0	1.1
Basic Arithmetic (% of Lessons)							
Single Digit Arithmetic	30.0	39.2	9.2	25.4	40.2	14.9	-5.7
Arithmetic with Multiples of Ten, Hundreds, Thousands	4.6	17.5	12.8	8.7	19.9	11.1	1.7
Completing Tens (Hundreds and Thousands)	0.9	1.0	0.1	2.0	2.0	0.1	-0.0
Bridging Tens (Hundreds and Thousands)	0.0	1.9	1.9	0.9	0.0	-0.9	2.8
Doubling and Halving	25.4	43.9	18.5	12.7	24.6	11.9	6.6
Addition/Subtraction Problems	6.6	7.9	1.3	1.9	3.0	1.0	0.2
Advanced Arithmetic (% of Lessons)							
Problems that Support Skills in Division	23.7	36.8	13.1	27.9	28.2	0.3	12.7
Multiplication Problems	15.7	17.0	1.3	19.4	17.8	-1.6	2.9

Variable	T Group			C Group			RAMP Impacts* (T2-T1)-(C2-C1)
	Baseline Mean (T1)	Endline Mean (T2)	Change Over Time (T2-T1)	Baseline Mean (C1)	Endline Mean (C2)	Change Over Time (C2-C1)	
Problems for Skills in Fractions, Ratio, Rate, Proportions	19.6	2.0	-17.5	9.4	0.0	-9.4	-8.2
Math Thinking or Calculating Strategies	46.4	52.7	6.3	50.9	43.8	-7.1	13.3
Number of Lessons	108	101		104	100		
Number of Teachers/ Schools	100	100		100	100		

Source: RAMP Impact Study – Midline and Endline Data 2017/2018 COTI Tool

Note: Columns T1, T2, C1, and C2 present ordinary least squares regression-adjusted group means (or percentages) at each time point, with weights accounting for school sampling probabilities. The RAMP Impacts column reflects the coefficient and significance of the difference-in-differences estimator. Errors are clustered at the classroom level to account for the clustering of observations within some classrooms.

*Difference in differences is statistically significant at the .05 level.

TABLE K.3. RAMP IMPACTS ON TEACHER FEEDBACK PRACTICES DURING MATH LESSONS

Variable	T Group			C Group			RAMP Impacts* (T2-T1)-(C2-C1)
	Baseline Mean (T1)	Endline Mean (T2)	Change Over Time (T2-T1)	Baseline Mean (C1)	Endline Mean (C2)	Change Over Time (C2-C1)	
Feedback When Students Asked Questions (%)							
General Feedback or Evaluative Feedback	89.3	100.0	10.7	97.1	96.6	-0.5	11.2*
Specific Feedback	65.0	70.0	5.0	62.6	61.6	-1.0	6.0
Strategic Feedback	37.9	59.5	21.6	49.2	39.9	-9.3	30.8*
Feedback When Students Completed Work (%)							
Corrected Answer Without Explaining Skill	9.9	21.5	11.6	11.9	20.5	8.6	3.0
General Encouragement	82.9	92.3	9.5	84.7	84.8	0.1	9.3
General Evaluative Feedback	56.2	70.9	14.7	59.4	70.2	10.8	3.9
Specific Feedback	36.4	43.9	7.6	35.0	29.5	-5.5	13.1
Strategic Feedback	29.6	47.1	17.4	33.5	27.6	-6.0	23.4*
Number of Lessons	108	101		104	100		
Number of Teachers/ Schools	100	100		100	100		

Source: RAMP Impact Study – Midline and Endline Data 2017/2018 COTI Tool

Note: Columns T1, T2, C1, and C2 present ordinary least squares regression-adjusted group means (or percentages) at each time point, with weights accounting for school sampling probabilities. The RAMP Impacts column reflects the coefficient and significance of the difference-in-differences estimator. Errors are clustered at the classroom level to account for the clustering of observations within some classrooms.

*Difference in differences is statistically significant at the .05 level.

ANNEX L. STUDENT READING HABITS, HOME ENVIRONMENTS, AND PERCEPTIONS OF TEACHERS' FEEDBACK AT ENDLINE

STUDENTS' READING HABITS AND HOME ENVIRONMENTS

In addition to examining impacts of RAMP on students' learning outcomes, the evaluation team explored endline impacts on students' reading habits and characteristics of their home environment, which may have changed in response to the community participation component of RAMP.

Similar to the results from baseline and midline, students in both grades reported reading habits and home environments that did not differ significantly between the two groups (Table L.1), except for the percentage of G2 students[9] who had books to read other than textbook. Students in the intervention group were eight percentage points more likely to report having books than comparison students and this difference was statistically significant. The groups did not differ in this respect at baseline or midline. It is possible that RAMP's community participation activities, which aimed to raise awareness about the importance of education, motivated parents to purchase more books for children to read at home. At midline, while 15 percent of G2 students in intervention schools reported taking private lessons in reading or math, only 10 percent did in comparison schools (results not shown). This difference was not found to be statistically significant either at endline or at baseline (after or before reweighting).

At endline, nearly all children in both groups and grades reported they liked to read; over 90 percent reported reading in class or the library, and at least 76 percent reported having books to read other than textbooks. Over 60 percent of children read with others or read aloud at home. In both groups and grades, almost 60 percent of children reported being read to at home, making it the least common activity of all activities measured. This may be due to students' age (at endline, students were 8 years old, on average). Nearly all children reported doing math problems at home and most children reported getting help with homework.

TABLE L1. G1 STUDENTS' READING AND MATH HABITS AT ENDLINE

Variable	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Effect Size	Number of Students
Likes to Read (Percentage)	97.4	97.7	-0.3	0.70	-7.6	1929
Has Books to Read Other Than Textbooks (Percentage)	74.7	74.1	0.7	0.81	2.1	1928
Reads with Other Kids or Parents at Home (Percentage)	85.2	84.1	1.0	0.68	4.9	1929

9 Who were in G3 at endline.

Variable	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Effect Size	Number of Students
Often Reads Aloud to Another Person at Home (Percentage)	62.5	63.8	-1.3	0.75	-3.5	1626
Is Read to at Home (Percentage)	58.8	57.1	1.7	0.63	4.2	1929
Does Math Problems at Home (Percentage)	97.8	98.0	-0.2	0.77	-6.5	1927
Gets Help With Homework at Home (Percentage)	87.3	89.7	-2.4	0.27	-14.4	1929
Takes Private After-School Lessons in Reading or Math (Percentage)	20.1	20.0	0.1	0.98	0.2	1926
Number of Schools	117	120				

Source: Endline Data 2018 Student Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes are calculated as Cox indexes. Sample sizes vary due to item non-response.

*Difference in group means is statistically significant at the .05 level.

TABLE L2. G2 STUDENTS' READING AND MATH HABITS AT ENDLINE

Variable	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Effect Size	Number of Students
Likes to Read (Percentage)	98.3	97.0	1.2	0.18	32.7	1929
Has Books to Read Other Than Textbooks (Percentage)	82.9`	75.6	7.3*	0.002	27.2	1930
Reads with Other Kids or Parents at Home (Percentage)	81.4	78.8	2.6	0.31	10.0	1928
Often Reads Aloud to Another Person at Home (Percentage)	68.8	62.5	6.3	0.09	16.8	1560
Is Read to at Home (Percentage)	58.1	58.7	-0.6	0.85	-1.5	1927
Does Math Problems at Home (Percentage)	98.3	98.0	0.3	0.76	11.1	1929
Gets Help with Homework at Home (Percentage)	88.6	91.0	-2.4	0.23	-15.7	1929
Takes Private After-School Lessons in Reading or Math (Percentage)	21.5	16.8	4.7	0.07	18.6	1927
Number of Schools	118	119				

Source: Endline Data 2018 Student Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes are calculated as Cox indexes. Sample sizes vary due to item non-response.

*Difference in group means is statistically significant at the .05 level.

STUDENTS' REPORTS ON TEACHER FEEDBACK

Students also answered questions about the type of feedback they received from teachers. Overall, there were no statistically significant differences between students in the intervention and comparison groups, in either grade. The only two exceptions were for asking someone for help when having difficulties reading in Arabic and reporting that the teacher did nothing when the student could not answer or answered a question incorrectly. G1 students[10] in the intervention group were 10 percentage points more likely to ask someone for help than comparison students, and G2 intervention students[11] were 10 percentage points more likely than comparison students to report that the teacher did nothing when a student could not answer or answered a question incorrectly. By far, the most common type of feedback reported in both groups and grades was praise (over 80 percent in both groups), followed by teachers' doing nothing, which was reported by at least half of students in both groups.

TABLE L.3. G1 STUDENTS REPORT OF TEACHER FEEDBACK

Variable	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Effect Size	Number of Students
Teacher does nothing when student performs well in class or tests	11.4	11.3	0.0	0.99	0.2	1845
Teacher praises student when they perform well in class or tests	86.5	84.8	1.7	0.44	8.2	1845
Teacher does something else when student performs well in class or tests	2.2	3.9	-1.7	0.07	-36.1	1845
Teacher punishes student when question not answered or answered incorrectly	29.8	26.9	2.9	0.39	8.6	1819
Teacher does nothing when student can't answer or answers question incorrectly	47.2	49.3	-2.1	0.61	-5.1	1819
Teacher corrects student when they can't answer or answer question incorrectly	16.6	16.4	0.2	0.95	0.7	1819
Teacher does something else when student can't answer question correctly	6.3	7.3	-1.0	0.59	-9.3	1819
Arabic language teacher reads stories in class	61.6	62.7	-1.0	0.80	-2.7	1929

10 Who were in G2 at endline.

11 Who were in G3 at endline.

Student asks someone for help when having difficulties to read in Arabic	47.7	37.5	10.2*	0.005	25.4	1922
Number of schools	117	120				

Source: Endline Data 2018 Student Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes are calculated as Cox indexes. Sample sizes vary due to item non-response.

*Difference in group means is statistically significant at the .05 level.

TABLE L.4. G2 STUDENTS REPORT OF TEACHER FEEDBACK

Variable	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Effect Size	Number of Students
Teacher does nothing when student performs well in class or tests	14.7	13.1	1.6	0.51	8.3	1843
Teacher praises student when they perform well in class or tests	83.6	85.0	-1.4	0.59	-6.3	1843
Teacher does something else when student performs well in class or tests	1.7	1.9	-0.3	0.71	-8.6	1843
Teacher punishes student when question not answered or answered incorrectly	21.3	25.0	-3.7	0.29	-12.7	1812
Teacher does nothing when student can't answer or answers question incorrectly	57.8	48.7	9.1*	0.029	22.2	1812
Teacher corrects student when they can't answer or answer question incorrectly	14.7	17.7	-3.0	0.38	-13.4	1812
Teacher does something else when student can't answer question correctly	6.2	8.6	-2.4	0.26	-21.2	1812
Arabic language teacher reads stories in class	76.0	68.5	7.5	0.08	22.9	1927
Student asks someone for help when having difficulties to read in Arabic	45.4	44.9	0.5	0.88	1.3	1924
Number of schools	118	119				

Source: Endline Data 2018 Student Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes are calculated as Cox indexes. Sample sizes vary due to item non-response.

*Difference in group means is statistically significant at the .05 level.

ANNEX M. ZERO-SCORES IN READING AND MATH

The evaluation team tested whether RAMP had an impact on the proportion of students who obtained scores equal to zero in the reading and mathematics subtasks. RAMP had few statistically significant impacts on the proportion of students who scored zero in each of the subtasks and in both grades (see Tables M1 to M4). However, RAMP had an unexpected negative impact on students' performance on G1 phoneme isolation, oral passage reading, and addition. Additional zero scores analysis is provided in Tables M1-5.

EGRA RESULTS: PERCENTAGE OF ZERO SCORES BY GRADE

RAMP had a negative impact on G1 students' performance on the phoneme isolation and passage reading subtasks. Specifically, a higher percentage of G1 intervention students than comparison students obtained a score equal to zero in these subtasks. The differences of three and six percentage points, respectively, were statistically significant. There were no other statistically significant differences between the groups (see Table M.1).

TABLE M.1. IMPACTS ON THE PERCENTAGE OF G1 STUDENTS WHO OBTAINED ZERO SCORES IN READING SUBTASKS AT ENDLINE

Variable	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Number of Students
Phoneme Isolation (Percentage)	8.6	5.4	3.2*	0.020	1931
Syllable Segmentation (Percentage)	21.3	20.5	0.8	0.78	1931
Letter Sound Knowledge (Percentage)	13.7	10.7	3.0	0.13	1931
Reading Vocabulary (Percentage)	5.8	5.6	0.2	0.85	1931
Passage Reading (Percentage)	23.8	17.8	6.0*	0.005	1931
Reading Comprehension (Percentage)	43.2	37.8	5.4	0.06	1931
Number of Schools	117	120			

Source: RAMP Impact Study - Endline Data 2018 Student Assessments

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups.

*Difference in group means is statistically significant at the .05 level.

RAMP did not have a significant impact on the percentage of G2 students who obtained a score equal to zero in the reading subtasks (see Table M.2).

TABLE M.2. IMPACTS ON THE PERCENTAGE OF G2 STUDENTS WHO OBTAINED ZERO SCORES IN READING SUBTASKS AT ENDLINE

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Number of Students
Syllable Segmentation (Percentage)	16.2	19.3	-3.1	0.25	1931
Letter Sound Knowledge (Percentage)	16.3	12.8	3.5	0.14	1931
Non-Word Decoding (Percentage)	31.8	33.7	-1.9	0.53	1931
Reading Vocabulary (Percentage)	1.2	2.0	-0.8	0.12	1931
Passage Reading (Percentage)	12.8	13.1	-0.3	0.86	1931
Reading Comprehension (Percentage)	23.4	22.4	1.1	0.63	1931
Number of Schools	118	119			

Source: RAMP Impact Study - Endline Data 2018 Student Assessments

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups.

EGMA RESULTS: PERCENTAGE OF ZERO SCORES BY GRADE

G1 students in the intervention group were significantly more likely than students in the comparison group were to obtain a score equal to zero in the addition subtask. Fourteen percent of intervention students obtained a score of zero, while 11 percent did in the comparison group. The difference of 3 percentage points was statistically significant. This difference may be driving the overall negative impact of RAMP on G1 students' addition (see Table 20). There were no significant differences between the groups in any other subtask (see Table M.3).

TABLE M.3. IMPACTS ON THE PERCENTAGE OF G1 STUDENTS WHO OBTAINED ZERO SCORES IN MATH SUBTASKS AT ENDLINE

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Number of Students
Counting Numbers (Percentage)	-0.1	0.5	-0.6	0.29	1931
Enumerating Quantities (Percentage)	0.8	0.5	0.3	0.56	1931
Number Identification (Percentage)	0.3	0.4	-0.1	0.66	1931
Number Discrimination (Percentage)	3.1	3.2	-0.1	0.90	1931
Missing Numbers (Percentage)	6.6	6.4	0.2	0.88	1931
Addition Facts - L1 (Percentage)	14.0	10.6	3.3*	0.041	1931
Number of Schools	117	120			

Source: RAMP Impact Study - Endline Data 2018 Student Assessments

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Home language and whether

the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups.

*Difference in group means is statistically significant at the .05 level.

There were no statistically significant differences between the intervention and comparison groups in the percentage of G2 students who obtained zero-scores in the math subtasks (see Table M.4).

TABLE M.4. IMPACTS ON THE PERCENTAGE OF G2 STUDENTS WHO OBTAINED ZERO SCORES IN MATH SUBTASKS AT ENDLINE

Variable (Total # Of Items)	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Number of Students
Number Identification (Out of 20, Prorated)	0.0	0.0	0.0	0.33	1931
Number Discrimination (Out of 10)	0.3	0.4	-0.1	0.68	1931
Missing Numbers (Out of 10)	2.3	1.9	0.4	0.53	1931
Addition Facts - L1 (Out of 20, Prorated)	12.7	11.3	1.5	0.39	1931
Addition Facts - L2 (Out of 5)	28.1	25.6	2.5	0.30	1931
Subtraction Facts - L1 (Out of 20, Prorated)	25.8	26.5	-0.7	0.78	1931
Subtraction Facts - L2 (Out of 5)	37.2	38.5	-1.3	0.68	1931
Number of Schools	118	119			

Source: RAMP Impact Study - Endline Data 2018 Student Assessments

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups.

*Difference in group means is statistically significant at the .05 level.

EGRA AND EGMA ZERO-SCORE TABLES BY SUBGROUP

GENDER

READING

TABLE M5. IMPACT ON ZERO SCORES IN READING FOR GRADE 1 STUDENTS AT ENDLINE, BY GENDER

Variable (Total # of Items)	Impact for Girls (A)	*P*-Value	Impact for Boys (B)	*P*-Value	Difference in Ramp Impacts by Gender (P-Value)
Phoneme Isolation	4.0*	0.023	2.3	0.25	0.50
Syllable Segmentation	-1.1	0.71	3.0	0.45	0.35
Letter Sound Knowledge	6.8*	0.012	-1.4	0.60	0.026*
Reading Vocabulary	0.4	0.76	0.0	1.00	0.87

Passage Reading	4.4	0.06	7.8*	0.022	0.38
Reading Comprehension	6.8	0.05	3.6	0.34	0.48
Number of Students	**988**		**943**		
Number of Schools	**193**		**192**		

Source: Endline Data 2018 Student Assessments

Note: Columns A and B present the regression-adjusted difference on the percentage of girls (or boys) who obtained a score equal to zero in the intervention and comparison groups. The last column shows the p-value for the gender by treatment interaction. Regressions adjust for baseline reading and math scores as well as children's gender, whether the child had a meal before getting to school, liked to read, attended preschool, spoke Arabic as primary language, had books to read other than textbooks, was read to at home, and did math problems at home. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. There are fewer schools than in analyses for the overall sample because single-gender schools only contribute to estimating the impacts on girls or boys. There are 36 female-only schools and 48 male-only schools in the sample.

*Difference in group means is statistically significant at the .05 level.

TABLE M6. IMPACT ON ZERO SCORES IN READING FOR GRADE 2 STUDENTS AT ENDLINE, BY GENDER

Variable (Total # of Items)	Impact for Girls (A)	*P*-Value	Impact for Boys (B)	*P*-Value	Difference in RAMP Impacts by Gender (P-Value)
Syllable Segmentation	-1.9	0.49	-4.7	0.22	0.45
Letter Sound Knowledge	3.3	0.23	3.7	0.30	0.92
Non-Word Decoding	-6.6	0.07	4.3	0.29	0.029*
Reading Vocabulary	-0.2	0.70	-1.7	0.12	0.18
Passage Reading	-2.2	0.29	2.1	0.52	0.25
Reading Comprehension	-3.7	0.20	7.3*	0.041	0.021*
Number of Students	**994**		**937**		
Number of Schools	**191**		**178**		

Source: Endline Data 2018 Student Assessments

Note: Columns A and B present the regression-adjusted difference on the percentage of girls (or boys) who obtained a score equal to zero in the intervention and comparison groups. The last column shows the p-value for the gender by treatment interaction. Regressions adjust for baseline reading and math scores as well as children's gender, whether the child had a meal before getting to school, liked to read, attended preschool, spoke Arabic as primary language, had books to read other than textbooks, was read to at home, and did math problems at home. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. There are fewer schools than in analyses for the overall sample because single-gender schools only contribute to estimating the impacts on girls or boys. There are 36 female-only schools and 48 male-only schools in the sample.

*Difference in group means is statistically significant at the .05 level.

TABLE M7. IMPACT ON ZERO SCORES IN MATH FOR GRADE 1 STUDENTS AT ENDLINE, BY GENDER

Variable (Total # of Items)	Impact for Girls (A)	*P*-Value	Impact for Boys (B)	*P*-Value	Difference in RAMP Impacts by Gender (P-Value)
Counting Numbers	-1.1	0.29	0.0	0.95	0.28
Enumerating Quantities	0.0	0.97	0.6	0.53	0.50
Number Identification	0.1	0.45	-0.4	0.27	0.12
Number Discrimination	0.7	0.58	-1.0	0.33	0.27
Missing Numbers	-0.1	0.97	0.5	0.80	0.82
Addition Facts - L1	2.8	0.17	4.0	0.10	0.68
Number of Students	**988**		**943**		
Number of Schools	**193**		**192**		

Source: Endline Data 2018 Student Assessments

Note: Columns A and B present the regression-adjusted difference on the percentage of girls (or boys) who obtained a score equal to zero in the intervention and comparison groups. The last column shows the p-value for the gender by treatment interaction. Regressions adjust for baseline reading and math scores as well as children's gender, whether the child had a meal before getting to school, liked to read, attended preschool, spoke Arabic as primary language, had books to read other than textbooks, was read to at home, and did math problems at home. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. There are fewer schools than in analyses for the overall sample because single-gender schools only contribute to estimating the impacts on girls or boys. There are 36 female-only schools and 48 male-only schools in the sample.

*Difference in group means is statistically significant at the .05 level.

TABLE M8. IMPACT ON ZERO SCORES IN MATH FOR GRADE 2 STUDENTS AT ENDLINE, BY GENDER

Variable (Total # of Items)	Impact for Girls (A)	*P*-Value	Impact for Boys (B)	*P*-Value	Difference in RAMP Impacts by Gender (P-Value)
Number Identification	0.0	0.78	0.0	0.31	0.32
Number Discrimination	0.1	0.68	-0.3	0.55	0.50
Missing Numbers	0.2	0.82	0.8	0.53	0.66
Addition Facts - L1	0.5	0.80	2.7	0.31	0.50
Addition Facts - L2	0.0	1.00	5.7	0.11	0.16
Subtraction Facts - L1	-1.9	0.56	0.9	0.77	0.50
Subtraction Facts - L2	-2.9	0.43	0.8	0.85	0.44
Number of Students	**994**		**937**		
Number of Schools	**191**		**178**		

Source: Endline Data 2018 Student Assessments

Note: Columns A and B present the regression-adjusted difference on the percentage of girls (or boys) who obtained a score equal to zero in the intervention and comparison groups. The last column shows the p-value for the gender by treatment interaction. Regressions adjust for baseline reading and math scores as well as children's gender, whether the child had a meal before getting to school, liked to read, attended preschool, spoke Arabic as primary language, had books to read other than textbooks, was read to at home, and did math problems at home. Home language and whether

the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. There are fewer schools than in analyses for the overall sample because single-gender schools only contribute to estimating the impacts on girls or boys. There are 36 female-only schools and 48 male-only schools in the sample.

*Difference in group means is statistically significant at the .05 level.

SHIFT

READING

TABLE M9. IMPACT ON READING PERFORMANCE FOR GRADE 1 STUDENTS AT ENDLINE, BY NUMBER OF SCHOOL SHIFTS

Variable (Total # of Items)	Impact for Single Shift(A)	*P*-Value	Impact for Double Shift (B)	*P*-Value	Difference in RAMP Impacts by Number pf Shifts (P-Value)
Phoneme Isolation	2.8	0.09	6.0*	0.003	0.21
Syllable Segmentation	2.1	0.49	-2.6	0.64	0.45
Letter Sound Knowledge	2.2	0.33	6.0	0.19	0.45
Reading Vocabulary	1.4	0.28	-4.0	0.14	0.07
Passage Reading	8.0	0.001	-1.2	0.80	0.08
Reading Comprehension	8.2*	0.008	-5.5	0.38	0.050*
Number of Students	**1643**		**288**		
Number of Schools	**199**		**38**		

Source: Endline Data 2018 Student Assessments

Note: Columns A and B present the regression-adjusted difference on the percentage of students in single- or double-shift schools who obtained a score equal to zero in the intervention and comparison groups. The last column shows the p-value for the number of shifts by treatment interaction. Regressions adjust for baseline reading and math scores as well as children's gender, whether the child had a meal before getting to school, liked to read, attended preschool, spoke Arabic as primary language, had books to read other than textbooks, was read to at home, and did math problems at home. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups.

*Difference in group means is statistically significant at the .05 level.

TABLE M10. IMPACT ON ZERO SCORES IN READING FOR GRADE 2 STUDENTS AT ENDLINE, BY NUMBER OF SCHOOL SHIFTS

Variable (Total # of Items)	Impact for Single Shift(A)	*P*-Value	Impact for Double Shift (B)	*P*-Value	Difference in RAMP Impacts by Number Of Shifts (P-Value)
Syllable Segmentation	-0.6	0.83	-11.9*	0.047	0.09
Letter Sound Knowledge	6.1*	0.015	-9.0	0.13	0.021*
Non-Word Decoding	-0.5	0.88	-7.2	0.32	0.40
Reading Vocabulary	-0.9	0.17	0.0	0.98	0.29
Passage Reading	0.6	0.78	-2.8	0.41	0.40
Reading Comprehension	1.9	0.44	-0.2	0.97	0.68

Number of Students	**1636**		**295**		
Number of Schools	**198**		**39**		

Source: Endline Data 2018 Student Assessments

Note: Columns A and B present the regression-adjusted difference on the percentage of students in single- or double-shift schools who obtained a score equal to zero in the intervention and comparison groups. The last column shows the p-value for the number of shifts by treatment interaction. Regressions adjust for baseline reading and math scores as well as children's gender, whether the child had a meal before getting to school, liked to read, attended preschool, spoke Arabic as primary language, had books to read other than textbooks, was read to at home, and did math problems at home. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups.

*Difference in group means is statistically significant at the .05 level.

MATH

TABLE M11. IMPACT ON ZERO SCORES IN MATH FOR GRADE 1 STUDENTS AT ENDLINE, BY NUMBER OF SCHOOL SHIFTS

Variable (Total # of Items)	Impact for Single Shift(A)	*P*-Value	Impact for Double Shift (B)	*P*-Value	Difference in RAMP Impacts by Number of Shifts (P-Value)
Counting Numbers	-0.1	0.57	-2.9	0.27	0.27
Enumerating Quantities	0.1	0.83	0.9	0.16	0.31
Number Identification	-0.3	0.24	0.6	0.22	0.12
Number Discrimination	0.5	0.58	-2.1	0.48	0.40
Missing Numbers	0.5	0.74	-0.5	0.86	0.75
Addition Facts - L1	5.5*	0.001	-4.2	0.33	0.035*
Number of Students	**1643**		**288**		
Number of Schools	**199**		**38**		

Source: Endline Data 2018 Student Assessments

Note: Columns A and B present the regression-adjusted difference on the percentage of students in single- or double-shift schools who obtained a score equal to zero in the intervention and comparison groups. The last column shows the p-value for the number of shifts by treatment interaction. Regressions adjust for baseline reading and math scores as well as children's gender, whether the child had a meal before getting to school, liked to read, attended preschool, spoke Arabic as primary language, had books to read other than textbooks, was read to at home, and did math problems at home. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups.

*Difference in group means is statistically significant at the .05 level.

TABLE M12. IMPACT ON ZERO SCORES IN MATH FOR GRADE 2 STUDENTS AT ENDLINE, BY NUMBER OF SCHOOL SHIFTS

Variable (Total # of Items)	Impact for Single Shift(A)	*P*-Value	Impact for Double Shift (B)	*P*-Value	Difference in RAMP Impacts by Number of Shifts (P-Value)
Number Identification	0.0	0.32	0.0	0.59	0.92
Number Discrimination	-0.1	0.79	-0.1	0.57	0.95
Missing Numbers	0.6	0.51	0.4	0.59	0.87
Addition Facts - L1	2.1	0.30	-0.7	0.83	0.50
Addition Facts - L2	4.0	0.15	-1.9	0.60	0.19
Subtraction Facts - L1	2.7	0.31	-12.5*	0.028	0.016*
Subtraction Facts - L2	4.3	0.18	-23.0*	0.004	0.002*
Number of Students	**1636**		**295**		
Number of Schools	**198**		**39**		

Source: Endline Data 2018 Student Assessments

Note: Columns A and B present the regression-adjusted difference on the percentage of students in single- or double-shift schools who obtained a score equal to zero in the intervention and comparison groups. The last column shows the p-value for the number of shifts by treatment interaction. Regressions adjust for baseline reading and math scores as well as children's gender, whether the child had a meal before getting to school, liked to read, attended preschool, spoke Arabic as primary language, had books to read other than textbooks, was read to at home, and did math problems at home. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups.

*Difference in group means is statistically significant at the .05 level.

ANNEX N. SUBGROUP ANALYSES BY GENDER AND NUMBER OF SCHOOL SHIFTS

The evaluation team examined whether impacts of RAMP differed for boys and girls as well as for students in single- versus double-shift schools.

READING BY GENDER

The results revealed no statistically significant differences in RAMP impacts on reading between boys and girls in either grade (see the last column in Tables N1 and N2).

TABLE N.1. IMPACT ON READING PERFORMANCE SCORES FOR GRADE 1 STUDENTS AT ENDLINE, BY GENDER

Variable (Total # of Items)	Impact for Girls (A)	P-Value	Impact for Boys (B)	P-Value	P-Value for the Difference in RAMP Impacts by Gender
Phoneme Isolation (Out of 10)	-0.3	0.21	-0.2	0.43	0.82
Syllable Segmentation (Out of 10)	0.3	0.31	0.2	0.63	0.80
Letter Sound Knowledge (Out of 100, Prorated)	-4.8*	0.014	-2.6	0.16	0.40
Reading Vocabulary (Out of 10)	0.0	0.84	0.0	0.84	0.97
Passage Reading (Out of 41, Prorated Score)	0.1	0.95	-0.6	0.59	0.64
Reading Comprehension (Out of 6)	-0.2	0.27	-0.2	0.21	0.91
Number of Students	988		943		
Number of Schools	193		192		

Source: Endline Data 2018 Student Assessments

Note: Columns A and B present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well as students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline are equated into the midline scale. There are fewer schools than in analyses for the overall sample because single-gender schools only contribute to estimating the impacts on girls or boys.

*Difference in group means is statistically significant at the .05 level.

ANNEX O. DESCRIPTIVE STATISTICS AND HISTOGRAMS FOR READING AND MATH EQUATED SCORES, BY GRADE AND STUDY GROUP

Tables O.1 to O.6 show descriptive statistics for reading and math equated scores for students who were assessed in both baseline at endline.

DESCRIPTIVE STATISTICS

TABLE O.1. READING AND MATH DESCRIPTIVE STATISTICS FOR GRADE 1 STUDENTS, BOTH COHORTS

	Baseline					Midline					Endline				
Variable	Number of Students	Mean	Sd	Min	Max	Number of Students	Mean	Sd	Min	Max	Number of Students	Mean	Sd	Min	Max
Reading															
Orientation to Print	1952	2.3	1.4	0	5	1864	2.9	1.4	0	5	NA	NA	NA	NA	NA
Phoneme Isolation	1952	3.7	2.1	0	10	1864	4.1	2.6	0	10	1952	4.8	2.8	0	10
Letter Sound	1952	21	14	0	86	1864	30.8	18.9	0	88	1952	33	20.5	0	90
Syllable Segmentation	1952	4	4.2	0	10	1864	5.3	4.2	0	10	1952	6.5	3.9	0	10
Invented Word Decoding	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Reading Vocabulary	1952	3.9	3.1	0	10	1864	6.6	2.8	0	10	1952	8.1	2.5	0	10
Passage Reading	NA	NA	NA	NA	NA	1864	6	8.3	0	53.2	1952	14.9	12.8	0	68.3
Reading Comprehension	NA	NA	NA	NA	NA	1864	0.8	1.1	0	6	1952	1.5	1.7	0	6
Math															
Counting Numbers	1952	27.2	11.3	0	40	1864	32.3	11	1	40	1952	36.3	8.2	0	40
Counting Objects	1952	8.7	2.1	0	10	1864	9.7	1	1	10	1952	9.9	0.7	1	10
Number Identification	1952	26.5	13.7	0	92.3	1864	23.1	7.6	0	63.2	1952	32.4	12.4	0	85.7
Number Discrimination	1952	7.6	2.8	0	10	1864	8.4	2.8	0	10	1952	9.1	2.1	0	10
Missing Number	NA	NA	NA	NA	NA	1864	4.4	2.9	0	10	1952	6.8	2.8	0	10
Addition Facts-L1	NA	NA	NA	NA	NA	1864	4.8	6.2	0	24	1952	5.6	5.7	0	37.2

Source: RAMP Impact Study - Baseline, Midline, and Endline Data 2018 Student Assessments

Note: The table shows unadjusted, unweighted descriptive statistics (mean, standard deviation, and minimum and maximum values observed) for student reading and math scores. Sd stands for standard deviation. Midline and endline scores are equated into the baseline scale. Endline subtasks that were not administered at baseline are equated into the midline scale, and midline subtasks that were not administered at baseline are not equated. The analytic sample at all time points consists of students who were assessed at both baseline and endline. NA indicates that the subtask was not administered to students in that grade and time point.

TABLE O.2. READING AND MATH DESCRIPTIVE STATISTICS FOR GRADE 1 STUDENTS IN COHORT 2

	Baseline					Midline					Endline				
Variable	Number of Students	Mean	Sd	Min	Max	Number of Students	Mean	Sd	Min	Max	Number of Students	Mean	Sd	Min	Max
Reading															
Orientation to Print	973	2.4	1.4	0	5	941	3.1	1.3	0	5	NA	NA	NA	NA	NA
Phoneme Isolation	973	3.9	2.1	0	10	941	4.4	2.6	0	10	973	5	2.8	0	10
Letter Sound	973	22.2	14.8	0	86	941	32.5	19.4	0	88	973	33.2	21.1	0	90
Syllable Segmentation	973	4.9	4.3	0	10	941	6.1	4	0	10	973	7.2	3.6	0	10
Invented Word Decoding	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Reading Vocabulary	973	4.4	3.1	0	10	941	7.1	2.6	0	10	973	8.4	2.2	0	10
Passage Reading	NA	NA	NA	NA	NA	941	7.3	9.1	0	53.2	973	16.6	13.2	0	68.3
Reading Comprehension	NA	NA	NA	NA	NA	941	1	1.1	0	6	973	1.7	1.7	0	6
Math															
Counting Numbers	973	28.7	10.9	0	40	941	33.3	10.5	2	40	973	36.9	7.4	4	40
Counting Objects	973	8.9	1.9	0	10	941	9.8	0.8	1	10	973	9.9	0.8	1	10
Number Identification	973	28.4	13.5	0	76	941	24	7.5	1.3	57.1	973	33.8	12.5	0	85.7
Number Discrimination	973	8	2.5	0	10	941	8.7	2.6	0	10	973	9.3	1.9	0	10
Missing Number	NA	NA	NA	NA	NA	941	4.9	2.9	0	10	973	7.2	2.7	0	10
Addition Facts-L1	NA	NA	NA	NA	NA	941	5.7	6.6	0	24	973	6	5.7	0	37.2

Source: RAMP Impact Study - Baseline, Midline, And Endline Data 2018 Student Assessments

Note: The table shows unadjusted, unweighted descriptive statistics (mean, standard deviation, and minimum and maximum values observed) for student reading and math scores. Sd stands for standard deviation. Midline and endline scores are equated into the baseline scale. Endline subtasks that were not administered at baseline are equated into the midline scale, and midline subtasks that were not administered at baseline are not equated. The analytic sample at all time points consists of students who were assessed at both baseline and endline. NA indicates that the subtask was not administered to students in that grade and time point.

TABLE O.3. READING AND MATH DESCRIPTIVE STATISTICS FOR GRADE 1 STUDENTS IN COHORT 3

	Baseline					Midline					Endline				
Variable	Number of Students	Mean	Sd	Min	Max	Number of Students	Mean	Sd	Min	Max	Number of Students	Mean	Sd	Min	Max
Reading															
Orientation to Print	979	2.1	1.5	0	5	923	2.8	1.4	0	5	NA	NA	NA	NA	NA
Phoneme Isolation	979	3.4	2.1	0	9	923	3.8	2.4	0	10	979	4.5	2.7	0	10
Letter Sound	979	19.8	13.1	0	62	923	29.1	18.2	0	80	979	32.7	20	0	90
Syllable Segmentation	979	3	3.9	0	10	923	4.5	4.1	0	10	979	5.9	4	0	10
Invented Word Decoding	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Reading Vocabulary	979	3.5	3	0	10	923	6.1	2.9	0	10	979	7.8	2.6	0	10
Passage Reading	NA	NA	NA	NA	NA	923	4.7	7.1	0	46.8	979	13.1	12.1	0	61.6
Reading Comprehension	NA	NA	NA	NA	NA	923	0.7	1	0	5	979	1.3	1.6	0	6
Math															
Counting Numbers	979	25.6	11.3	0	40	923	31.3	11.4	1	40	979	35.8	8.8	0	40
Counting Objects	979	8.4	2.3	0	10	923	9.6	1.1	1	10	979	9.9	0.6	1	10
Number Identification	979	24.7	13.8	0	92.3	923	22.2	7.6	0	63.2	979	31.1	12.1	0	75
Number Discrimination	979	7.1	2.9	0	10	923	8.1	3	0	10	979	8.9	2.3	0	10
Missing Number	NA	NA	NA	NA	NA	923	4	2.8	0	10	979	6.4	2.9	0	10
Addition Facts-L1	NA	NA	NA	NA	NA	923	3.9	5.7	0	20.7	979	5.2	5.7	0	36.4

Source: RAMP Impact Study - Baseline, Midline, and Endline Data 2018 Student Assessments

Note: The table shows unadjusted, unweighted descriptive statistics (mean, standard deviation, and minimum and maximum values observed) for student reading and math scores. Sd stands for standard deviation. Midline and endline scores are equated into the baseline scale. Endline subtasks that were not administered at baseline are equated into the midline scale, and midline subtasks that were not administered at baseline are not equated. The analytic sample at all time points consists of students who were assessed at both baseline and endline. NA indicates that the subtask was not administered to students in that grade and time point.

TABLE O.4. READING AND MATH DESCRIPTIVE STATISTICS FOR GRADE 2 STUDENTS, BOTH COHORTS

	BASELINE					MIDLINE					ENDLINE				
Variable	Number of Students	Mean	Sd	Min	Max	Number of Students	Mean	Sd	Min	Max	Number of Students	Mean	Sd	Min	Max
Reading															
Orientation to Print	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Phoneme Isolation	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Letter Sound	1944	28	18.7	0	86	1872	33.6	21.6	0	105.3	1944	35.3	23.2	0	101.7
Syllable Segmentation	1944	4.4	3.7	0	10	1872	5.4	3.6	0	10	1944	5.6	3.5	0	10
Invented Word Decoding	1944	7	6.8	0	38	1872	8.2	7.5	0	50	1944	12.2	11.2	0	50
Reading Vocabulary	1944	6.4	3.2	0	10	1872	7.9	2.7	0	10	1944	9.1	2.1	0	10
Passage Reading	1944	10.8	12.5	0	100	1872	16.7	11.8	0	84.8	1944	22.6	14.3	0	72.7
Reading Comprehension	1944	1.3	1.3	0	6	1872	1.8	1.8	0	6	1944	2.7	1.8	0	6
Math															
Counting Numbers	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Counting Objects	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Number Identification	1944	22.4	12.1	0	92.3	1872	29.2	14.1	0	100	1944	35.2	15.2	0	120
Number Discrimination	1944	7.3	2.4	0	10	1872	7.6	2.5	0	10	1944	8.6	2	0	10
Missing Number	1944	5.6	2.9	0	10	1872	6.9	3	0	10	1944	8.3	2.4	0	10
Addition Facts-L1	1944	8.2	5.5	0	29.3	1872	10.2	5.6	0	42.9	1944	9.3	6.1	0	34.3

Source: RAMP Impact Study - Baseline, Midline, and Endline Data 2018 Student Assessments

Note: The table shows unadjusted, unweighted descriptive statistics (mean, standard deviation, and minimum and maximum values observed) for student reading and math scores. Sd stands for standard deviation. Midline and endline scores are equated into the baseline scale. Endline subtasks that were not administered at baseline are equated into the midline scale, and midline subtasks that were not administered at baseline are not equated. The analytic sample at all time points consists of students who were assessed at both baseline and endline. NA indicates that the subtask was not administered to students in that grade and time point.

TABLE O.5. READING AND MATH DESCRIPTIVE STATISTICS FOR GRADE 2 STUDENTS IN COHORT 2

	Baseline					Midline					Endline				
Variable	Number of Students	Mean	Sd	Min	Max	Number of Students	Mean	Sd	Min	Max	Number of Students	Mean	Sd	Min	Max
Reading															
Orientation to Print	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Phoneme Isolation	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Letter Sound	966	29.3	19.3	0	84	936	34.2	21.7	0	105.3	966	35.1	24.1	0	101.7
Syllable Segmentation	966	5.3	3.6	0	10	936	6.4	3.4	0	10	966	6.3	3.4	0	10
Invented Word Decoding	966	7.9	7.3	0	38	936	9.2	7.8	0	50	966	13.3	11.7	0	50
Reading Vocabulary	966	6.9	2.9	0	10	936	8.1	2.5	0	10	966	9.3	1.8	0	10
Passage Reading	966	12.2	13.1	0	100	936	18.1	11.8	0	82	966	23.8	14.1	0	68.3
Reading Comprehension	966	1.4	1.3	0	6	936	2	1.8	0	6	966	2.8	1.8	0	6
Math															
Counting Numbers	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Counting Objects	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Number Identification	966	24.1	12	1	60	936	30.7	13.7	0	80	966	37	14.5	3	80
Number Discrimination	966	7.7	2.1	0	10	936	7.8	2.4	0	10	966	8.8	1.8	0	10
Missing Number	966	5.9	2.8	0	10	936	7.3	2.8	0	10	966	8.6	2.2	0	10
Addition Facts-L1	966	9.1	5.5	0	29.3	936	10.7	5.5	0	42.9	966	9.7	6.2	0	33.3

Source: RAMP Impact Study - Baseline, Midline, and Endline Data 2018 Student Assessments

Note: The table shows unadjusted, unweighted descriptive statistics (mean, standard deviation, and minimum and maximum values observed) for student reading and math scores. Sd stands for standard deviation. Midline and endline scores are equated into the baseline scale. Endline subtasks that were not administered at baseline are equated into the midline scale, and midline subtasks that were not administered at baseline are not equated. The analytic sample at all time points consists of students who were assessed at both baseline and endline. NA indicates that the subtask was not administered to students in that grade and time point.

TABLE O.6. READING AND MATH DESCRIPTIVE STATISTICS FOR GRADE 2 STUDENTS IN COHORT 3

	Baseline					Midline					Endline				
Variable	Number of Students	Mean	Sd	Min	Max	Number of Students	Mean	Sd	Min	Max	Number of Students	Mean	Sd	Min	Max
Reading															
Orientation to Print	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Phoneme Isolation	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Letter Sound	978	26.7	18	0	86	936	33	21.4	0	99	978	35.5	22.2	0	100.3
Syllable Segmentation	978	3.4	3.6	0	10	936	4.5	3.5	0	10	978	4.9	3.5	0	10
Invented Word Decoding	978	6.1	6.3	0	34	936	7.1	7	0	35	978	11	10.6	0	44
Reading Vocabulary	978	6	3.4	0	10	936	7.7	2.9	0	10	978	8.9	2.4	0	10
Passage Reading	978	9.4	11.7	0	92.3	936	15.4	11.5	0	84.8	978	21.4	14.3	0	72.7
Reading Comprehension	978	1.3	1.3	0	6	936	1.6	1.7	0	6	978	2.6	1.8	0	6
Math															
Counting Numbers	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Counting Objects	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Number Identification	978	20.8	12.1	0	92.3	936	27.7	14.3	0	100	978	33.5	15.6	0	120
Number Discrimination	978	6.9	2.7	0	10	936	7.4	2.6	0	10	978	8.3	2.2	0	10
Missing Number	978	5.3	2.9	0	10	936	6.6	3.1	0	10	978	8.1	2.6	0	10
Addition Facts-L1	978	7.3	5.4	0	27.3	936	9.6	5.7	0	26.1	978	8.8	6	0	34.3

Source: RAMP Impact Study - Baseline, Midline, and Endline Data 2018 Student Assessments

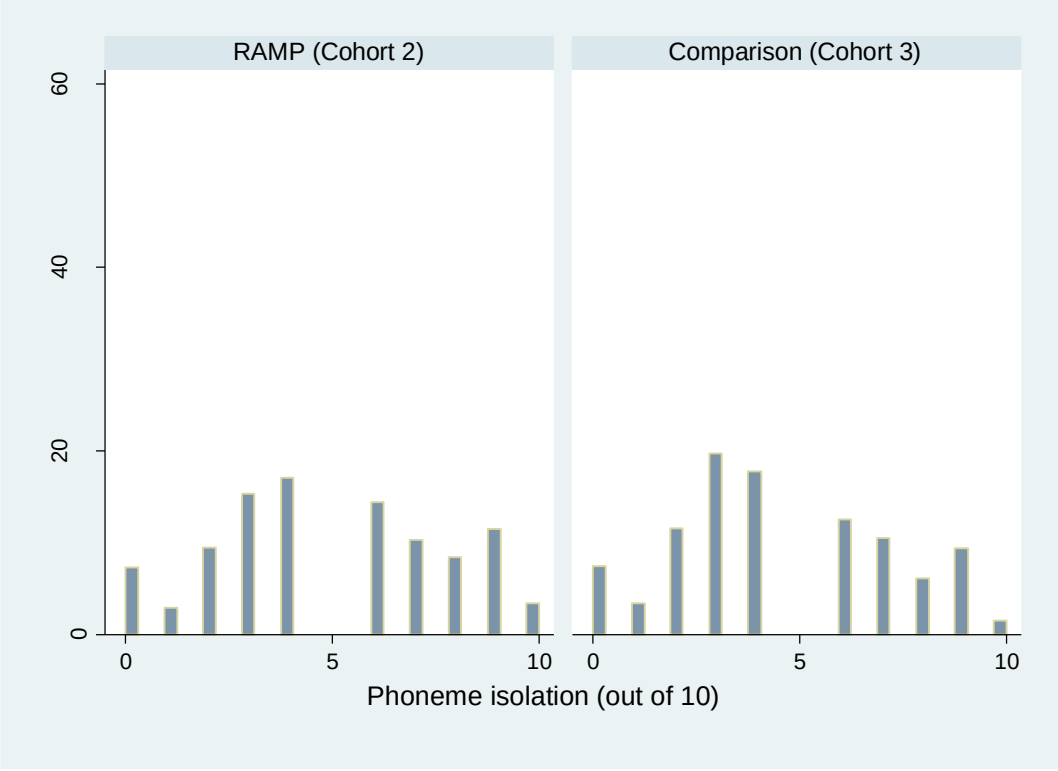
Note: The table shows unadjusted, unweighted descriptive statistics (mean, standard deviation, and minimum and maximum values observed) for student reading and math scores. Sd stands for standard deviation. Midline and endline scores are equated into the baseline scale. Endline subtasks that were not administered at baseline are equated into the midline scale, and midline subtasks that were not administered at baseline are not equated. The analytic sample at all time points consists of students who were assessed at both baseline and endline. NA indicates that the subtask was not administered to students in that grade and time point.

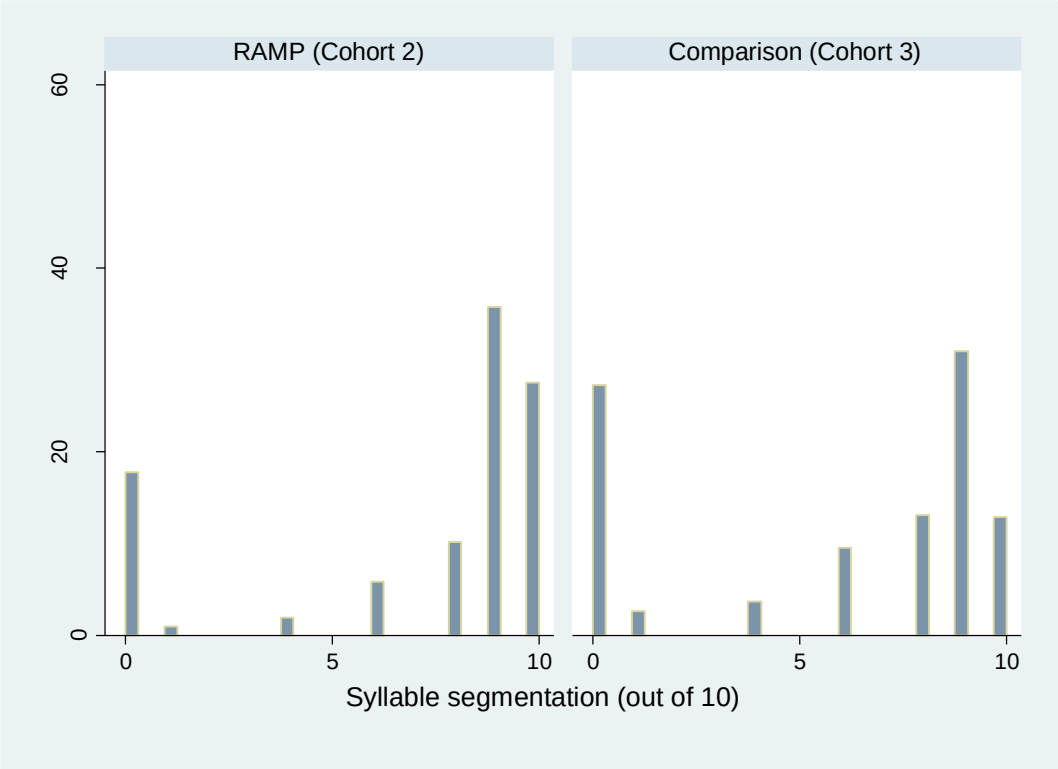
HISTOGRAMS FOR READING AND MATH SCORES AT ENDLINE

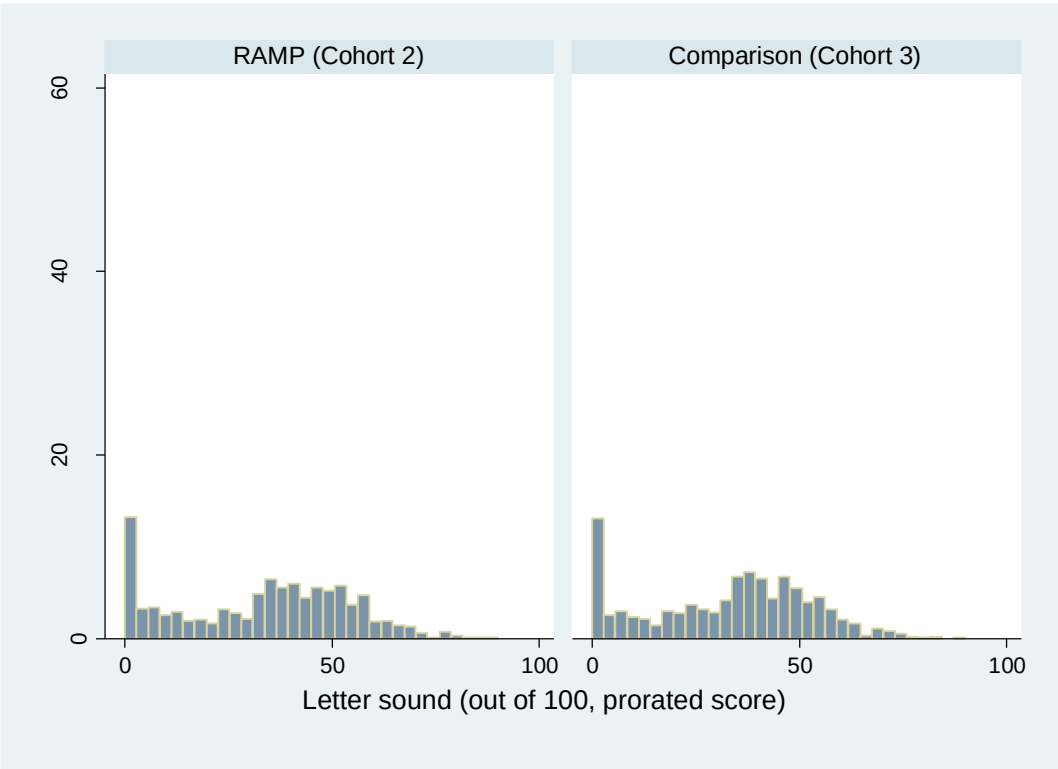
The histograms in this section provide a visual illustration of the distribution of endline scores by grade and study group for each subtask. Reading by grade histograms are presented first, followed by math histograms (Panels O.1-O.4). All scores are equated into the baseline scale, except for G1 passage reading, reading comprehension, missing numbers, and addition level 1, which were not administered to G1 students at baseline and were equated into the midline scale instead.

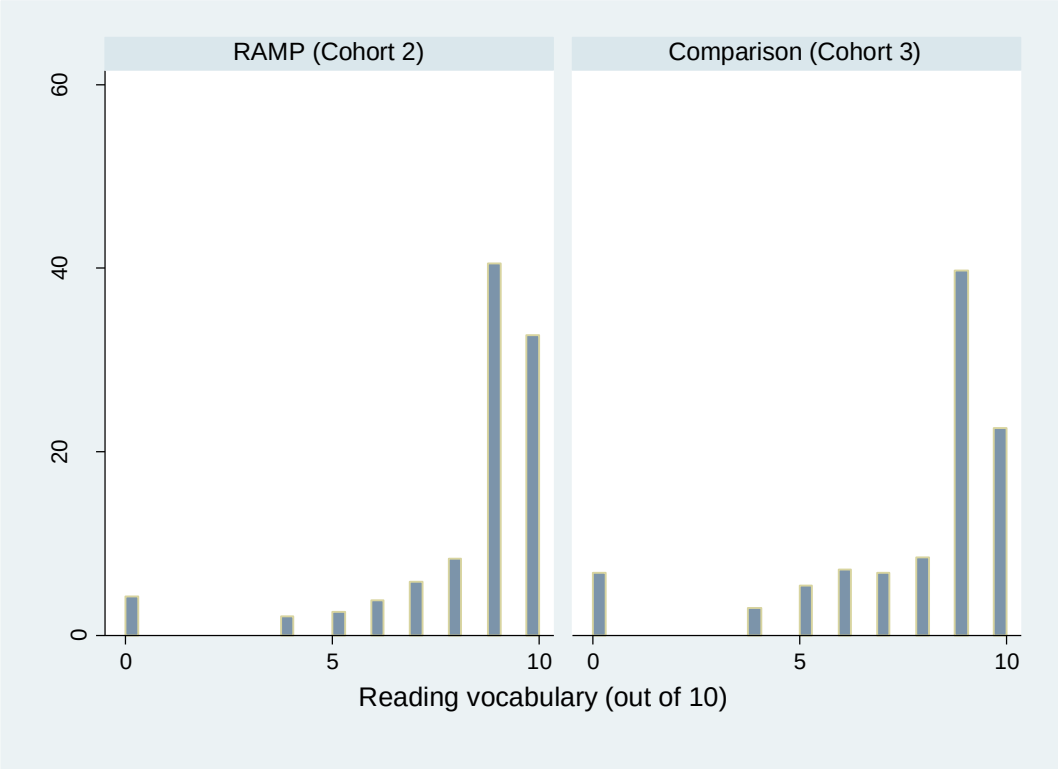
In some subtasks most students obtained the maximum possible score (for example, counting numbers and counting objects for students who were in G1 at baseline), while in others a large proportion of students obtained a score equal to zero (for example, non-word decoding and level 1 subtraction for students who were in G2 at baseline). These ceiling and floor effects can make it difficult to detect the impact of RAMP on those outcomes. There were, however, other subtasks where the scores approximated a normal distribution (for example, number identification). In addition, regressions were fitted to estimate RAMP impacts on the proportion of zero scores, allowing the evaluation to obtain useful information in the context of floor effects.

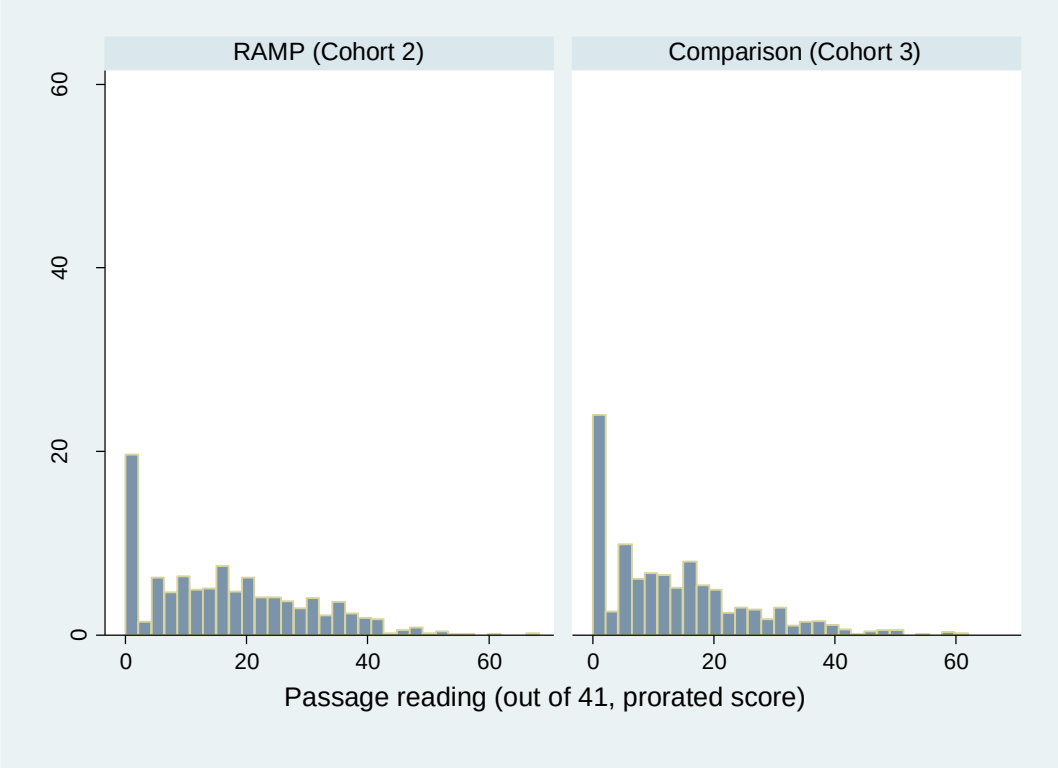
PANEL O.1. G1 EGRA HISTOGRAMS BY STUDY GROUP

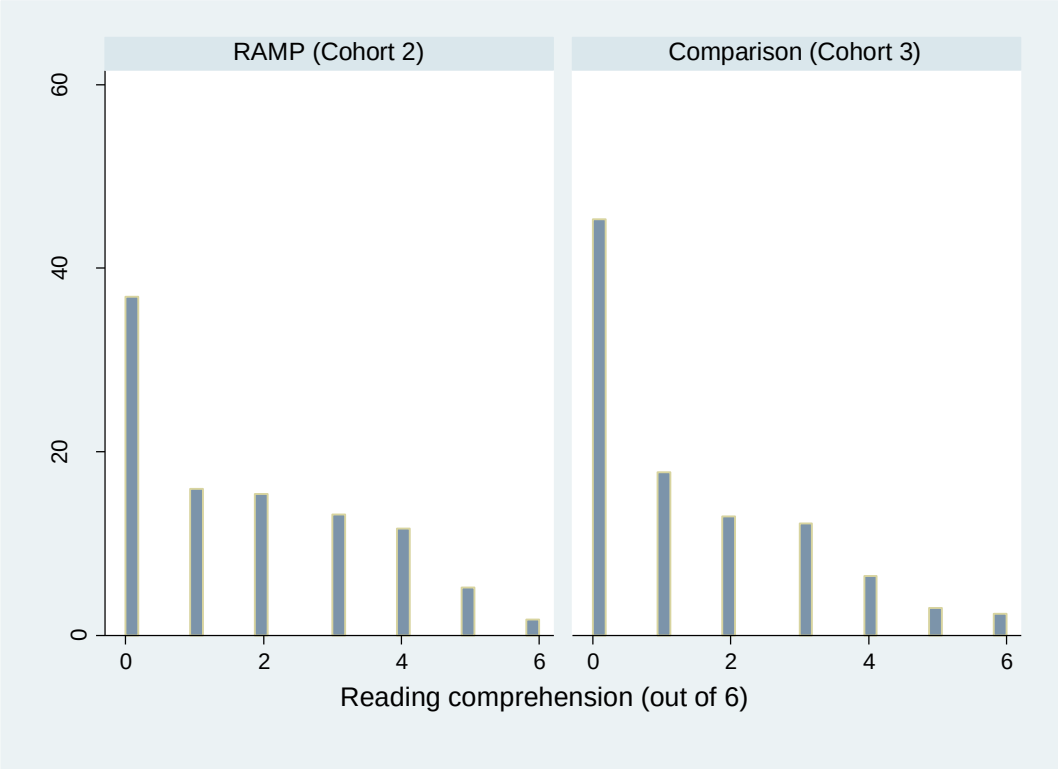



RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
5
10
Syllable segmentation (out of 10)

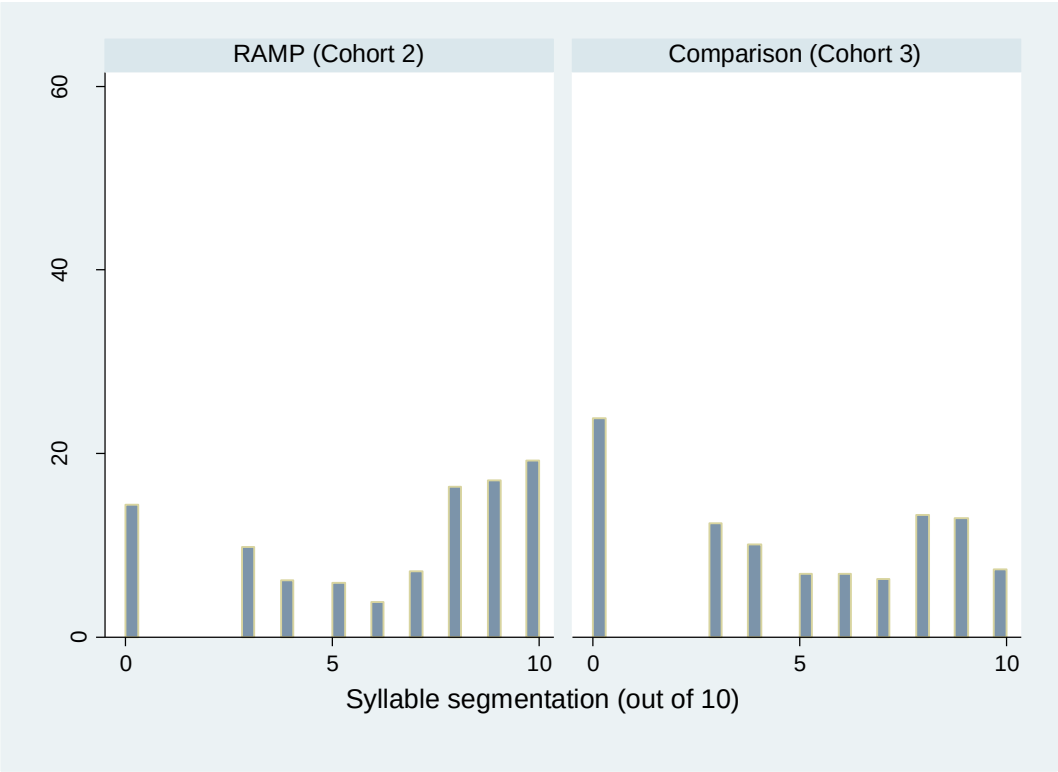

RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
50
100
Letter sound (out of 100, prorated score)

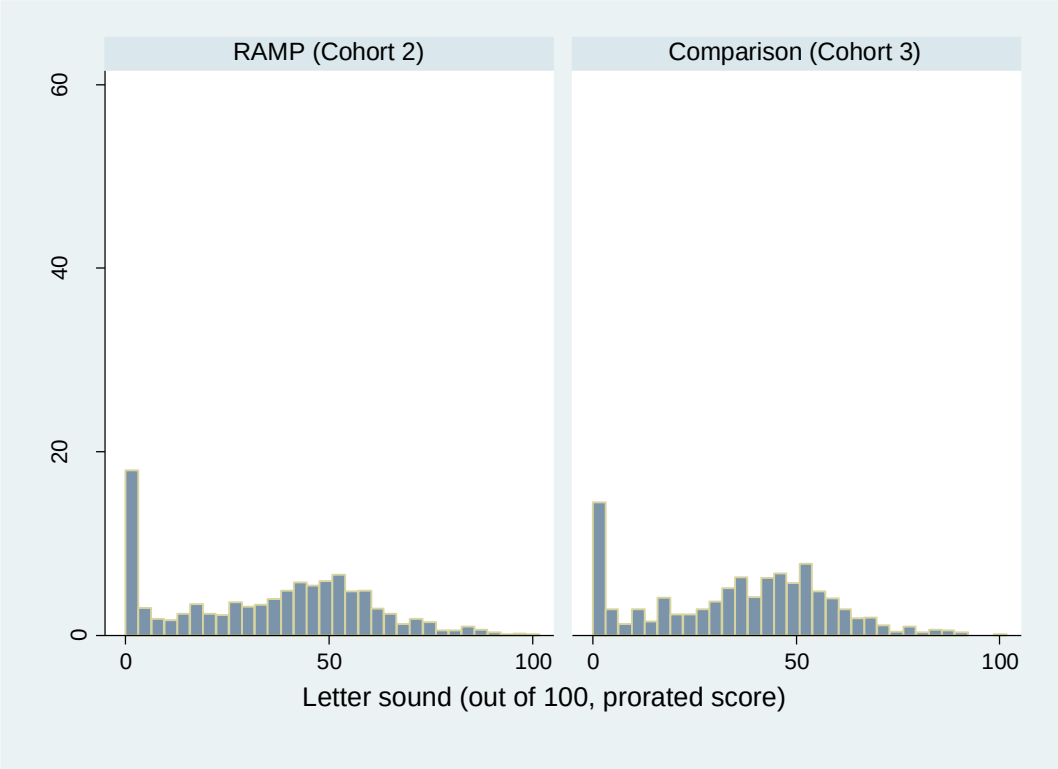

RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
5
10
Reading vocabulary (out of 10)

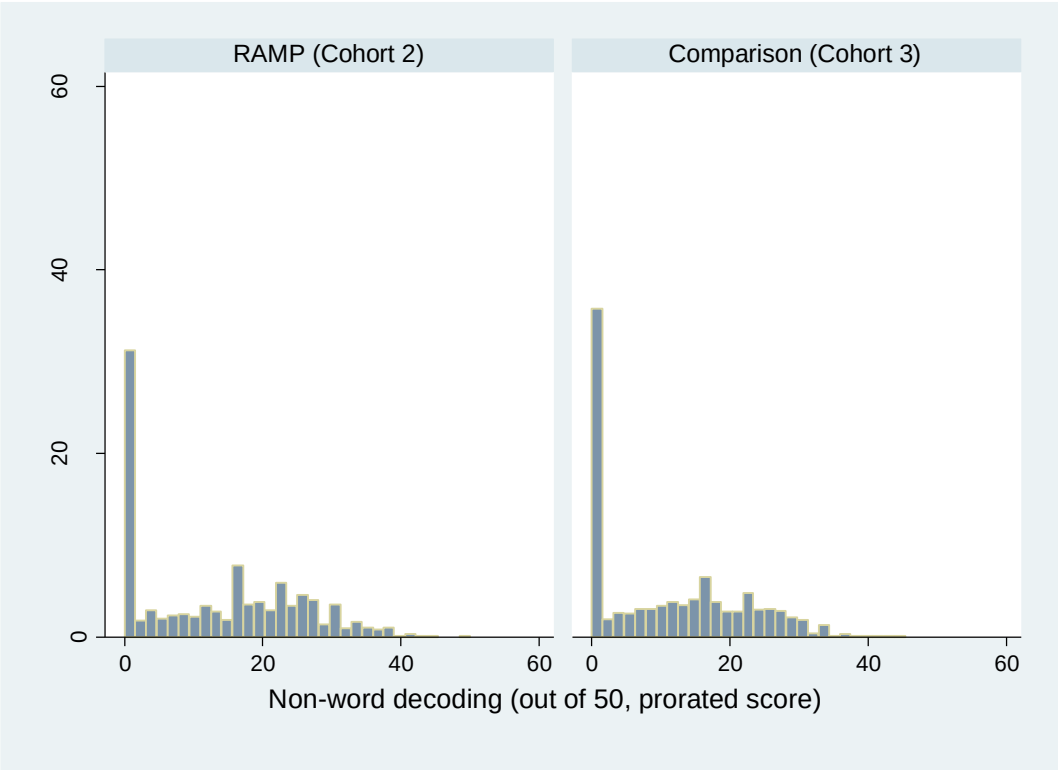

RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
20
40
60
Passage reading (out of 41, prorated score)

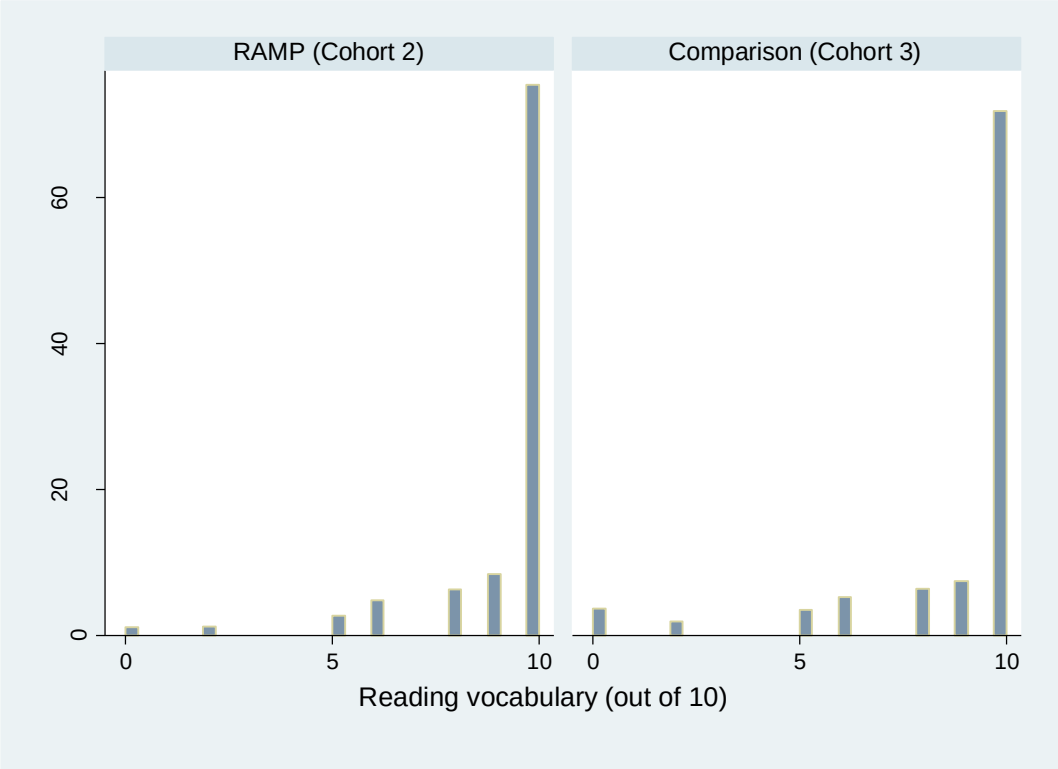


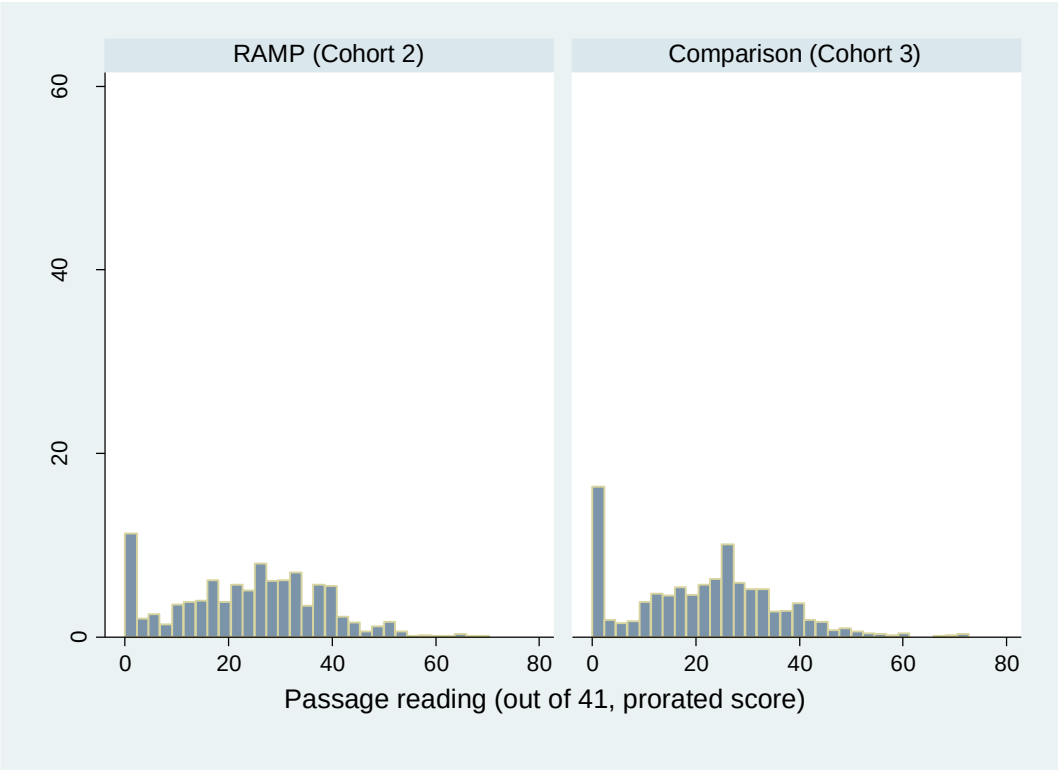
PANEL O.2: G2 EGRA HISTOGRAMS FOR READING SCORES BY STUDY GROUP

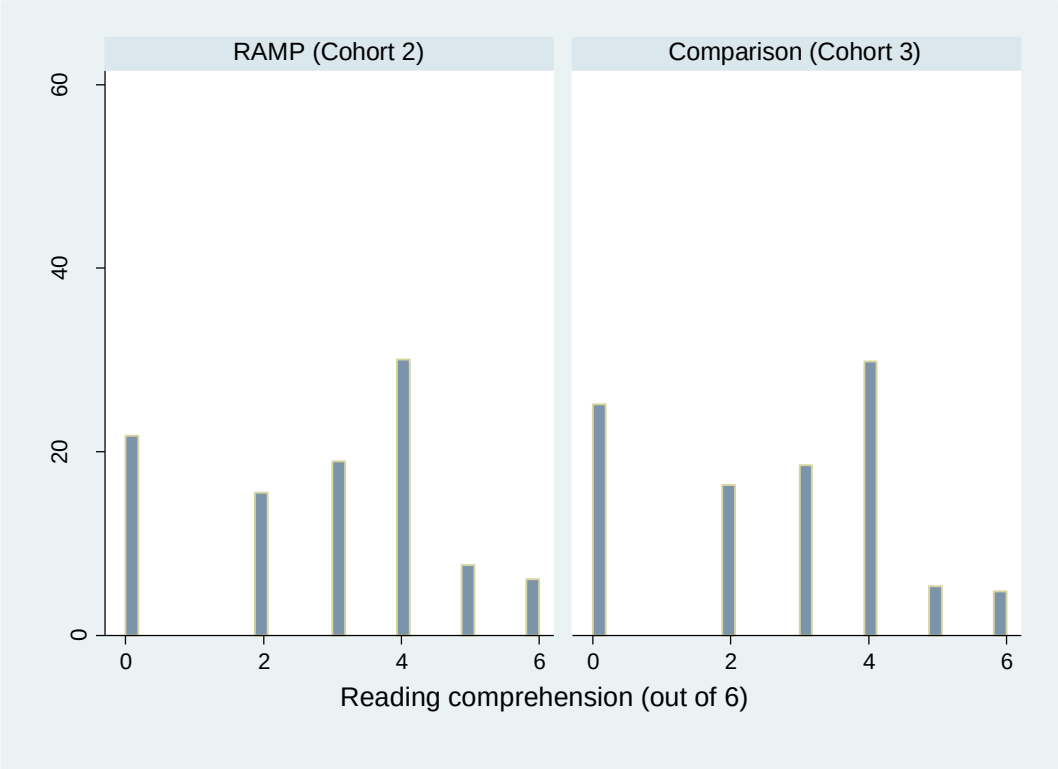



RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
50
100
Letter sound (out of 100, prorated score)

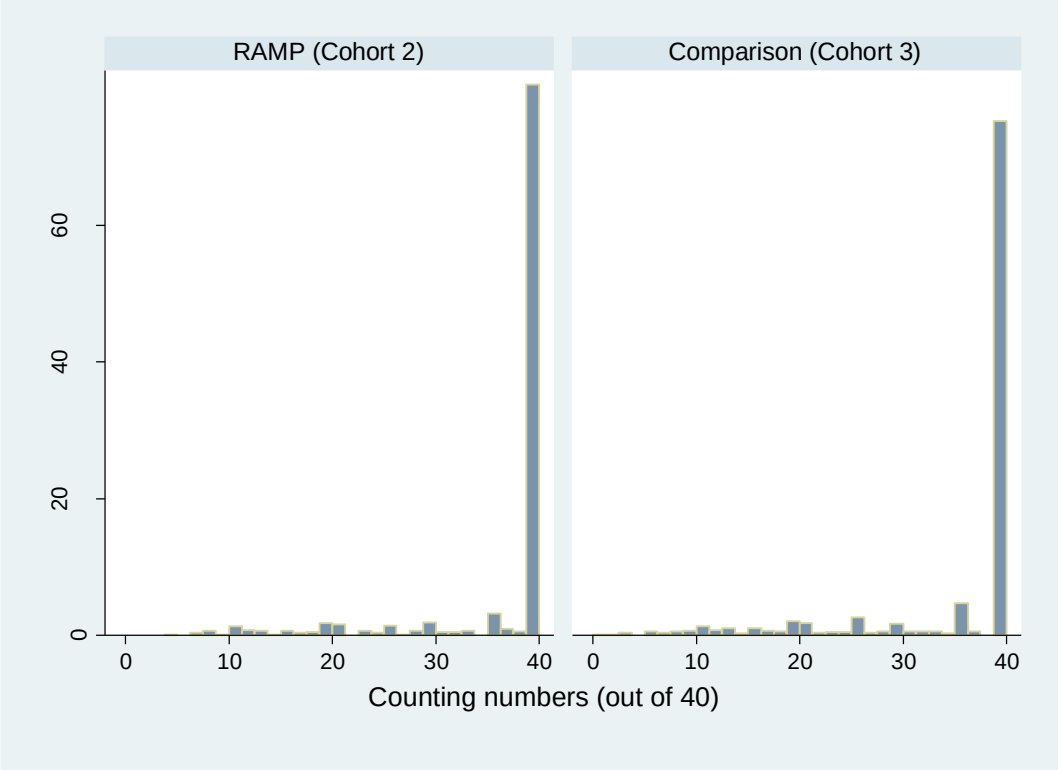

RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
20
40
60
Non-word decoding (out of 50, prorated score)

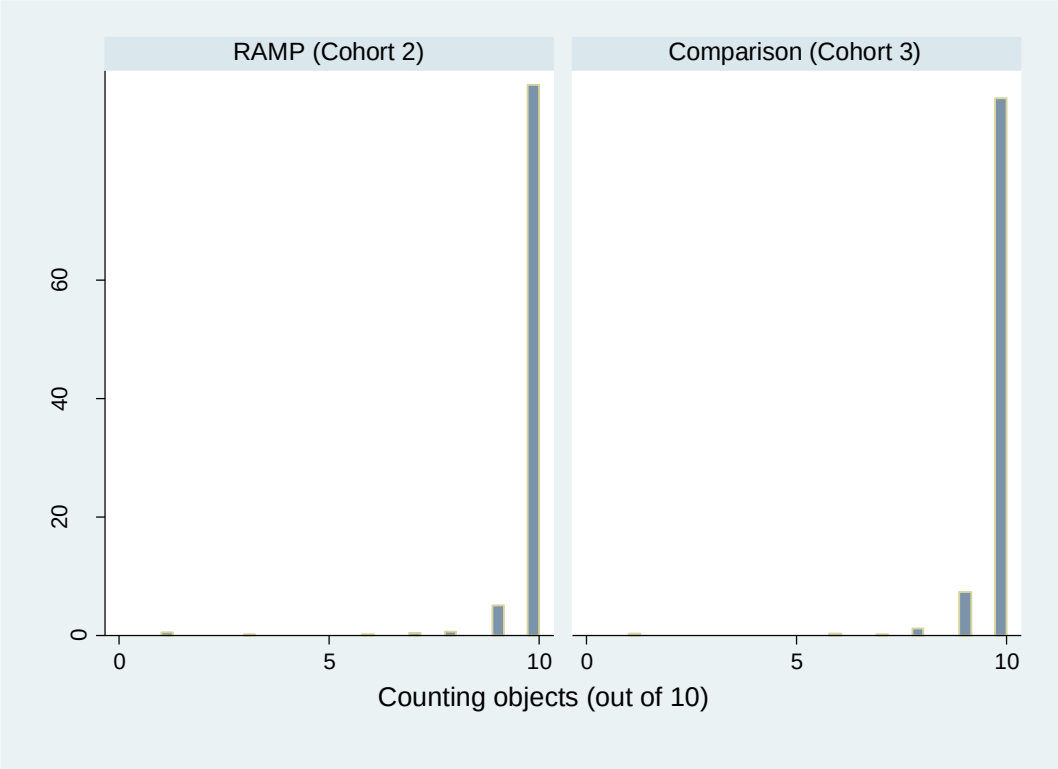

RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
5
10
Reading vocabulary (out of 10)

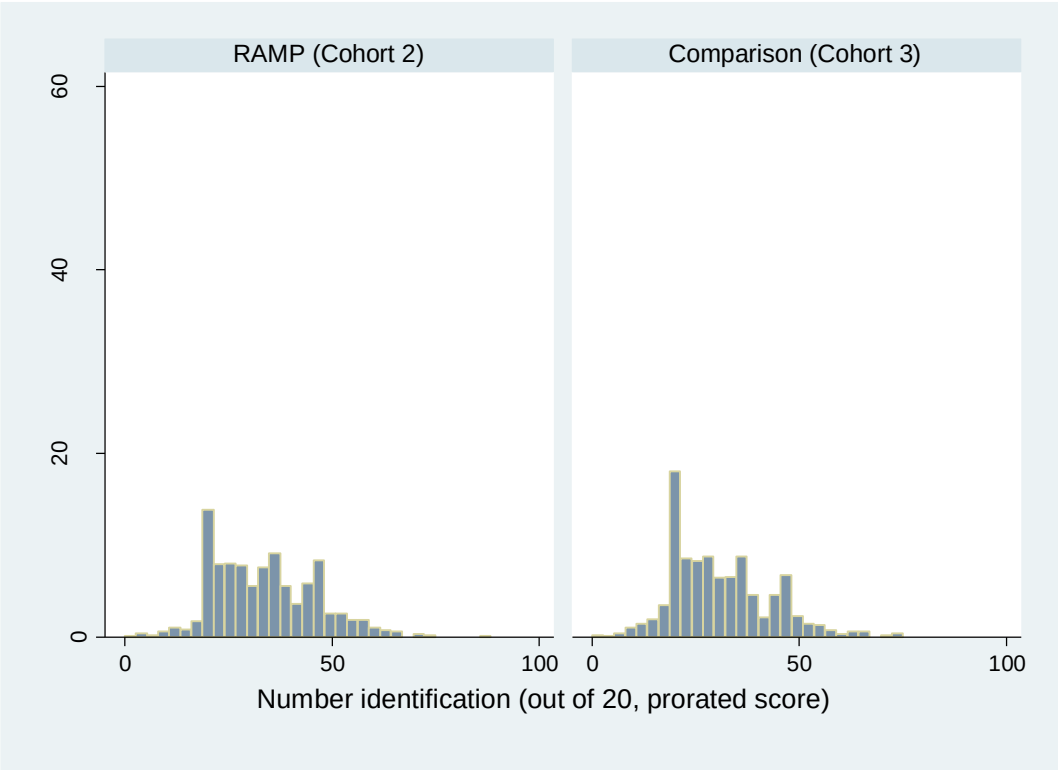

RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
20
40
60
80
Passage reading (out of 41, prorated score)

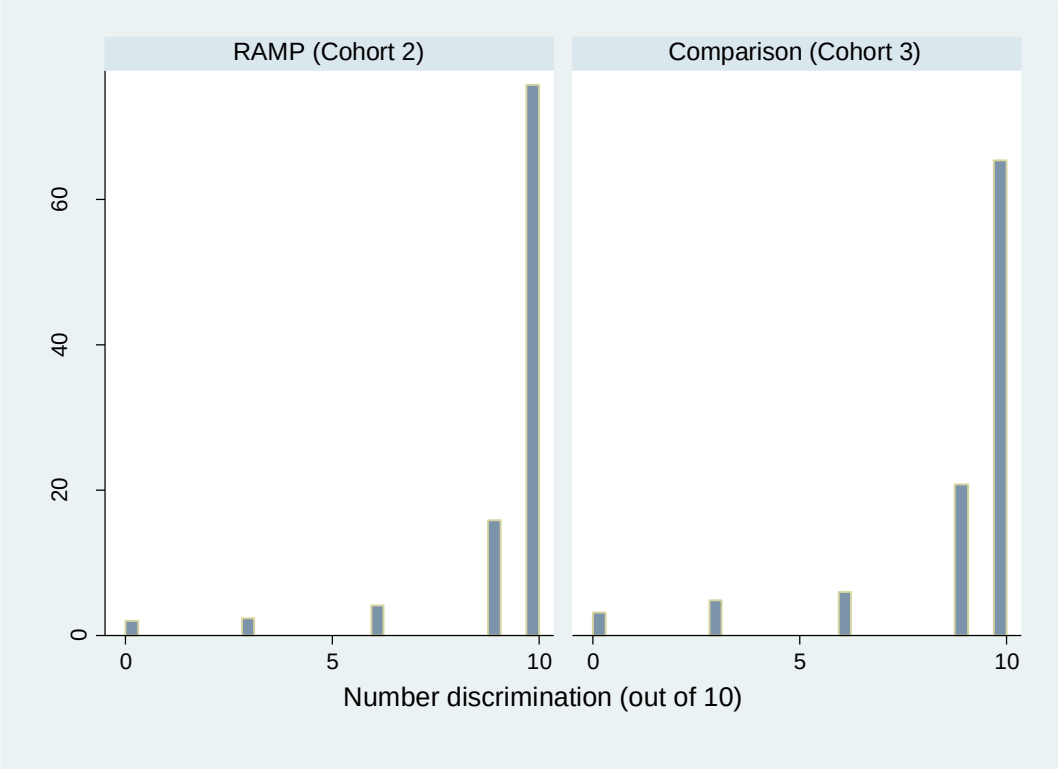


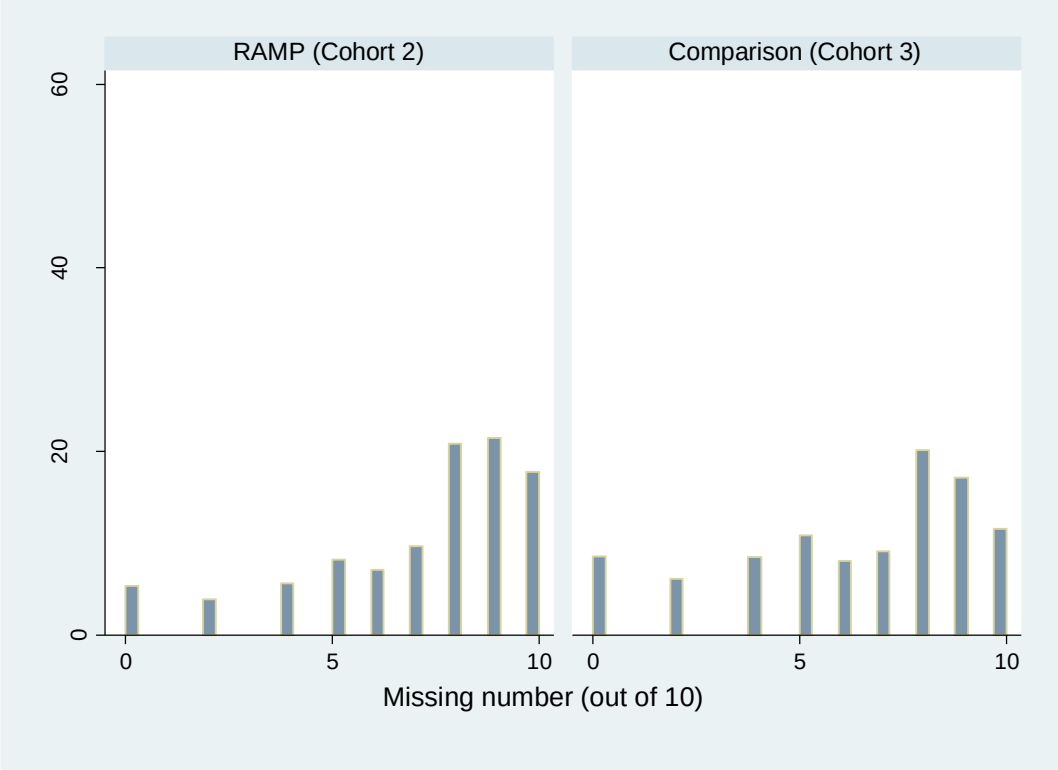
PANEL O.3. G1 EGMA HISTOGRAMS FOR MATH SCORES BY STUDY GROUP

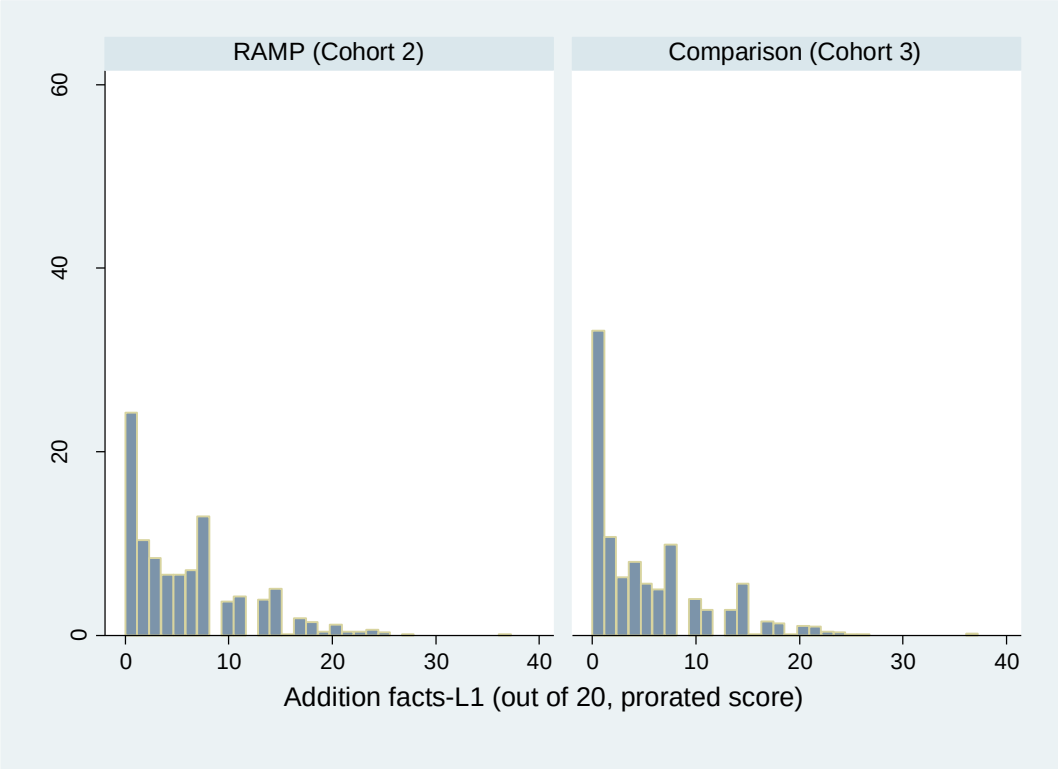



RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
5
10
Counting objects (out of 10)

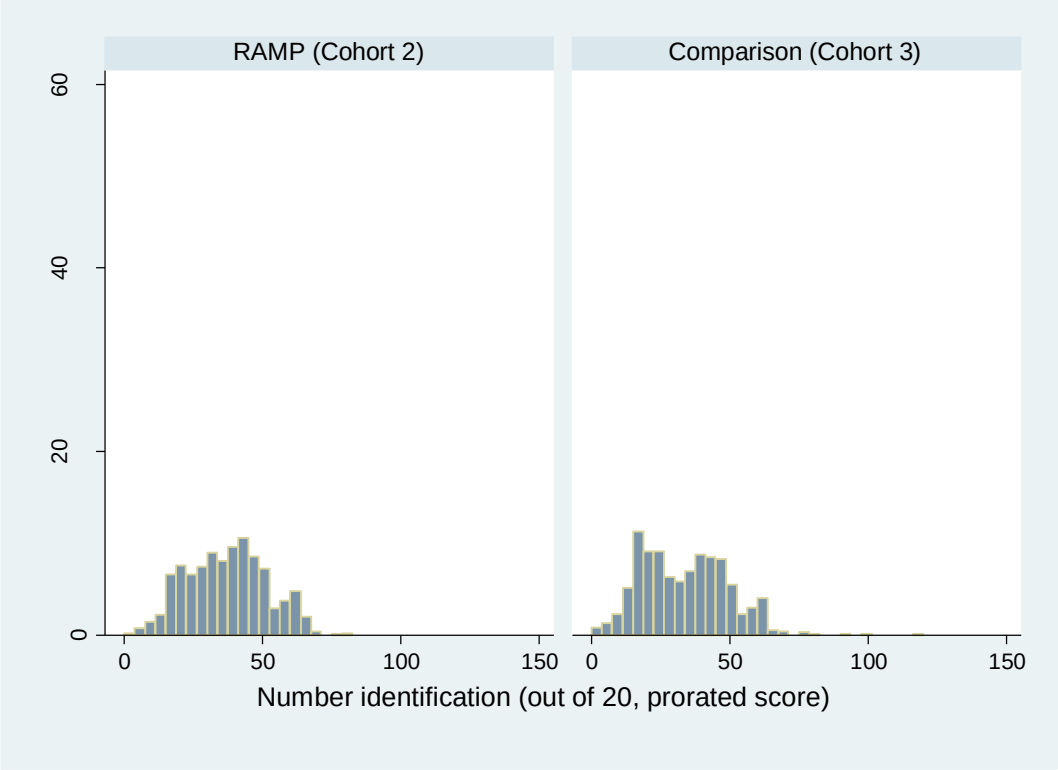

RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
50
100
Number identification (out of 20, prorated score)

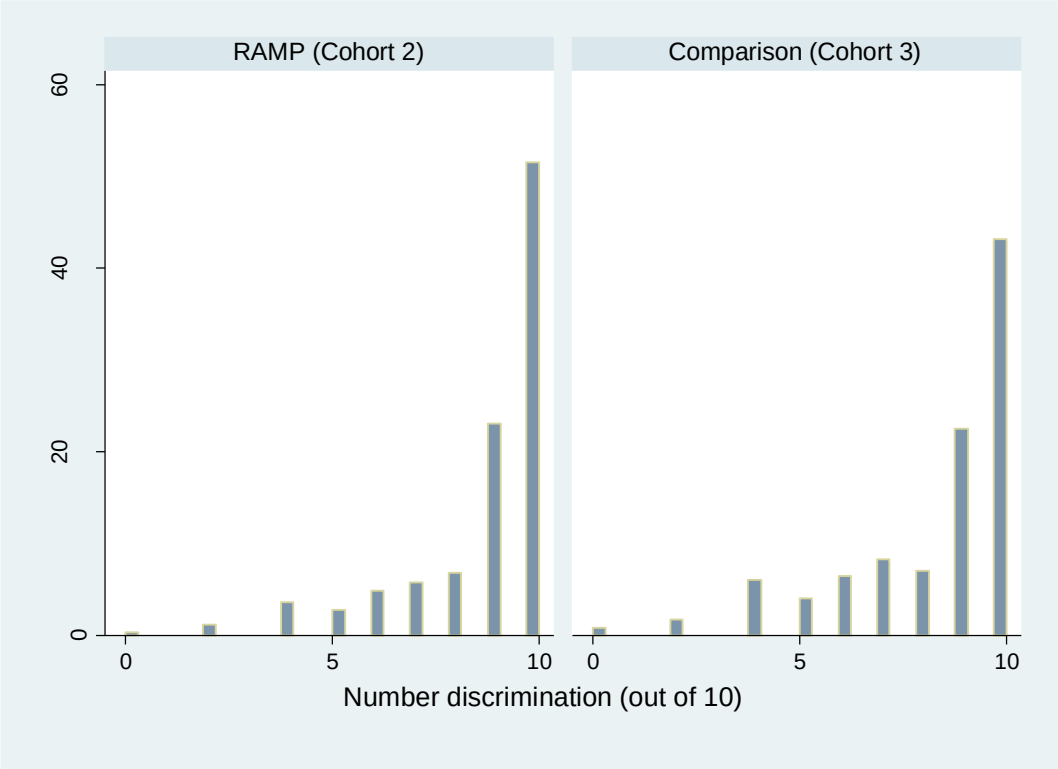

RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
5
10
Number discrimination (out of 10)

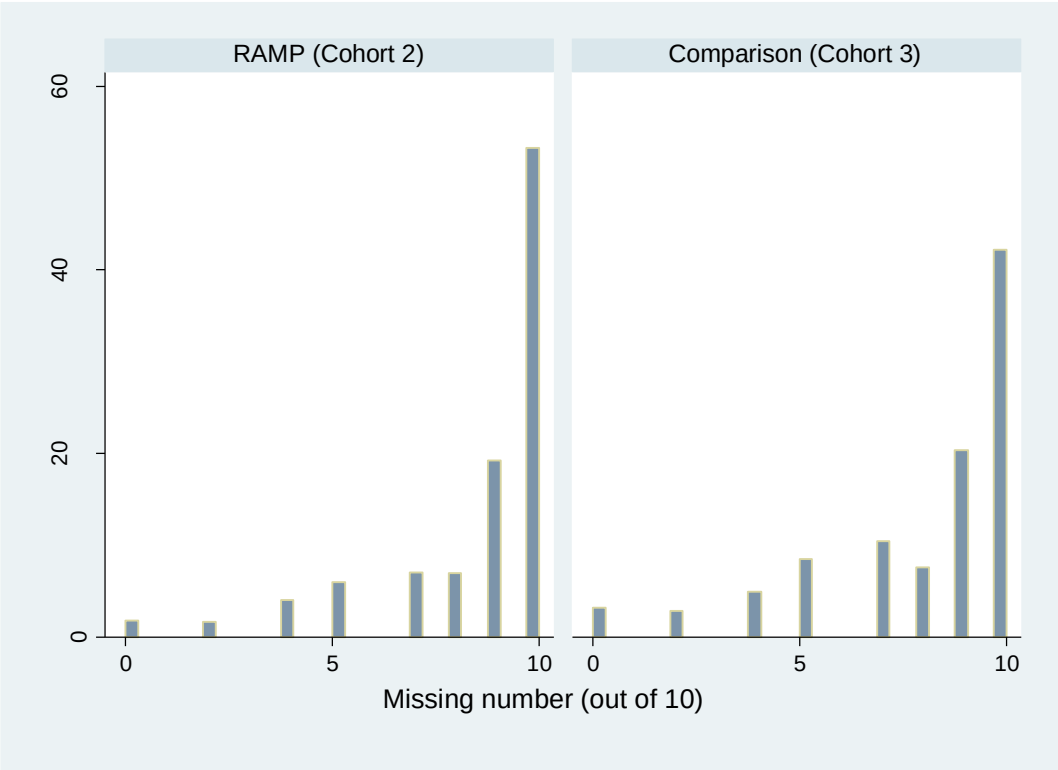

RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
5
10
Missing number (out of 10)

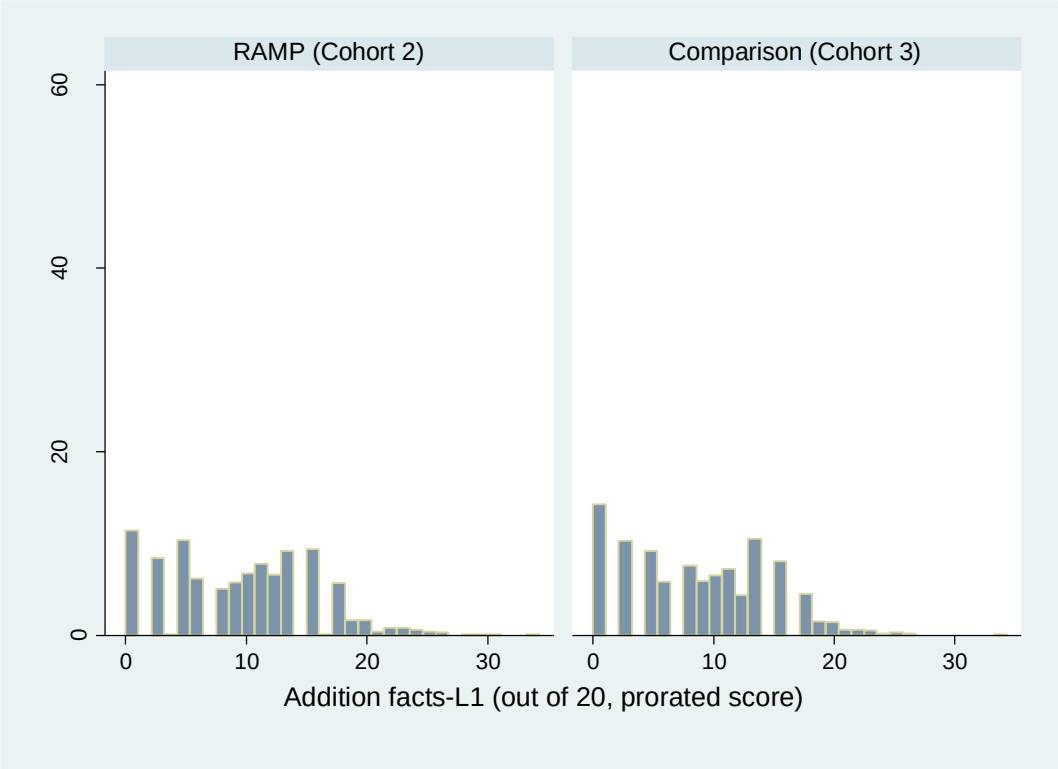


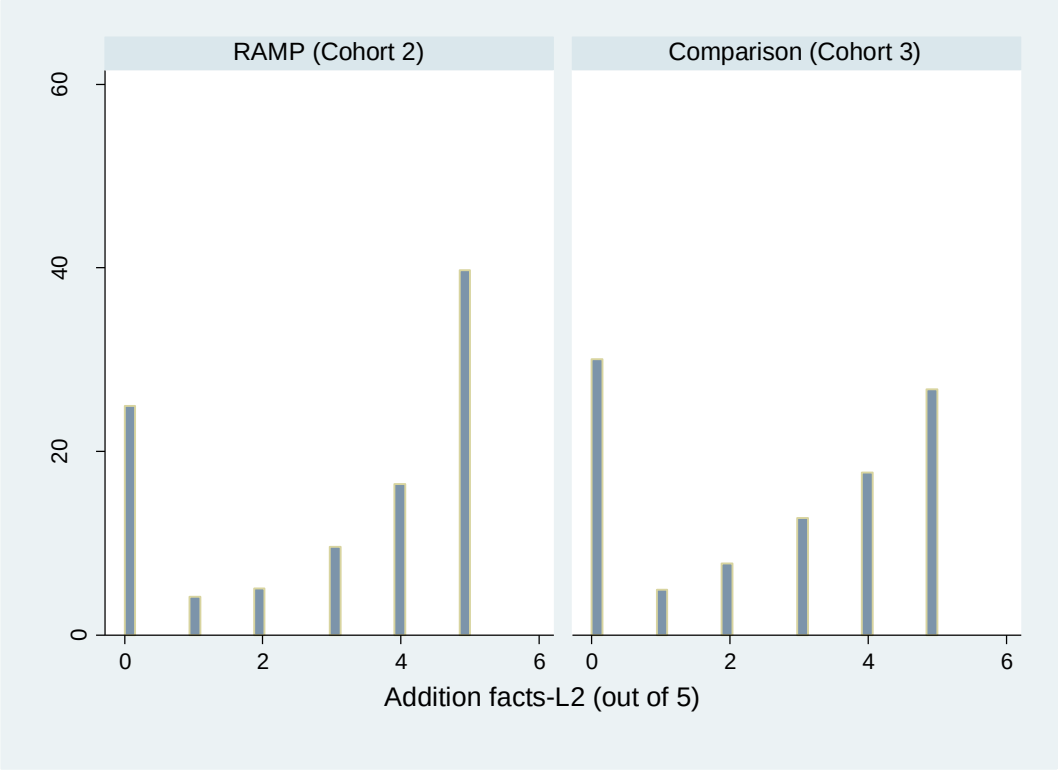
PANEL O.4. G2 EGMA HISTOGRAMS BY STUDY GROUP

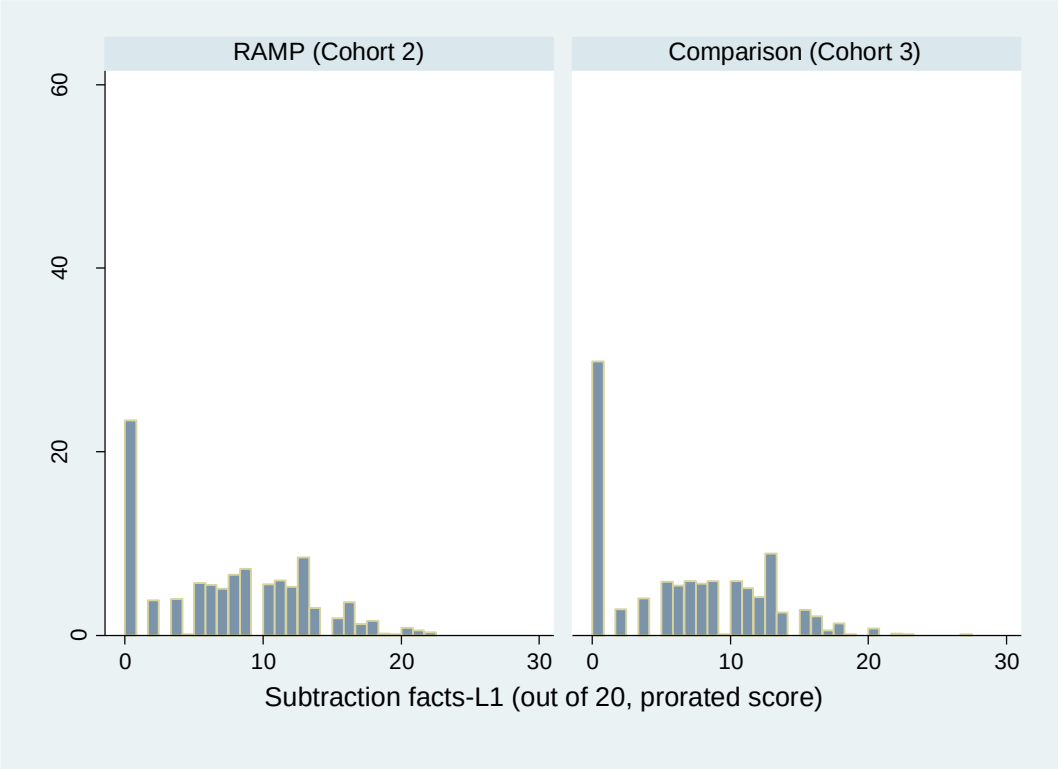


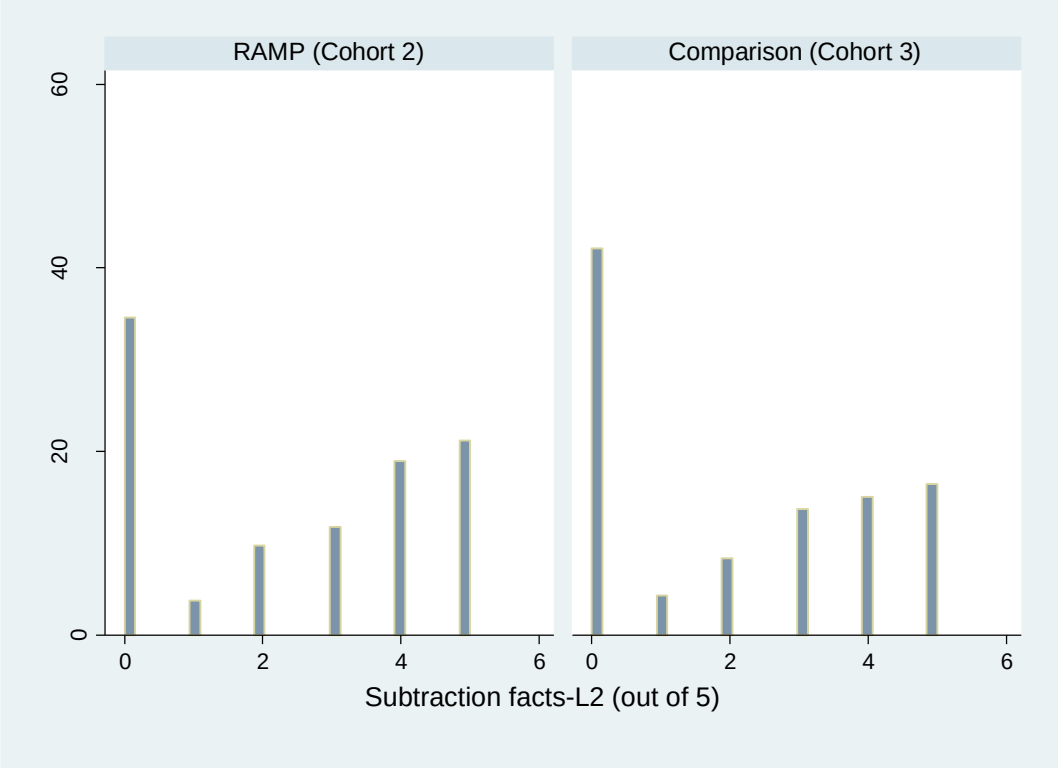

RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
5
10
Number discrimination (out of 10)


RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
5
10
Missing number (out of 10)


RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
10
20
30
Addition facts-L1 (out of 20, prorated score)


RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
2
4
6
Addition facts-L2 (out of 5)

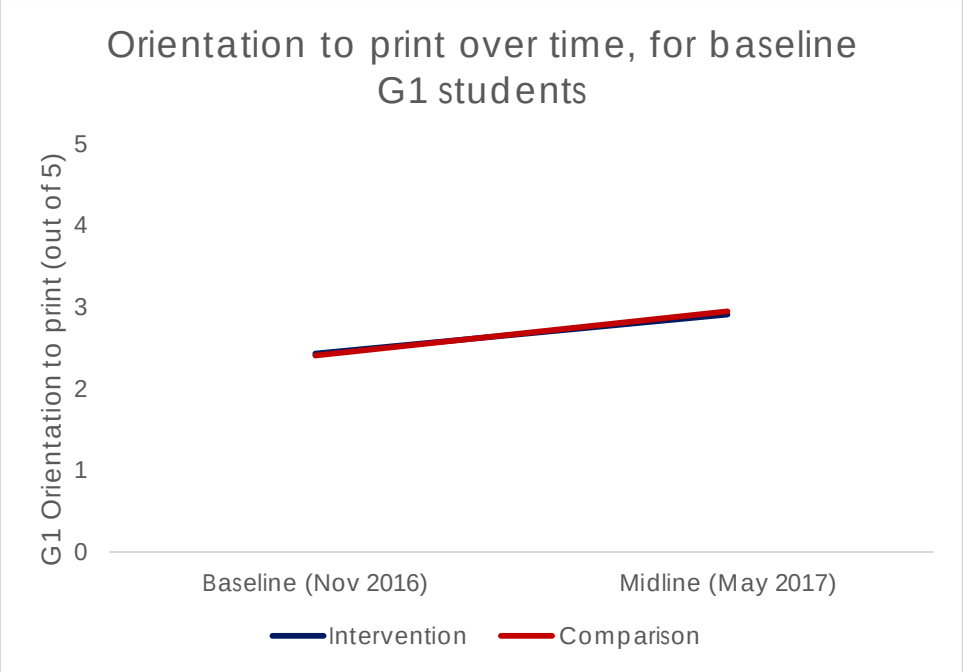

RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
10
20
30
Subtraction facts-L1 (out of 20, prorated score)

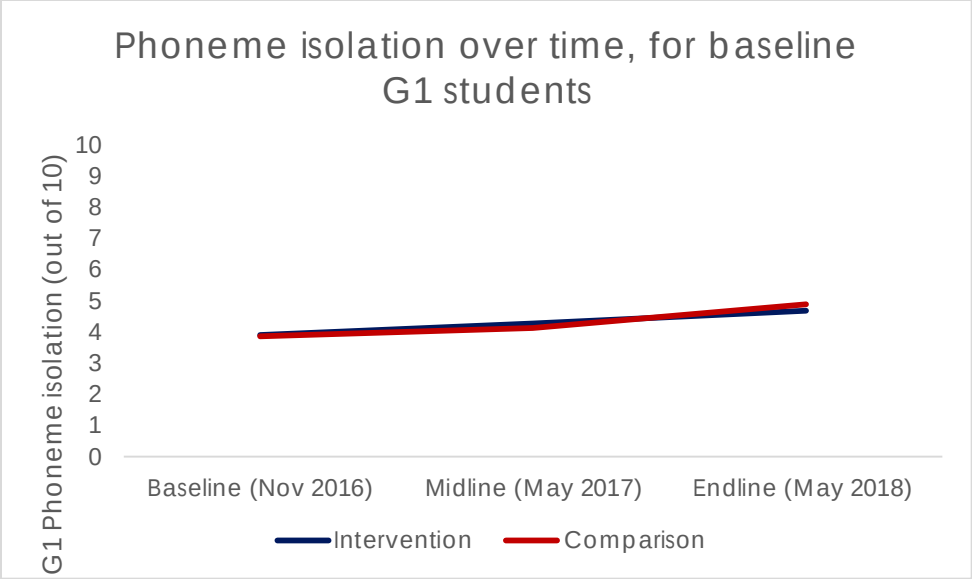

RAMP (Cohort 2)
Comparison (Cohort 3)
0
20
40
60
0
2
4
6
Subtraction facts-L2 (out of 5)

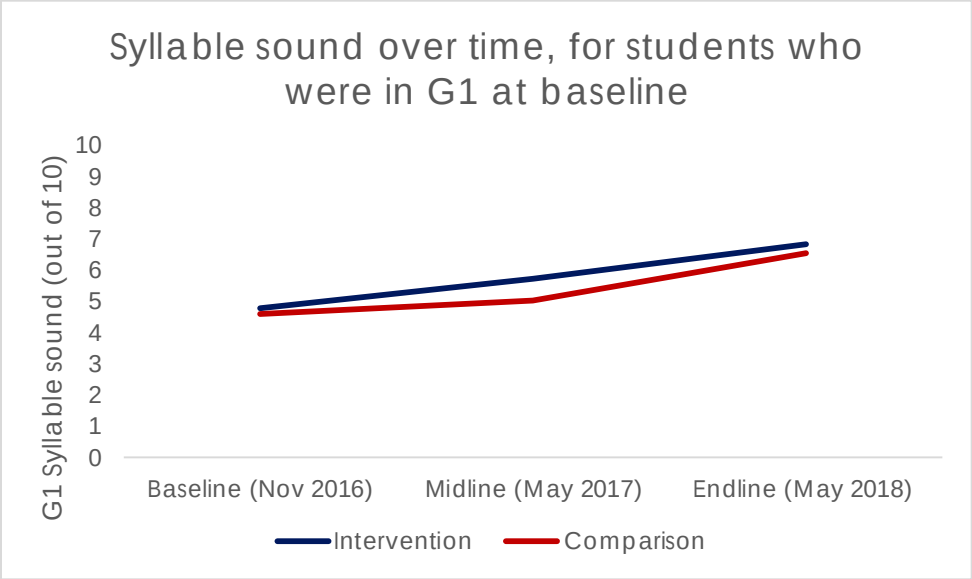
READING AND MATH PERFORMANCE OVER TIME

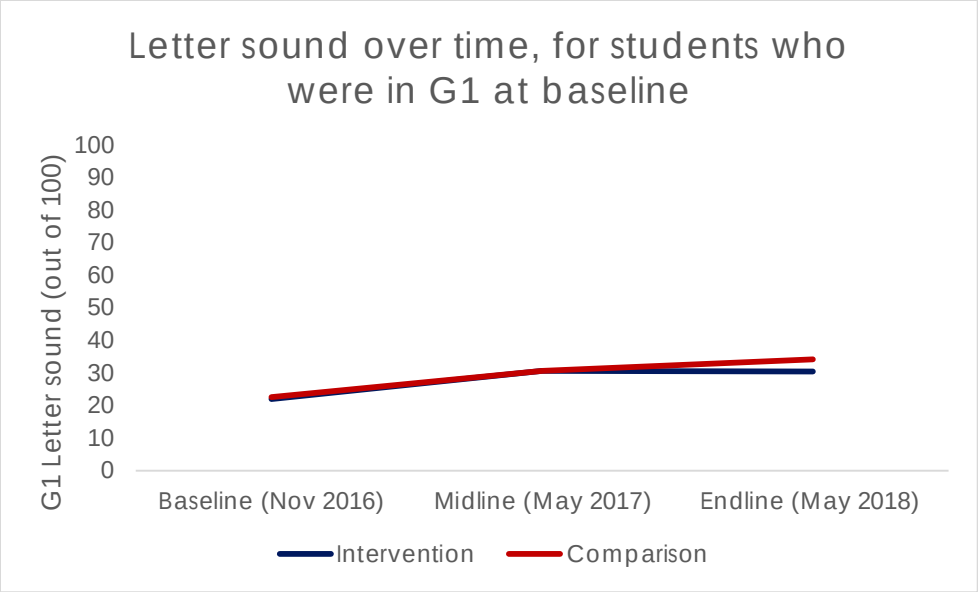
To provide a visual descriptive illustration of change in students' scores over the course of the evaluation, the graphs in panels O.1 to O.4 show point estimates at baseline, midline, and endline. Estimates are ordinary least squares regression-adjusted group means and, in the case of midline and endline, adjust for baseline scores. Analyses were not performed to test whether differences between time points were statistically significant. Midline and endline reading and math scores were transformed (or equated) into the baseline scale. Some estimates are not available (for example, endline orientation to print) because not all subtasks were administered at all three time points.

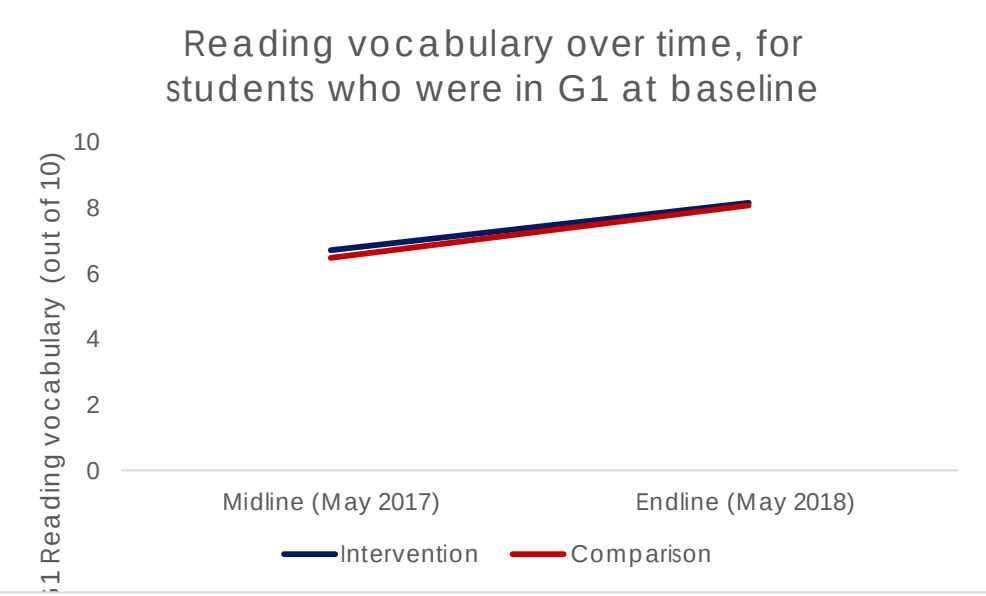
PANEL O.1. G1 STUDENTS' READING SCORES OVER TIME

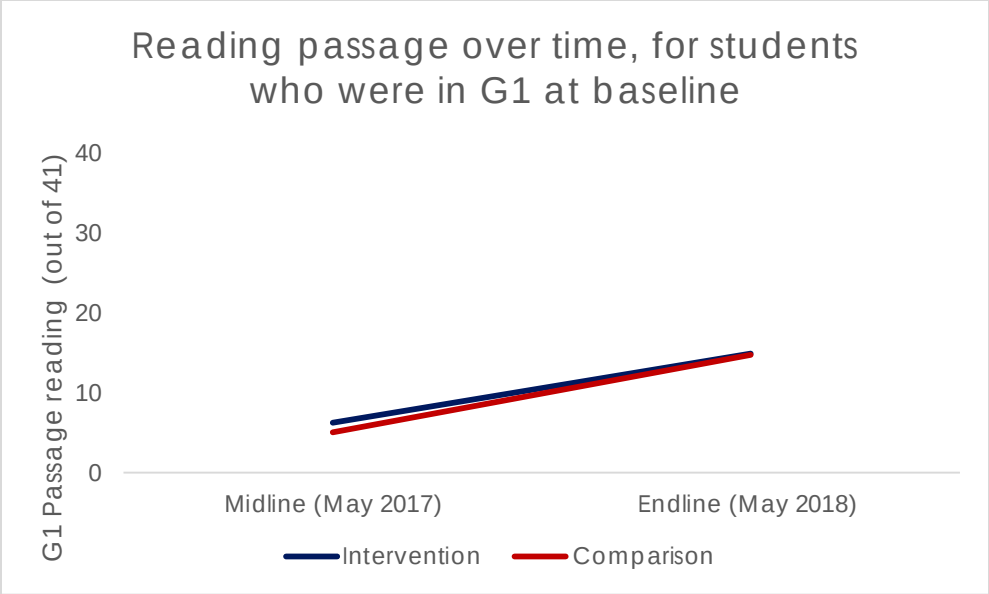


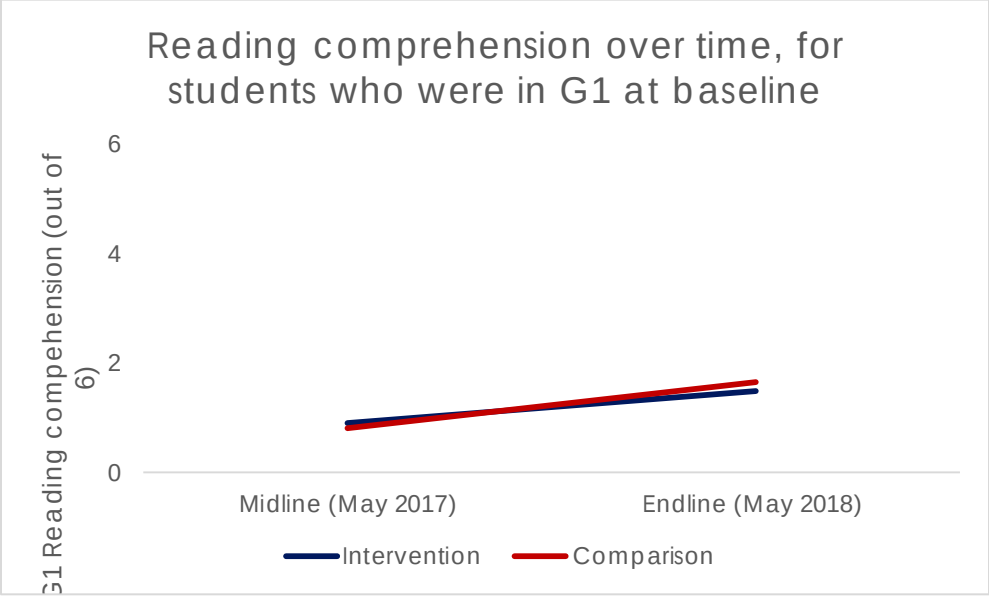

Phoneme isolation over time, for baseline G1 students
G1 Phoneme isolation (out of 10)
10
9
8
7
6
5
4
3
2
1
0
Baseline (Nov 2016)
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison


Syllable sound over time, for students who were in G1 at baseline
G1 Syllable sound (out of 10)
10
9
8
7
6
5
4
3
2
1
0
Baseline (Nov 2016)
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison

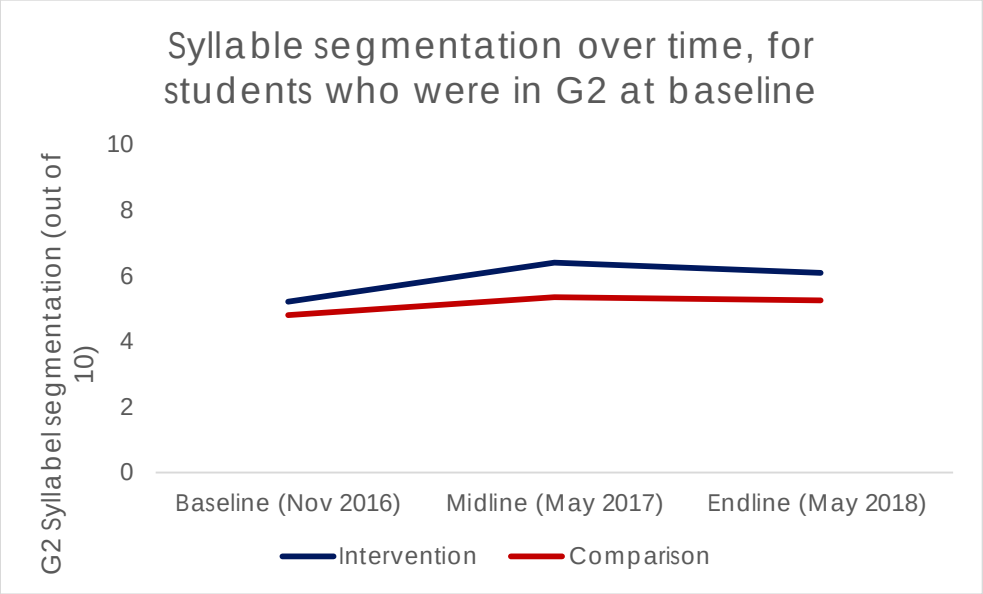

Letter sound over time, for students who were in G1 at baseline
G1 Letter sound (out of 100)
100
90
80
70
60
50
40
30
20
10
0
Baseline (Nov 2016)
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison

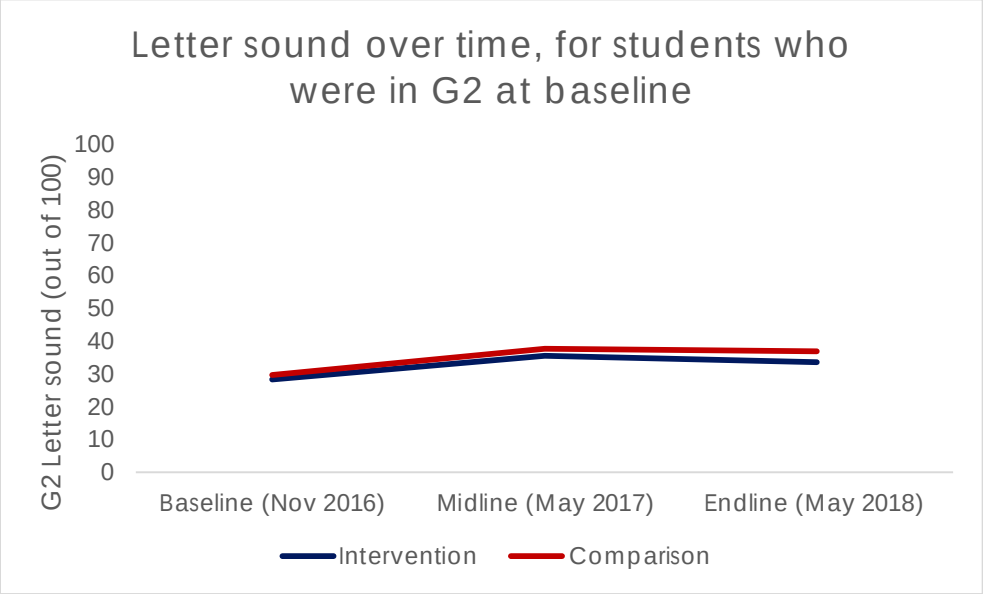

Reading vocabulary over time, for students who were in G1 at baseline
10
8
6
4
2
0
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison

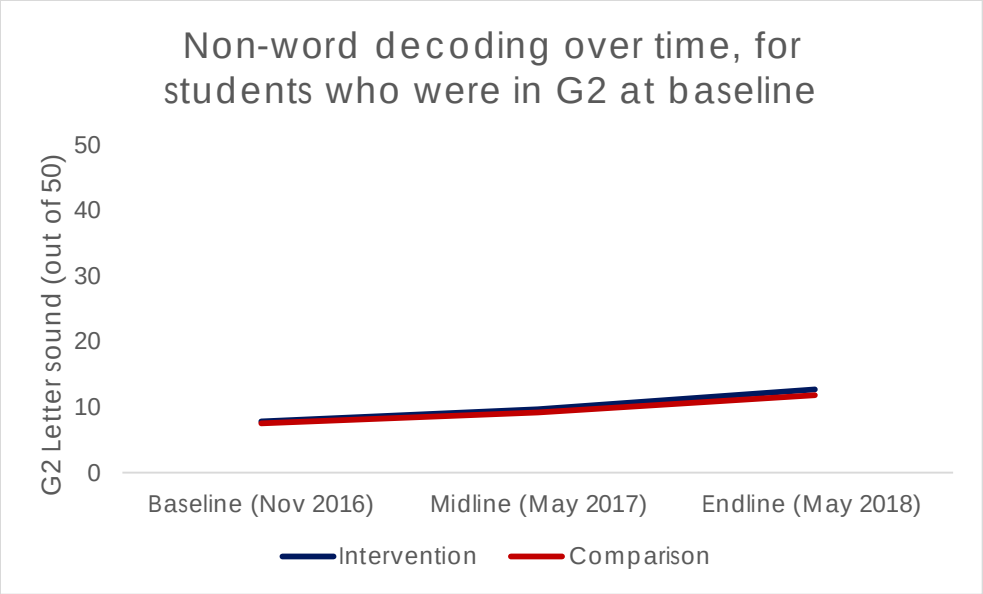

Reading passage over time, for students who were in G1 at baseline
G1 Passage reading (out of 41)
40
30
20
10
0
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison

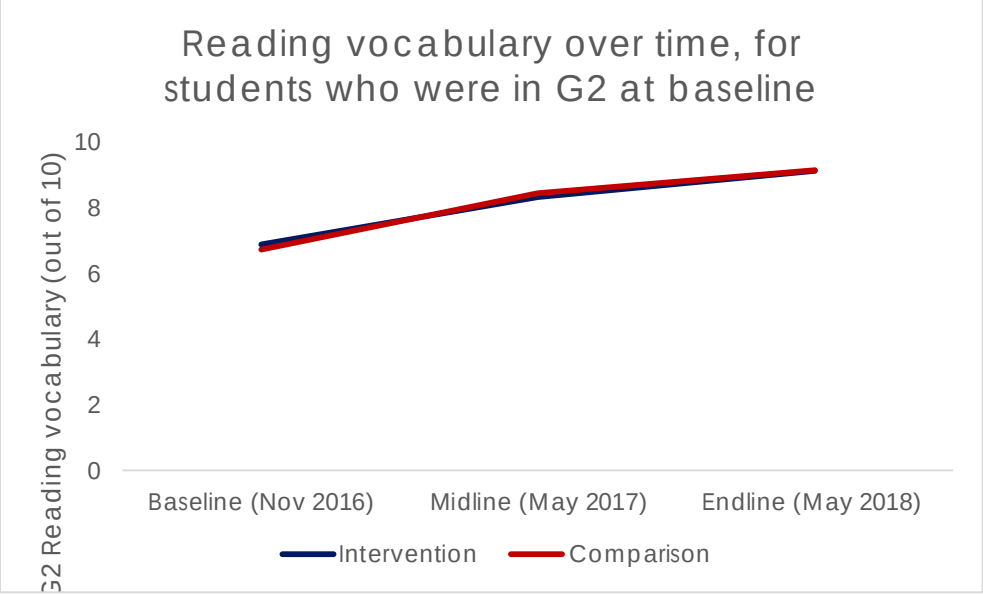

Reading comprehension over time, for students who were in G1 at baseline
G1 Reading compehension (out of 6)
6
4
2
0
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison

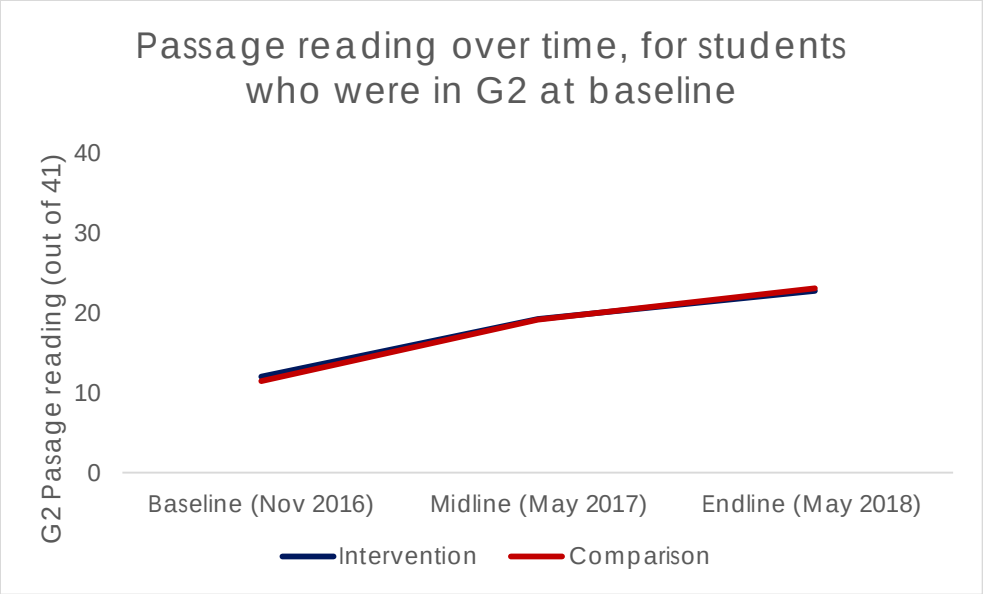
PANEL O.2. G2 STUDENTS' READING SCORES OVER TIME

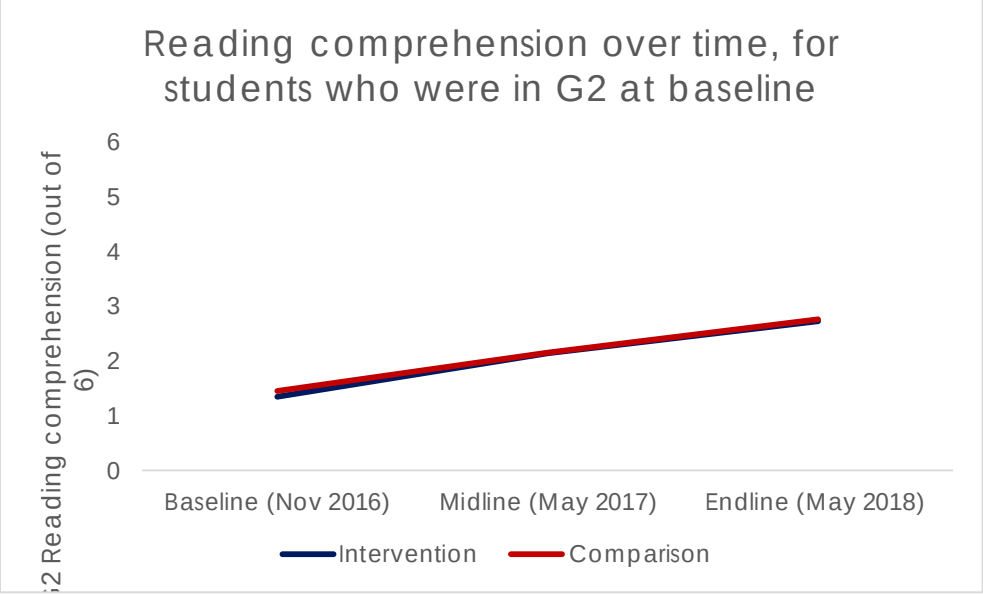




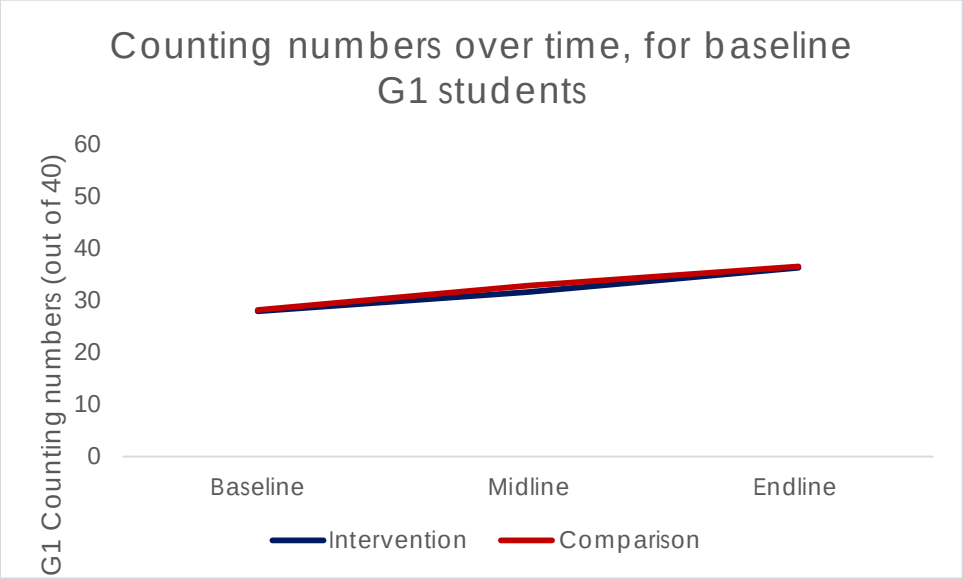

Non-word decoding over time, for students who were in G2 at baseline
G2 Letter sound (out of 50)
50
40
30
20
10
0
Baseline (Nov 2016)
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison

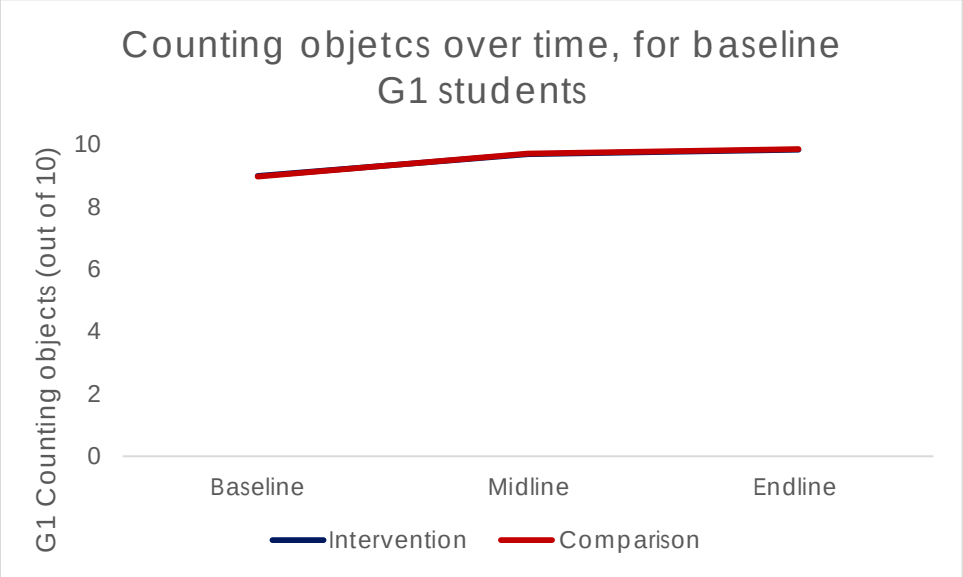

Reading vocabulary over time, for students who were in G2 at baseline
G2 Reading vocabulary (out of 10)
10
8
6
4
2
0
Baseline (Nov 2016)
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison

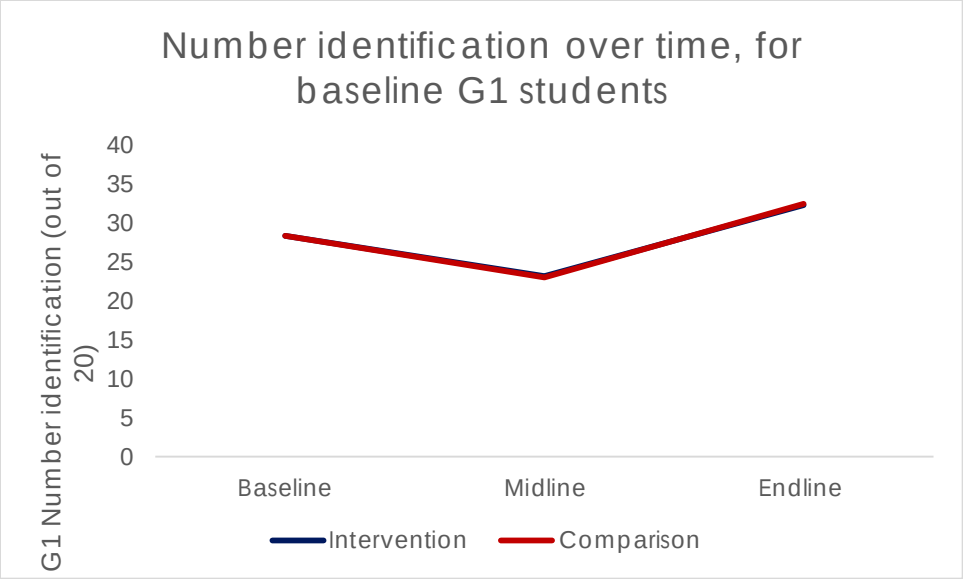

Passage reading over time, for students who were in G2 at baseline
G2 Pasage reading (out of 41)
40
30
20
10
0
Baseline (Nov 2016)
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison

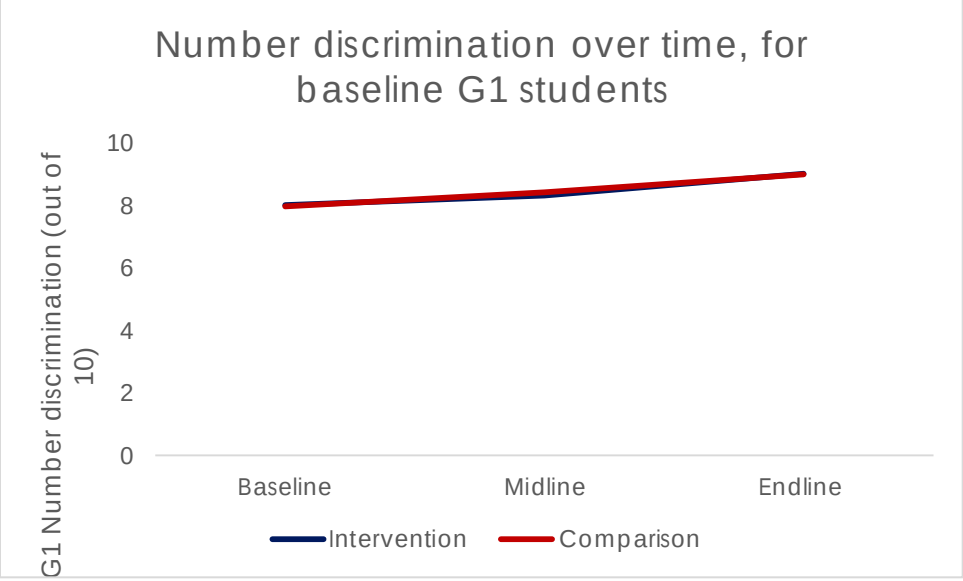

Reading comprehension over time, for students who were in G2 at baseline
6
5
4
3
2
1
0
Baseline (Nov 2016)
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison

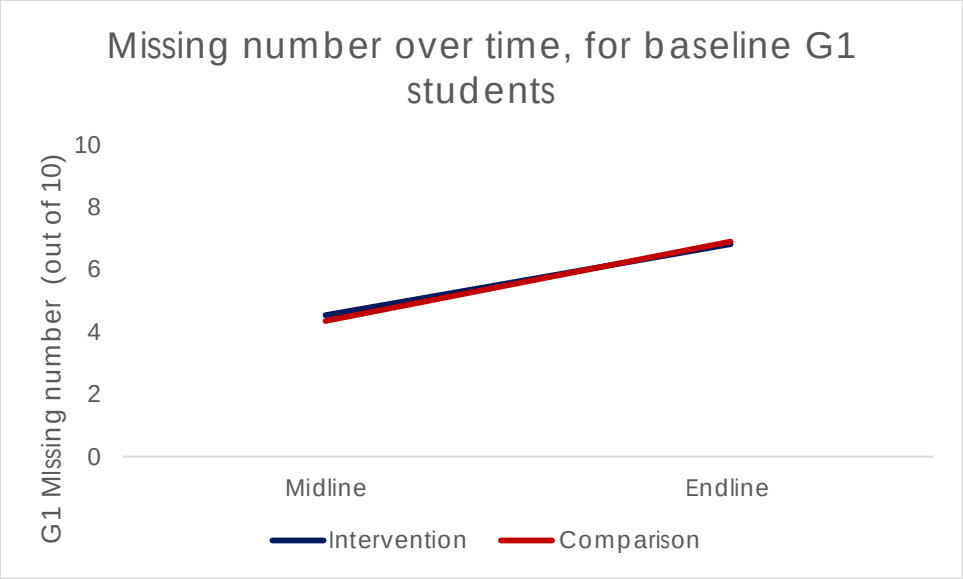
PANEL O.3. G1 STUDENTS' MATHEMATICS SCORES OVER TIME

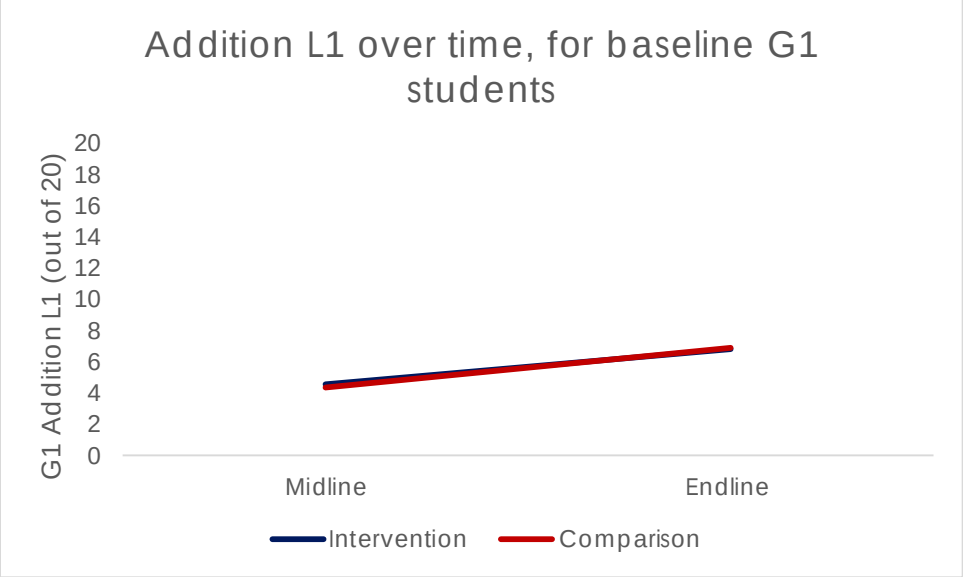




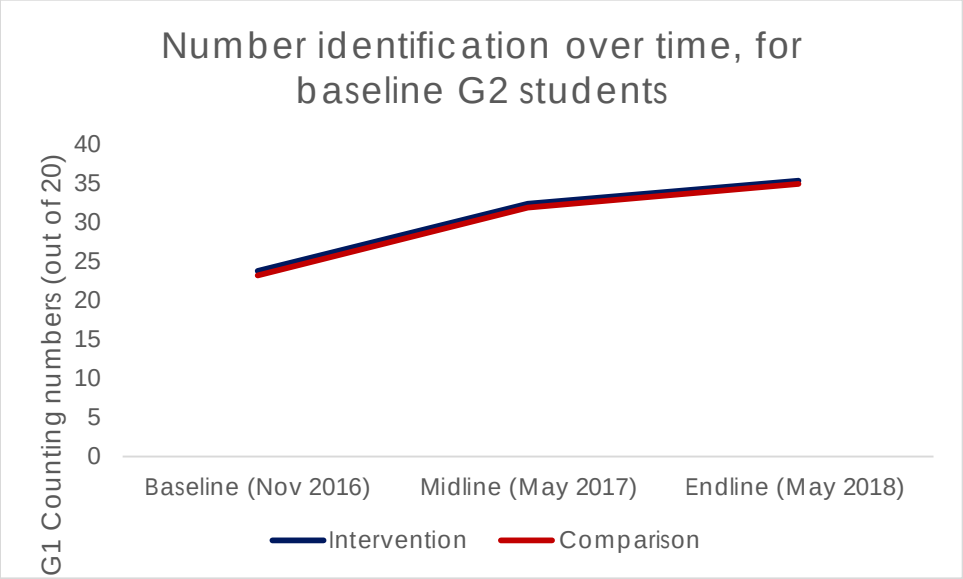

Number identification over time, for baseline G1 students
G1 Number identification (out of 20)
40
35
30
25
20
15
10
5
0
Baseline
Midline
Endline
Intervention
Comparison

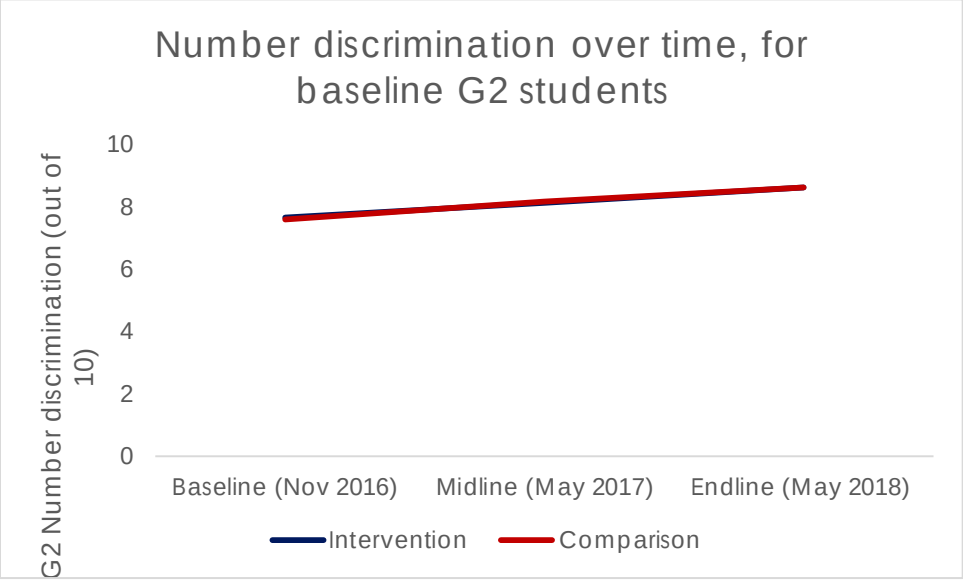

Number discrimination over time, for baseline G1 students
G1 Number discrimination (out of 10)
10
8
6
4
2
0
Baseline
Midline
Endline
Intervention
Comparison

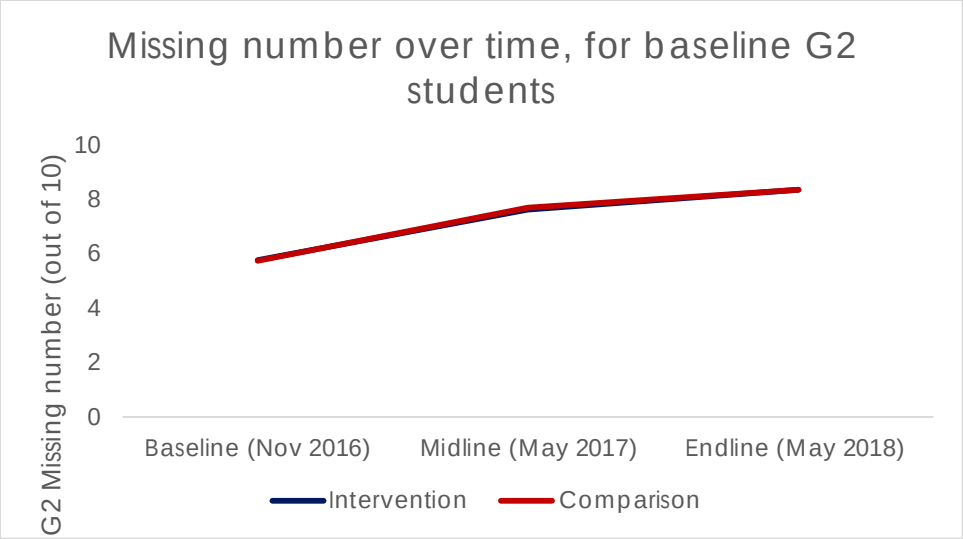

Missing number over time, for baseline G1 students
G1 Missing number (out of 10)
10
8
6
4
2
0
Midline
Endline
Intervention
Comparison

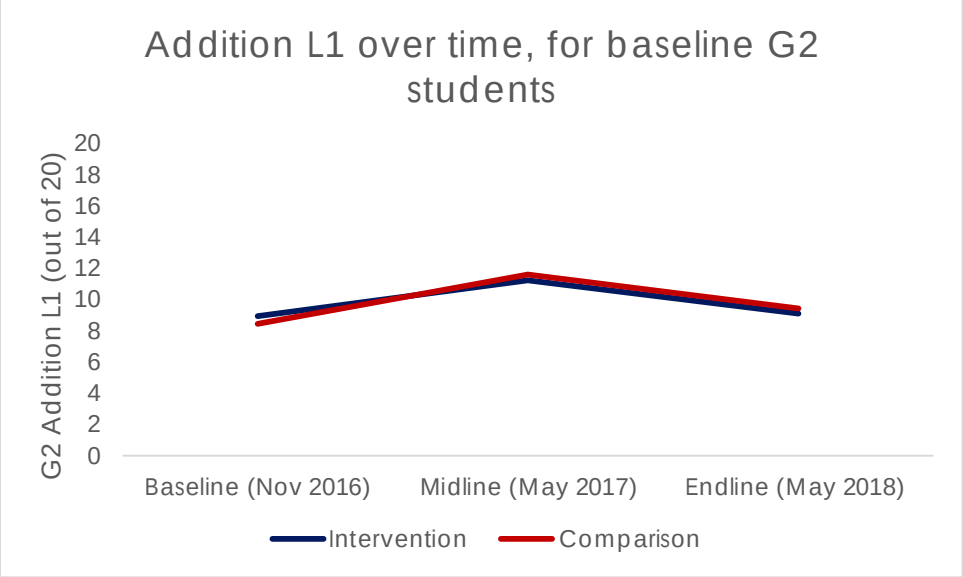

Addition L1 over time, for baseline G1 students
G1 Addition L1 (out of 20)
20
18
16
14
12
10
8
6
4
2
0
Midline
Endline
Intervention
Comparison

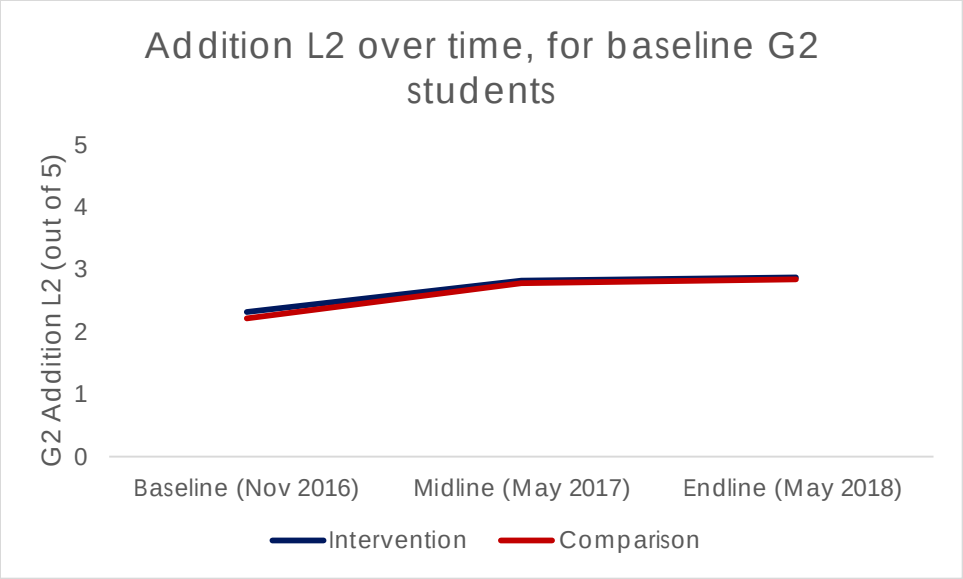
PANEL O.4. G2 STUDENTS' MATHEMATICS SCORES OVER TIME

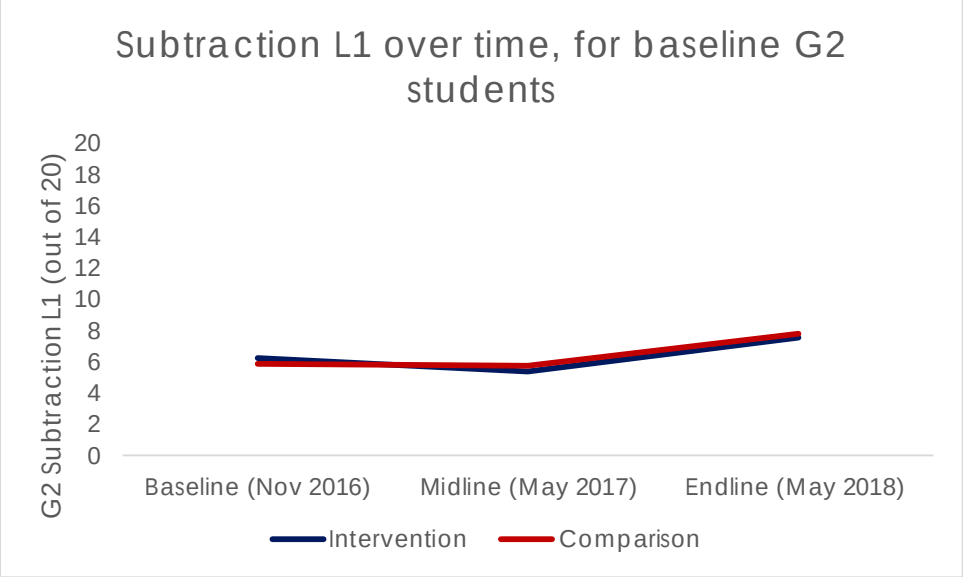


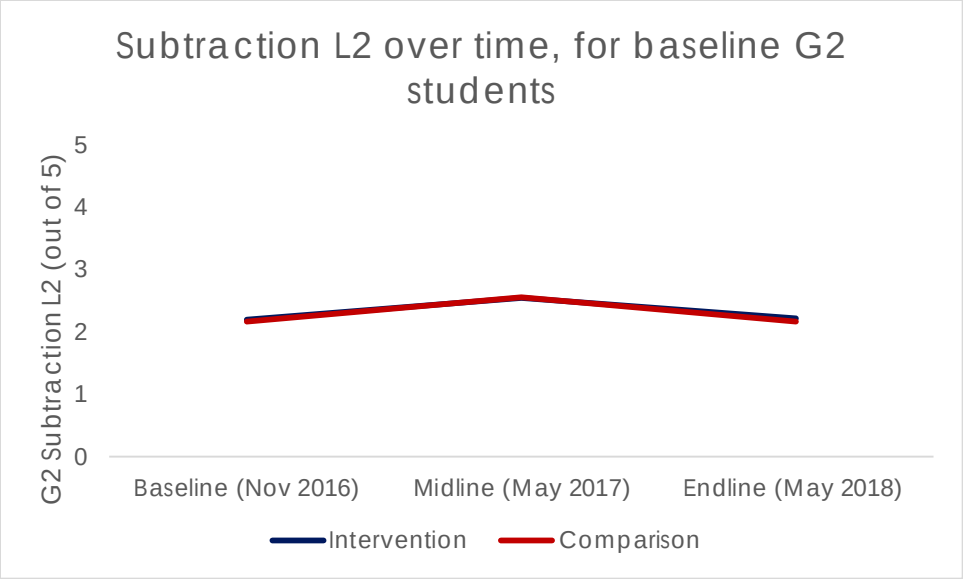



Missing number over time, for baseline G2 students
G2 Missing number (out of 10)
10
8
6
4
2
0
Baseline (Nov 2016)
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison


Addition L1 over time, for baseline G2 students
G2 Addition L1 (out of 20)
20
18
16
14
12
10
8
6
4
2
0
Baseline (Nov 2016)
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison


Addition L2 over time, for baseline G2 students
G2 Addition L2 (out of 5)
5
4
3
2
1
0
Baseline (Nov 2016)
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison


Subtraction L1 over time, for baseline G2 students
G2 Subtraction L1 (out of 20)
20
18
16
14
12
10
8
6
4
2
0
Baseline (Nov 2016)
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison


Subtraction L2 over time, for baseline G2 students
G2 Subtraction L2 (out of 5)
5
4
3
2
1
0
Baseline (Nov 2016)
Midline (May 2017)
Endline (May 2018)
Intervention
Comparison

ANNEX P. TESTING WHETHER RESULTS CHANGE WHEN BASELINE ADJUSTMENTS ARE NOT INCLUDED

EGRA RESULTS: READING BY GRADE, GENDER, AND NUMBER OF SCHOOL SHIFTS WITHOUT BASELINE ADJUSTMENTS

This section presents the impacts of RAMP by grade, gender, and number of school shifts without adjusting for students' baseline performance.[12] The evaluation team considered it important to test whether results changed when baseline scores were not included in the regression because, as mentioned in Annex G, baseline scores could potentially introduce bias in the impact estimates if they reflected early impacts of RAMP. The main analyses, however, adjust for baseline scores for a variety of reasons. First, adjusting for baseline characteristics is required of quasi-experimental designs, where pre-existing differences between the intervention and comparison groups could compromise the study's ability to attribute endline differences to the intervention. Second, including baseline information can substantially increase the precision of impact estimates (Schochet, 2010). Third and last, there was no persuasive evidence to suggest that RAMP had an early impact on students' outcomes after only 6 to 8 weeks of implementation.

Tables P.1-P.4 show that the results by grade did not change when baseline adjustments were not included. RAMP had a statistically significant negative impact on G1 and G2 students' knowledge of letter-sound correspondence, and a positive impact on G2 students' ability to segment words into syllables. In terms of math, RAMP had a negative impact on students' ability to solve basic addition problems, although the impact coefficient was only significant at the 10 percent level when baseline adjustments were not included. RAMP did not have other significant impacts on math scores.

TABLE P.1. ENDLINE READING PERFORMANCE SCORES FOR GRADE 1 STUDENTS, WITHOUT BASELINE ADJUSTMENTS

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Effect Size	Number of Students
Phoneme Isolation (Out of 10)	4.9	5.1	-0.2	0.42	-0.1	1931
Syllable Segmentation (Out of 10)	7.1	6.7	0.4	0.19	0.1	1931
Letter Sound Knowledge (Out of 100, Prorated)	31.7	35.1	-3.5*	0.028	-0.2	1931
Reading Vocabulary (Out of 10)	8.4	8.3	0.1	0.44	0.1	1931
Passage Reading (Out of 41, Prorated Score)	16.2	15.9	0.3	0.75	0.0	1931

12 The results in the body of the report are from regressions that adjust for all baseline scores.

Reading Comprehension (Out of 6)	1.7	1.8	-0.1	0.41	-0.1	1931
Number of Schools	117	120				

Source: RAMP Impact Study - Endline Data 2018 Student Assessments

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes for continuous outcomes are Hedges' g. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated into the midline scale.

*Difference in group means is statistically significant at the .05 level.

TABLE P.2. ENDLINE READING PERFORMANCE SCORES FOR GRADE 2 STUDENTS, WITHOUT BASELINE ADJUSTMENTS

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Effect Size	Number of Students
Syllable Segmentation (out of 10)	6.2	5.3	0.9*	0.004	0.3	1931
Letter Sound Knowledge (out of 100, Prorated)	33.7	37.3	-3.6*	0.037	-0.2	1931
Non-Word Decoding (out of 50, Prorated)	13.0	12.2	0.8	0.34	0.1	1931
Reading Vocabulary (out of 10)	9.2	9.2	0.0	0.82	0.0	1931
Passage Reading (out of 41, Prorated Score)	23.0	23.7	-0.7	0.51	0.0	1931
Reading Comprehension (out of 6)	2.8	2.8	-0.1	0.60	0.0	1931
Number of Schools	118	119				

Source: RAMP Impact Study - Endline Data 2018 Student Assessments

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes for continuous outcomes are Hedges' g. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated into the midline scale.

*Difference in group means is statistically significant at the .05 level

TABLE P.3. MATH PERFORMANCE SCORES FOR GRADE 1 STUDENTS AT ENDLINE, WITHOUT BASELINE ADJUSTMENTS

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Effect Size	Number of Students
Counting Numbers (Out of 40)	36.8	36.9	-0.1	0.86	0.0	1931
Enumerating Quantities (Out of 10)	9.9	9.9	0.0	0.52	0.0	1931
Number Identification (Out of 20, Prorated)	33.8	33.5	0.3	0.76	0.0	1931
Number Discrimination (Out of 10)	9.2	9.2	0.1	0.62	0.0	1931
Missing Numbers (Out of 10)	7.1	7.1	0.0	0.87	0.0	1931

Addition Facts - L1 (Out of 20, Prorated)	5.7	6.5	-0.8	0.08	-0.1	1931
Number of Schools	117	120				

Source: RAMP Impact Study - Endline Data 2018 Student Assessments

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes for continuous outcomes are Hedges' g. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated into the midline scale.

*Difference in group means is statistically significant at the .05 level

TABLE P.4. MATH PERFORMANCE SCORES FOR GRADE 2 STUDENTS AT ENDLINE, WITHOUT BASELINE ADJUSTMENTS

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Effect Size	Number of Students
Number Identification (out of 20, Prorated)	36.1	35.7	0.4	0.72	0.0	1931
Number Discrimination (out of 10)	8.8	8.7	0.0	0.85	0.0	1931
Missing Numbers (out of 10)	8.5	8.5	0.0	0.99	0.0	1931
Addition Facts - L1 (out of 20, Prorated) Timed	9.5	9.6	-0.1	0.79	0.0	1931
Addition Facts - L2 (out of 5)	3.0	2.9	0.0	0.76	0.0	1931
Subtraction Facts - L1 (out of 20, Prorated)	7.3	7.5	-0.1	0.79	0.0	1931
Subtraction Facts - L2 (out of 5)	2.3	2.3	0.1	0.63	0.0	1931
Number of Schools	118	119				

Source: RAMP Impact Study - Endline Data 2018 Student Assessments

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes for continuous outcomes are Hedges' g. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated into the midline scale.

*Difference in group means is statistically significant at the .05 level

The results of the subgroup analysis by gender and number of school shifts also remained unchanged when baseline adjustments were excluded (full results available upon request). The impacts of RAMP did not vary depending on students' gender in either grade, but there were differences in RAMP impacts depending on the number of school shifts. The findings on reading vocabulary and addition were consistent to results that adjusted for baseline scores (G1 reading vocabulary: no impact in single-shift schools and positive impact in double-shift schools; G1 addition: negative impact in single-shift schools and no impact in double-shift schools). The models that did not adjust for baseline scores found a difference that was significant at the 10 percent level only for G1 passage reading and reading comprehension, whereas the models that adjusted for baseline scores showed statistically significant differences at the 5 percent level.

Despite differences in the level of statistical significance, the findings were consistent with the overall pattern of negative impacts in single-shift schools and positive impacts (or no impacts) in double-shift schools.

ANNEX Q. TEACHER QUESTIONNAIRES AS PART OF THE IMPACT STUDY OF STUDENTS' LEARNING

Teachers of G2 and G3 students who were assessed as part of the impact study of students' learning were administered an endline questionnaire to complement other data sources. Teachers were asked about their background, classroom resources, and support received from school and RAMP staff. This section presents comparisons between teachers' responses in intervention and comparison schools, to provide context for findings from the impact study of students' learning. Causal conclusions about the impact of RAMP cannot be drawn from these results, as baseline equivalence was not established for these outcomes.

When interpreting these results, RAMP's implementation timeline should also be borne in mind. G2 teachers in the comparison group were expected to start receiving RAMP support in 2017, while G3 teachers were expected to start in 2018 (after completion of data collection for this evaluation; see Annex A). Thus, G2 comparisons reflect two years of RAMP for teachers in the intervention group versus one year of exposure for teachers in the comparison group, whereas G3 comparisons reflect two years of RAMP for teachers in intervention schools versus no exposure for teachers in comparison schools.

TEACHER AND CLASSROOM CHARACTERISTICS

At endline, G2 and G3 teachers in intervention and comparison schools did not significantly differ in characteristics that were not expected to change as a result of RAMP, including gender, years of teaching experience, having a permanent teaching position, and level of education. There were also no significant differences between the groups in the number of students in the classroom, the percentage of students who were repeating a grade, or the percentage absent on a regular day (see Tables Q.1 and Q.2).

There were some statistically significant differences in classroom resources, but it is unclear if such differences may have been the result of RAMP. In both grades, intervention teachers were more likely than comparison teachers to report having a classroom library they used (this difference was statistically significant at the 10 percent level for G2), a white board, and tablets (this was not statistically significant for G3). The differences were still significant after adjusting for school- and teacher-level covariates, except for the presence of tablets in the classroom, which was no longer significant when school-level covariates were included (results from robustness checks available upon request).

G3 intervention teachers, but not G2 teachers, were 26 to 27 percentage points more likely than comparison teachers were to report having teacher guides for language and reading, and mathematics. Given that teacher guides were part of the materials that RAMP planned to develop, it is noteworthy that less than half of all teachers reported having teacher guides for either subject, despite the statistically significant difference in favor of G3 teachers in RAMP schools. These

differences were still statistically significant after adjusting for school and teacher characteristics (results from robustness checks available upon request).

TABLE Q.1. G2 TEACHER AND CLASSROOM CHARACTERISTICS AT ENDLINE

Variable	Intervention (T)	Comparison (C)	Difference (T-C) *	P-Value	Effect Size	Number of Teachers
Teacher Characteristics						
Teacher is Female	82.8	85.0	-2.3	0.65	-0.1	227
Teacher Teaches a Multi-Grade Classroom (Students in Different Grades)	0.0	2.7	-2.7	0.08	NA	227
Number of Years Teaching	10.8	10.3	0.6	0.52	0.1	226
Teacher has a Permanent Teaching Position	90.4	87.5	3.0	0.48	0.2	227
Highest Level of Education						
Diploma	3.3	4.4	-1.1	0.67	-0.2	227
Baccalaureate	71.8	74.3	-2.6	0.66	-0.1	227
Higher Diploma, Masters or Ph.D. Degree	24.9	21.2	3.7	0.51	0.1	227
Classroom Characteristics						
Number of Students in the Classroom	25.8	25.0	0.7	0.53	0.1	227
Percentage of Students Repeating a Grade	2.4	0.6	1.9	0.38	0.1	227
Percentage of Students Absent on a Regular Day	7.5	7.7	-0.3	0.77	0.0	225
Classroom Resources						
Classroom Library Used by Teacher	55.1	43.3	11.8	0.08	0.3	227
Enough Reading Books for Students	69.1	64.6	4.5	0.48	0.1	226
Enough Textbooks for Students	100.0	99.1	0.9	0.32	NA	227
Language or Reading Teacher Guides	45.9	37.8	8.1	0.22	0.2	227
Mathematics Teacher Guides	45.0	38.1	6.9	0.29	0.2	226
White Board	83.0	66.9	16.1*	0.006	0.5	222
Smart Board	3.6	4.6	-1.1	0.69	-0.2	222
Tablets	6.2	0.9	5.3*	0.034	1.2	222
Other	42.5	52.4	-9.9	0.14	-0.2	222
Number of Schools	112	113				

Source: Ramp Impact Study - Endline Data 2018 Teacher Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). All regressions include school sampling weights. Effect sizes for continuous outcomes are Hedges' g. Effect sizes for dichotomous outcomes are Cox index. Effect sizes are "NA" (not applicable) when the standard deviation for one or both groups equals zero. Sample sizes vary due item nonresponse.

*Difference in group means is statistically significant at the .05 level.

TABLE Q.2. G3 TEACHER AND CLASSROOM CHARACTERISTICS AT ENDLINE

Variable	Intervention (T)	Comparison (C)	Difference (T-C) *	P-Value	Effect Size	Number of Teachers
Teacher Characteristics						
Teacher is Female	82.2	83.1	-0.9	0.86	0.0	234
Teacher Teaches a Multi-Grade Classroom (Students in Different Grades)	0.9	0.0	0.9	0.32	NA	234
Number of Years Teaching	11.0	11.4	-0.4	0.64	-0.1	234
Teacher has a Permanent Teaching Position	85.4	83.0	2.4	0.61	0.1	234
Highest Level of Education						
Diploma	4.3	6.7	-2.3	0.43	-0.3	234
Baccalaureate	71.4	74.5	-3.0	0.61	-0.1	234
Higher Diploma, Masters or Ph.D. Degree	24.2	18.9	5.4	0.32	0.2	234
Classroom Characteristics						
Number of Students in the Classroom	25.8	24.6	1.1	0.35	0.1	234
Percentage of Students Repeating a Grade	0.9	0.3	0.6	0.06	0.3	234
Percentage of Students Absent on a Regular Day	6.5	7.7	-1.1	0.11	-0.2	233
Classroom Resources						
Classroom Library Used by Teacher	55.0	38.1	16.9*	0.010	0.4	234
Enough Reading Books for Students	69.3	61.1	8.1	0.19	0.2	234
Enough Textbooks for Students	99.1	99.1	0.0	1.00	0.0	233
Language or Reading Teacher Guides	44.4	17.7	26.7*	0.000	0.8	233
Mathematics Teacher Guides	44.4	17.8	26.5*	0.000	0.8	232
White Board	81.5	62.2	19.3*	0.001	0.6	229
Smart Board	5.3	4.3	1.0	0.72	0.1	229
Tablets	3.5	0.9	2.6	0.18	0.9	229
Other	41.7	51.7	-10.1	0.13	-0.2	229

Source: Ramp Impact Study - Endline Data 2018 Teacher Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). All regressions include school sampling weights. Effect sizes for continuous outcomes are Hedges' g. Effect sizes for dichotomous outcomes are Cox index. Effect sizes are "NA" (not applicable) when the standard deviation for one or both groups equals zero. Sample sizes vary due item nonresponse.

*Difference in group means is statistically significant at the .05 level.

TEACHER TRAINING

Overall, there were no statistically significant differences between G2 teachers in the intervention and comparison groups, but there were several differences for G3 teachers.

The proportion of G2 teachers who participated in pre-service and in-service trainings was not statistically significantly different for teachers in intervention versus comparison schools. In fact, most G2 teachers in both groups reported participating in in-service trainings. Pre-service training was far less common, with only nine percent of teachers in both groups having participated (see Table Q.3). The proportion of G2 teachers who received RAMP training, the number of training days received, whether a RAMP coach or supervisor had observed the teacher, and the frequency of observations, also did not differ significantly between the groups.

In contrast, G3 teachers in the intervention group were 27 to 35 percentage points more likely than comparison teachers to report participating in in-service trainings (see Table Q.4). The proportion of G3 teachers who received RAMP training, the number of training days received, whether a RAMP coach or supervisor had observed the teacher, and the frequency of observations, also differed significantly between the groups. Contrary to expectations, there was a non-zero proportion of G3 teachers in the comparison group who reported receiving RAMP support. This indicates contamination between the intervention and comparison groups that could hamper the evaluation's ability to detect the true impact of RAMP. These results were robust to the inclusion of school- and teacher-level covariates (results from robustness checks are available upon request).

TABLE Q.3. G2 RAMP TRAINING

Variable	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Effect Size	Number of Teachers
Pre-Service Training						
Teacher Received Pre-Service Teaching Training in Teaching Reading	8.9	8.9	0.0	0.99	0.0	227
Teacher Received Pre-Service Training in Teaching Math	6.2	8.9	-2.7	0.45	-0.2	227
In-Service Training						
Teacher Received in-Service Teaching Training	93.8	92.8	1.0	0.77	0.1	227
Teacher Received in-Service Training In Phonics	71.1	70.6	0.6	0.93	0.0	226
Teacher Received in-Service Training In Teaching Math	89.4	91.9	-2.5	0.51	-0.2	227
RAMP Training						
Teacher Received Training From RAMP	90.2	91.0	-0.8	0.84	-0.1	227
Number of Days in Which Teacher Received RAMP Training	3.9	3.8	0.1	0.09	0.2	206
RAMP Coach/Supervisor Observed Teacher	92.0	92.9	-0.9	0.80	-0.1	227
Number of Times RAMP Coach/Supervisor Observed Teacher						

1-3 Times	27.7	20.2	7.5	0.21	0.3	210
2-3 Times	18.9	30.2	-11.3	0.06	-0.4	210
4-6 Times	53.5	49.7	3.8	0.58	0.1	210
Number of Schools	114	113				

Source: Ramp Impact Study - Endline Data 2018 Teacher Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). The regressions use weights to account for school sampling probabilities. Effect sizes for continuous outcomes are Hedges' g. Effect sizes for dichotomous outcomes are Cox index. Sample sizes vary due item nonresponse.

TABLE Q.4. G3 RAMP TRAINING

Variable	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Effect Size	Number of Teachers
Pre-Service Training						
Teacher Received Pre-Service Teaching Training in Teaching Reading	5.3	10.2	-5.0	0.16	-0.4	234
Teacher Received Pre-Service Training in Teaching Math	5.2	6.0	-0.8	0.79	-0.1	234
In-Service Training						
Teacher Received in-Service Teaching Training	91.4	61.0	30.4*	0.000	1.2	234
Teacher Received in-Service Training In Phonics	65.3	30.6	34.7*	0.000	0.9	225
Teacher Received in-Service Training In Teaching Math	87.9	60.9	27.0*	0.000	0.9	234
RAMP Training						
Teacher Received Training from RAMP	86.2	18.8	67.4*	0.000	2.0	234
Number of Days in which Teacher Received RAMP Training	3.9	3.1	0.8*	0.005	1.2	122
RAMP Coach/Supervisor Observed Teacher	95.7	38.2	57.5*	0.000	2.2	234
Number of Times RAMP Coach/Supervisor Observed Teacher						
1-3 Times	28.2	82.4	-54.3*	0.000	-1.5	156
2-3 Times	31.8	8.5	23.3*	0.000	1.0	156
4-6 Times	40.0	9.1	31.0*	0.000	1.1	156
Number Of Schools	116	118				

Source: Ramp Impact Study - Endline Data 2018 Teacher Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted school-level group means (or percentages). The regressions use weights to account for school sampling probabilities. Effect sizes for continuous outcomes are Hedges' g. Effect sizes for dichotomous outcomes are Cox index. Sample sizes vary due item nonresponse.

*Difference in group means is statistically significant at the .05 level.

TEACHERS' USE OF ASSESSMENTS AND SUPPORTS FOR STUDENTS WITH ACADEMIC OR BEHAVIORAL DIFFICULTIES

RAMP's development hypothesis proposes that training teachers in the use of assessments would improve their ability to respond to students' individual needs and, therefore, lead to improvements in students' learning. G2 and G3 teachers were asked about their use of assessment strategies, as well as their approaches to supporting students with academic or behavioral difficulties.

Overall, the results revealed no statistically significant differences between the groups in either grade, in the proportion of teachers who used a series of assessment methods, nor in how teachers used the results from those assessments (see Tables Q.5 and Q.6). As expected, nearly 90 percent

of G2 teachers in both the intervention and comparison reported using the RAMP fine-grain assessment tool, although there was substantial variation in how frequently they used it. The proportion of G3 intervention teachers who used the fine-grain tool was statistically significantly higher than the proportion in the comparison group, but a non-trivial number of comparison teachers reported using it.

TABLE Q.5. G2 TEACHERS' USE OF ASSESSMENTS AT ENDLINE

Variable	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Effect Size	Number of Teachers
Methods Used to Assess Student Performance						
Written Assessments	72.0	73.4	-1.4	0.82	0.0	227
Oral Assessments	58.4	50.8	7.6	0.25	0.2	227
Research and Projects	4.5	3.5	0.9	0.72	0.1	227
Homework	22.0	18.8	3.2	0.55	0.1	227
Participation and Discussion	40.1	34.2	5.9	0.36	0.2	227
Worksheets	56.0	58.5	-2.6	0.70	-0.1	227
End of Semester Tests	46.7	42.5	4.2	0.53	0.1	227
Other	17.7	12.9	4.7	0.32	0.2	227
Results from Oral and Written Assessments are Used to						
Grade Students	47.4	50.3	-2.9	0.67	-0.1	220
Assess Students' Learning and Understanding	54.9	46.2	8.7	0.20	0.2	220
Plan for Teaching	23.7	21.1	2.6	0.65	0.1	220
Adapt Teaching Methods	39.9	35.5	4.5	0.50	0.1	220
Develop New Teaching Strategies	11.8	11.1	0.7	0.87	0.0	220
Other	18.8	12.0	6.8	0.16	0.3	220
Tools and Strategies to Assess Reading Abilities						
Reading Aloud in Class	91.3	86.5	4.9	0.25	0.3	226
Fluency (Speed and Accuracy) Exercises	27.3	19.6	7.6	0.18	0.3	226
Listening Comprehension Questions	24.5	24.1	0.4	0.95	0.0	226
Reading Comprehension Questions	23.9	24.1	-0.2	0.97	0.0	226
Retelling Stories	31.5	25.0	6.6	0.28	0.2	226
Worksheets	23.1	21.3	1.7	0.76	0.1	226
Does Not Use Tools and Strategies To Assess Reading Abilities	0.0	0.9	-0.9	0.32	NA	227
Use of RAMP Tools						
Teacher Used RAMP Fine-Grain Assessment Tools	87.8	88.3	-0.6	0.89	0.0	227
Number of Times Teacher Used RAMP Assessment Tools in the Past Year						
One Time	45.4	40.4	4.9	0.46	0.1	219
2-3 Times	30.0	32.0	-2.0	0.75	-0.1	219

More Than 3 Times	15.4	19.1	-3.7	0.47	-0.2	219
Number Of Schools	114	113				

Source: Ramp Impact Study - Endline Data 2018 Teacher Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted school-level group means (or percentages). The regressions use weights to account for school sampling probabilities. Effect sizes for dichotomous outcomes are Cox index. Effect sizes are "NA" (not applicable) when the standard deviation for one or both groups equals zero. Sample sizes vary due item nonresponse.

*Difference in group means is statistically significant at the .05 level.

TABLE Q.6. G3 TEACHERS' USE OF ASSESSMENTS AT ENDLINE

Variable	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Effect Size	Number of Teachers
Methods Used to Assess Student Performance						
Written Assessments	76.8	72.7	4.1	0.47	0.1	233
Oral Assessments	54.2	51.0	3.3	0.62	0.1	233
Research and Projects	5.1	4.3	0.8	0.78	0.1	233
Homework	29.8	23.8	6.0	0.31	0.2	233
Participation and Discussion	41.3	39.3	2.0	0.75	0.1	233
Worksheets	53.7	59.3	-5.6	0.39	-0.1	233
End of Semester Tests	42.5	40.9	1.7	0.80	0.0	233
Other	19.2	12.0	7.3	0.13	0.3	233
Results from Oral and Written Assessments are Used to						
Grade Students	50.9	54.8	-3.8	0.56	-0.1	230
Assess Students' Learning and Understanding	45.6	46.5	-1.0	0.89	0.0	230
Plan for Teaching	22.7	16.4	6.2	0.24	0.2	230
Adapt Teaching Methods	35.1	34.7	0.4	0.95	0.0	230
Develop New Teaching Strategies	13.0	14.5	-1.4	0.75	-0.1	230
Other	20.9	14.7	6.2	0.22	0.3	230
Tools and Strategies to Assess Reading Abilities						
Reading Aloud in Class	96.5	97.4	-0.9	0.70	-0.2	231
Fluency (Speed and Accuracy) Exercises	28.1	18.2	9.9	0.08	0.3	231
Listening Comprehension Questions	28.9	22.2	6.7	0.25	0.2	231
Reading Comprehension Questions	21.2	17.1	4.1	0.43	0.2	231
Retelling Stories	29.2	21.4	7.8	0.18	0.2	231
Worksheets	23.9	14.5	9.4	0.07	0.4	231
Does Not Use Tools and Strategies to Assess Reading Abilities	1.7	0.9	0.8	0.58	0.4	234
Use of RAMP Tools						
Teacher Used RAMP Fine-Grain Assessment Tools	81.8	14.6	67.2*	0.000	2.0	234

Variable	Intervention (T)	Comparison (C)	Impact (T-C) *	P-Value	Effect Size	Number of Teachers
Number of Times Teacher Used RAMP Assessment Tools in the Past Year						
One Time	40.4	10.5	29.9	0.000	1.06	212
2-3 Times	36.1	3.8	32.3	0.000	1.60	212
More Than 3 Times	13.1	1.9	11.2	0.002	1.24	212
Number of Schools	116	118				

Source: Ramp Impact Study - Endline Data 2018 Teacher Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted school-level group means (or percentages). The regressions use weights to account for school sampling probabilities. Effect sizes for dichotomous outcomes are Cox index. Effect sizes are "NA" (not applicable) when the standard deviation for one or both groups equals zero. Sample sizes vary due item nonresponse.

*Difference in group means is statistically significant at the .05 level.

The results also showed few statistically significant differences between the groups in the strategies that G2 and G3 teachers used to support students with academic or behavioral difficulties. There were three exceptions: G2 teachers in intervention schools were 10 percentage points more likely than teachers in comparison schools to request assistance from parents to support students with reading difficulties; G3 intervention teachers were 14 percentage points less likely than comparison teachers to assign additional homework to disobedient or disruptive students; and G3 teachers in intervention schools were also 5 percentage points less likely to report using corporal punishment. These differences were statistically significant, but it is unclear if these were topics addressed during RAMP trainings. Also, only the difference in the use of corporal punishment was still statistically significant after adjusting for school-level covariates.

TABLE Q.7. G2 TEACHERS' STRATEGIES TO SUPPORT STUDENTS WITH ACADEMIC OR BEHAVIORAL DIFFICULTIES, AT ENDLINE

Variable	Intervention (T)	Comparison (C)	Difference (T-C) *	P-Value	Effect Size	Number of Teachers
Methods to Teach Underachieving Students						
Focus on Weak Students	88.7	81.6	7.0	0.14	0.3	227
Administer Daily Quizzes	11.3	9.6	1.7	0.68	0.1	227
Encouragement	31.1	24.6	6.5	0.28	0.2	227
Communicate with Parents	20.2	21.2	-1.1	0.84	0.0	227
Peer Tutoring	23.1	16.9	6.3	0.24	0.2	227
Cooperation with Colleagues	6.2	5.3	0.8	0.79	0.1	227
Move Students to the Resource Room	5.1	12.3	-7.2	0.05	-0.6	227
Do Not Treat Underachieving Students Differently	0.9	1.8	-0.9	0.55	-0.4	227
Other	19.5	11.3	8.2	0.09	0.4	227
Methods to Support Students with Reading Difficulties						
Use Appropriate Material	41.0	41.0	0.0	1.00	0.0	227
Teach Basic Reading Skills	58.3	57.6	0.7	0.92	0.0	227
Use Small Group Sessions	26.2	24.6	1.5	0.79	0.0	227

Variable	Intervention (T)	Comparison (C)	Difference (T-C) *	P-Value	Effect Size	Number of Teachers
Use Individual Sessions	34.9	33.5	1.4	0.83	0.0	227
Request Assistance from Specialist	1.7	1.8	-0.1	0.97	0.0	227
Request Assistance from Parents	16.1	6.2	9.9*	0.019	0.6	227
Teacher does not Provide Support to Students with Reading Difficulties	2.7	1.8	0.9	0.66	0.2	227
Methods to Support Disobedient/Disruptive Students						
Call Parents	13.3	10.8	2.5	0.57	0.1	226
Talk to or Advice Students	39.6	43.7	-4.0	0.54	-0.1	226
Assign Additional Homework	30.5	26.9	3.6	0.55	0.1	226
Discipline Students	47.5	56.0	-8.5	0.20	-0.2	226
Use Corporal Punishment	0.7	0.9	-0.2	0.86	-0.2	226
Other	36.1	34.7	1.4	0.83	0.0	226
Parents Follow Up on Students' Homework	64.0	65.1	-1.2	0.85	0.0	222
Number of Schools	114	113				

Source: Ramp Impact Study - Endline Data 2018 Teacher Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted school-level group means (or percentages). The regressions use weights to account for school sampling probabilities. Effect sizes for dichotomous outcomes are Cox index. Sample sizes vary due item nonresponse.

*Difference in group means is statistically significant at the .05 level.

TABLE Q.8. G3 TEACHERS' STRATEGIES TO SUPPORT STUDENTS WITH ACADEMIC OR BEHAVIORAL DIFFICULTIES, AT ENDLINE

Variable	Intervention (T)	Comparison (C)	Difference (T-C) *	P-Value	Effect Size	Number of Teachers
Methods to Teach Underachieving Students						
Focus on Weak Students	81.8	83.3	-1.5	0.76	-0.1	233
Administer Daily Quizzes	14.6	17.3	-2.7	0.58	-0.1	233
Encouragement	34.0	28.1	5.9	0.34	0.2	233
Communicate with Parents	20.9	17.2	3.7	0.48	0.1	233
Peer Tutoring	24.2	16.0	8.2	0.12	0.3	233
Cooperation with Colleagues	2.6	2.6	0.0	1.00	0.0	233
Move Students to the Resource Room	8.6	13.5	-4.9	0.24	-0.3	233
Do Not Treat Underachieving Students Differently	0.0	1.7	-1.7	0.16	NA	233
Other	20.0	15.2	4.8	0.34	0.2	233
Methods to Support Students with Reading Difficulties						
Use Appropriate Material	44.5	43.3	1.2	0.86	0.0	234
Teach Basic Reading Skills	60.5	52.8	7.7	0.24	0.2	234
Use Small Group Sessions	28.8	21.5	7.3	0.20	0.2	234
Use Individual Sessions	30.3	32.9	-2.5	0.68	-0.1	234

Request Assistance from Specialist	2.6	2.5	0.1	0.97	0.0	234
Request Assistance from Parents	17.4	11.1	6.4	0.17	0.3	234
Teacher does not Provide Support to Students with Reading Difficulties	1.8	5.7	-3.9	0.11	-0.7	234
Methods to Support Disobedient/Disruptive Students						
Call Parents	15.9	16.4	-0.5	0.92	0.0	231
Talk to or Advice Students	46.9	40.7	6.1	0.35	0.2	231
Assign Additional Homework	21.0	34.6	-13.7*	0.021	-0.4	231
Discipline Students	54.8	49.8	5.0	0.45	0.1	231
Use Corporal Punishment	0.0	5.2	-5.2*	0.013	NA	231
Other	30.7	24.3	6.4	0.28	0.2	231
Parents Follow Up on Students' Homework	61.8	64.0	-2.1	0.74	-0.1	232
Number of Schools	116	117				

Source: Ramp Impact Study - Endline Data 2018 Teacher Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted school-level group means (or percentages). The regressions use weights to account for school sampling probabilities. Effect sizes for dichotomous outcomes are Cox index. Effect sizes are "NA" (not applicable) when the standard deviation for one or both groups equals zero. Sample sizes vary due item nonresponse.

*Difference in group means is statistically significant at the .05 level.

TEACHER SUPPORT

Teachers were also asked about the availability of support at the school, as well as the number of times they cooperated with colleagues to prepare lesson plans. These were outcomes that were not necessarily expected to be directly affected by RAMP, but that could affect teachers' job performance and students' learning.

There were no statistically significant differences between the groups, in either grade (see Tables Q.9. and Q.10.). The only exception was that G2 teachers in the intervention group were 5 percentage points more likely than comparison teachers to report having cooperated with colleagues once. There were no statistically significant differences in the endorsement of other response choices, such as never, weekly, or daily.

TABLE Q.9. G2 TEACHER SUPPORT AT ENDLINE

Variable	Intervention (T)	Comparison (C)	Difference (T-C) *	P-Value	Effect Size	Number of Teachers
If Teacher Needs Help						
Teacher Has No One to Reach Out to	7.1	8.1	-1.0	0.77	-0.1	227
Teacher Meets With Other Teachers	24.5	24.7	-0.2	0.97	0.0	227
Teacher Discusses With Other Teachers	61.3	55.9	5.4	0.41	0.1	227

Teacher Reaches Out to the Principal or Principal's Assistant	12.4	11.3	1.1	0.80	0.1	227
Teacher Reaches Out to the Education Support or Subject Coordinator	25.5	24.4	1.2	0.84	0.0	227
Teacher Reaches Out to Others	11.5	9.8	1.7	0.68	0.1	227
Teacher Never Needs Help	5.3	6.3	-1.1	0.73	-0.1	227
Number of Time Teacher Cooperated with Others to Prepare Lesson Plans						
Never	25.8	37.3	-11.5	0.06	-0.3	225
Once	6.2	0.9	5.3*	0.031	1.2	224
Every 2-3 Months	10.7	11.8	-1.2	0.79	-0.1	224
Monthly	16.0	16.1	-0.2	0.97	0.0	224
Once Every 2 Weeks	11.6	8.2	3.4	0.40	0.2	224
Weekly	19.2	17.1	2.1	0.68	0.1	224
Daily	10.5	8.2	2.4	0.54	0.2	224
Number Of Schools	114	113				

Source: Ramp Impact Study - Endline Data 2018 Teacher Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). All regressions include school sampling weights. Effect sizes for dichotomous outcomes are Cox index.

*Difference in group means is statistically significant at the .05 level.

TABLE Q.10. G3 TEACHER SUPPORT AT ENDLINE

Variable	Intervention (T)	Comparison (C)	Difference (T-C) *	P-Value	Effect Size	Number of Teachers
If Teacher Needs Help						
Teacher Has No One to Reach Out to	5.9	11.9	-6.0	0.11	-0.5	233
Teacher Meets With Other Teachers	28.1	22.3	5.8	0.31	0.2	233
Teacher Discusses With Other Teachers	61.3	55.9	5.4	0.41	0.1	227
Teacher Reaches Out to the Principal Or Principal's Assistant	13.0	14.8	-1.7	0.70	-0.1	233
Teacher Reaches Out to the Education Support Or Subject Coordinator	20.7	28.0	-7.3	0.20	-0.2	233
Teacher Reaches Out to Others	14.6	6.9	7.6	0.06	0.5	233
Teacher Never Needs Help	6.8	11.9	-5.1	0.18	-0.4	233
Number Of Time Teacher Cooperated With Others to Prepare Lesson Plans						
Never	28.0	27.7	0.4	0.95	0.0	231
Once	6.1	1.7	4.4	0.09	0.8	231
Every 2-3 Months	11.6	10.3	1.2	0.77	0.1	231
Monthly	16.8	20.6	-3.8	0.46	-0.2	231
Once Every 2 Weeks	7.9	6.9	1.0	0.77	0.1	231
Weekly	19.3	22.4	-3.1	0.57	-0.1	231
Daily	10.2	10.3	-0.1	0.97	0.0	231
Number Of Schools	116	117				

Source: Ramp Impact Study - Endline Data 2018 Teacher Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). All regressions include school sampling weights. Effect sizes for dichotomous outcomes are Cox index. Sample sizes vary due item nonresponse.

*Difference in group means is statistically significant at the .05 level.

TEACHER SUPERVISION

Finally, teachers were asked about the frequency and nature of supervision they received from the principal or assistant principal at their school, as well as the level supervisor. The RAMP training was not expected to dramatically change supervision received by teachers, but it could have shifted as other processes were changing in the school or due to broader policy changes related to the MoE.

The results showed no statistically significant differences between the groups, except for a difference of 4 percentage points in the proportion of G2 teachers who reported receiving visits by the level supervisor once every two weeks. This difference may offset the higher proportion of teachers in the intervention group reporting never receiving visits from the level supervisor. This difference, however, was not statistically significant (see Table Q.11).

TABLE Q.11. G2 TEACHER SUPERVISION AT ENDLINE

Variable	Intervention (T)	Comparison (C)	Difference (T-C) *	P-Value	Effect Size	Number of Teachers
Principal or Assistant Principal Observed Teacher's Classroom in Current School Year						
Never	2.7	0.9	1.8	0.33	0.7	227
Once	8.0	9.8	-1.8	0.63	-0.1	227
Every 2-3 Months	26.0	28.3	-2.2	0.71	-0.1	227
Monthly	24.8	23.2	1.6	0.78	0.1	227
Once Every 2 Weeks "Twice a Month or More"	16.4	19.5	-3.1	0.54	-0.1	227
Weekly	15.9	13.2	2.8	0.56	0.1	227
Daily	6.3	5.2	1.1	0.73	0.1	227
Level Supervisor Visited the School in Current School Year						
Never	17.5	10.3	7.1	0.12	0.4	226
Once	27.4	34.4	-7.0	0.26	-0.2	226
Every 2-3 Months	20.6	16.0	4.6	0.38	0.2	226
Monthly	7.1	9.0	-1.9	0.60	-0.2	226
Once Every 2 Weeks	3.6	0.0	3.6*	0.044	NA	226
Supervisor Offered Advice and Tips on Teaching Methods	84.9	91.0	-6.1	0.23	-0.4	165
Supervisor Offered Advice and Tips on Assessment of Student Performance	80.3	84.4	-4.1	0.49	-0.2	163
Number of Schools	114	113				

Source: Ramp Impact Study - Endline Data 2018 Teacher Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). All regressions include school sampling weights. Effect sizes for dichotomous outcomes are Cox index. Effect sizes are "NA" (not applicable) when the standard deviation for one or both groups equals zero. Sample sizes vary due item nonresponse.* Difference in group means is statistically significant at the .05 level.

TABLE Q.12. G3 TEACHER SUPERVISION AT ENDLINE

Variable	Intervention (T)	Comparison (C)	Difference (T-C) *	P-Value	Effect Size	Number of Teachers
Principal or Assistant Principal Observed Teacher's Classroom in Current School Year						
Never	2.6	1.7	0.9	0.65	0.3	227
Once	10.4	8.5	1.9	0.63	0.1	227
Every 2-3 Months	24.7	23.1	1.7	0.77	0.1	227
Monthly	20.8	27.5	-6.7	0.24	-0.2	227
Once Every 2 Weeks "Twice a Month Or More"	12.1	17.4	-5.2	0.26	-0.3	227
Weekly	24.2	15.2	9.0	0.08	0.4	227
Daily	5.2	6.7	-1.5	0.64	-0.2	227
Level Supervisor Visited the School in Current School Year						
Never	14.8	9.1	5.7	0.18	0.3	226
Once	29.8	39.1	-9.3	0.14	-0.2	226

Variable	Intervention (T)	Comparison (C)	Difference (T-C) *	P-Value	Effect Size	Number of Teachers
Every 2-3 Months	15.9	10.0	5.9	0.19	0.3	226
Monthly	5.4	4.3	1.1	0.71	0.1	226
Once Every 2 Weeks	2.7	0.0	2.7	0.08	NA	226
Supervisor Offered Advice and Tips on Teaching Methods	87.2	83.6	3.6	0.53	0.2	165
Supervisor Offered Advice and Tips on Assessment of Student Performance	74.4	72.6	1.8	0.80	0.1	163
Number of Schools	116	117				

Source: Ramp Impact Study - Endline Data 2018 Teacher Questionnaire

Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). All regressions include school sampling weights. Effect sizes for dichotomous outcomes are Cox index. Effect sizes are "NA" (not applicable) when the standard deviation for one or both groups equals zero. Sample sizes vary due item nonresponse.

*Difference in group means is statistically significant at the .05 level.

ANNEX R. MIDLINE IMPACTS OF RAMP ON STUDENTS' READING AND MATH

This annex includes key midline math and reading results by grade, as well as midline results from subgroup analysis by student gender and number of school shifts. Results with reading and math scores are presented first, followed by the analysis of zero-scores. A full discussion of the results is provided in the RAMP Impact Evaluation Midline Report (MESP, 2017).

RESULTS BY GRADE

MIDLINE GRADE 1 READING RESULTS

G1 students were assessed on five foundational skills required to master fluent reading and comprehension. RAMP had impacts on only two out of seven reading subtasks. G1 intervention students scored significantly better than comparison students did on syllable segmentation (See Table R.1). While intervention students sounded out 5.7 words, comparison students sounded 5.0 out of 10 words correctly. RAMP also had a small significant impact on students' ability to read a short passage; intervention students read approximately one more word per minute than comparison students did. RAMP did not have detectable impacts on G1 students' knowledge of early print concepts, ability to segment words into phonemes, knowledge of letter-sound correspondence, reading vocabulary, or reading comprehension.

TABLE R.1. READING PERFORMANCE SCORES FOR GRADE 1 STUDENTS AT MIDLINE

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C *	Effect Size	Number of Students
Orientation to Print (Out of 5)	2.9	3.0	-0.05	-0.03	2172
			0.53		
Phoneme Isolation (Out of 10)	4.3	4.1	0.14	0.05	2172
			0.32		
Syllable Segmentation (Out of 10)	5.7	5.0	0.69*	0.17	2172
			0.01		
Letter Sound Knowledge (Out of 100, Prorated Score)	30.6	30.5	0.10	0.01	2172
			0.93		
Reading Vocabulary (Out of 10)	6.7	6.5	0.23	0.08	2172
			0.16		
Passage Reading (Out of 52, Prorated Score)	6.3	5.1	1.16*	0.14	2172
			0.02		
Reading Comprehension (Out of 6)	0.9	0.8	0.09	0.08	2172
			0.21		

Source: RAMP Impact Study – Midline Data 2017 Student Assessments

Note: The table presents ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors and student propensity score weights. Columns T and C present group means for first grade students in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the different in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown below the corresponding difference in means in the T minus C column. Effect sizes are the standardized mean difference below the corresponding difference in means in the T minus C column. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and denominator is the pooled standard deviation. Sample size refers to the student-level analytical sample used for difference testing.

* Difference in group means is statistically significant at the .05 level.

MIDLINE GRADE 2 READING RESULTS

RAMP had impacts on only one out of six EGRA subtasks assessed in G2. Specifically, RAMP had a statistically significant impact on G2 students' ability to segment words into syllables, equivalent to 0.31standard deviations. Intervention students segmented 6.4 words whereas comparison students segmented 5.3 out of 10 words (see Table R.2). RAMP did not have other detectable impacts on the reading skills of G2 students.

TABLE R.2. READING PERFORMANCE SCORES FOR GRADE 2 STUDENTS AT MIDLINE

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C *	Effect Size	Number of Students
Syllable segmentation (out of 10)	6.4	5.3	1.05*	0.31	1741
			0.00		
Letter sound knowledge (out of 100, prorated score)	35.6	37.7	-2.07	-0.10	1741
			0.15		
Non-word decoding (out of 50, prorated score)	9.7	9.2	0.46	0.06	1741
			0.36		
Reading vocabulary (out of 10)	8.3	8.4	-0.12	-0.04	1741
			0.31		
Passage reading (out of 52, prorated score)	19.2	19.2	0.02	0.00	1741
			0.97		
Reading comprehension (out of 6)	2.2	2.2	-0.01	-0.01	1741
			0.90		

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors and student propensity score weights. Columns T and C present group means for first grade students in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. Sample size refers to the student-level analytical sample used for difference testing.

*Difference in group means is statistically significant at the .05 level.

MIDLINE GRADE 1 MATH RESULTS

G1 students were assessed on their ability to (1) count from zero to sixty, (2) count objects, (3) identify numerals (4) compare numerical magnitudes, (5) detect number patterns, and (6) solve grade-level addition problems. The data revealed minimal differences between intervention and comparison students' foundational math skills (see Table R.3).

TABLE R.3. MATH PERFORMANCE SCORES FOR GRADE 1 STUDENTS AT MIDLINE

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C *	Effect Size	Number of Students
Counting Numbers (Out of 60)	31.6	32.8	-1.25*	-0.12	2172
			0.02		
Counting Objects (Out of 10)	9.7	9.7	-0.02	-0.03	2172
			0.52		
Number Identification (Out of 20 , Prorated Score)	23.2	23.0	0.18	0.02	2172
			0.70		
Number Discrimination (Out of 10)	8.3	8.4	-0.10	-0.04	2172
			0.43		
Missing Numbers (Out of 10)	4.6	4.4	0.18	0.06	2172
			0.28		
Addition Facts - L1 (Out of 20, Prorated Score)	5.4	4.7	0.63	0.10	2172
			0.12		

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors and student propensity score weights. Columns T and C present group means for first grade students in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. Sample size refers to the student-level analytical sample used for difference testing.

*Difference in group means is statistically significant at the .05 level.

MIDLINE GRADE 2 MATH RESULTS

G2 students were tested on their ability to identify numerals and compare numerical magnitudes, detect number patterns, and solve grade-level addition and subtraction problems. G2 intervention and comparison students performed similarly and the data revealed no detectable impacts in math (see Table R.4).

TABLE R.4. MATH PERFORMANCE SCORES FOR GRADE 2 STUDENTS AT MIDLINE

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T- C *	Effect Size	Number of Students
Number Identification (Out of 20, Prorated Score)	32.4	31.9	0.53	0.04	1741
			0.42		
Number Discrimination (Out of 10)	8.1	8.2	-0.05	-0.02	1741
			0.58		
Missing Numbers (Out of 10)	7.6	7.7	-0.08	-0.03	1741
			0.46		
Addition Facts - L1 (Out of 20, Prorated Score)	11.2	11.6	-0.39	-0.07	1741
			0.15		
Addition Facts - L2 (Out of 5)	2.8	2.8	0.04	0.02	1635
			0.73		
Subtraction Facts - L1 (Out of 20, Prorated Score)	5.4	5.8	-0.39	-0.10	1741
			0.11		
Subtraction Facts - L2 (Out of 5)	2.6	2.6	-0.01	-0.01	1553
			0.91		

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors and student propensity score weights. Columns T and C present group means for first grade students in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown in parentheses. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. Sample size refers to the analytical student-level sample used for difference testing.

*Difference in group means is statistically significant at the .05 level.

RESULTS BY GENDER

There were no differences in RAMP impacts between boys and girls, except for impacts in reading comprehension for G1 students and number identification for G2 students (see the last column in Tables R.5 - R.8).

TABLE R.5. READING PERFORMANCE SCORES FOR GRADE 1 STUDENTS BY GENDER AT MIDLINE

	Intervention Group		Comparison Group		Boys in Intervention vs. Comparison	Girls in Intervention vs. Comparison	Difference in RAMP Impacts by Gender (Coefficient and P-Value)
Variable (Total # of Items)	Boys (A)	Girls (B)	Boys (C)	Girls (D)	A minus C	B minus D	
Orientation to Print (Max 5 Items)	2.9	2.9	2.9	3.0	0.00	-0.09	-0.09
	(525)	(568)	(553)	(526)	0.99	0.42	0.56
Phoneme Isolation (Max 10 Items)	3.9	4.6	3.8	4.5	0.15	0.13	-0.02
	(525)	(568)	(553)	(526)	0.46	0.47	0.94
Syllable Segmentation (Max 10 Items)	5.5	6.0	4.6	5.5	0.94*	0.47	-0.47
	(525)	(568)	(553)	(526)	0.01	0.16	0.25
Letter Sound Knowledge (Max 100 Items)	28.5	32.7	28.2	32.8	0.36	-0.14	-0.51
	(525)	(568)	(553)	(526)	0.80	0.91	0.73
Reading Vocabulary (Max 10 Items)	6.5	7.0	6.2	6.8	0.28	0.18	-0.09
	(525)	(568)	(553)	(526)	0.19	0.35	0.70
Passage Reading (Prorated Score)	5.0	7.5	3.7	6.4	1.29	1.04	-0.25
	(525)	(568)	(553)	(526)	0.05	0.09	0.77
Reading Comprehension (Max 6 Items)	0.8	1.0	0.6	1.0	0.23*	-0.03	-0.26*
	(525)	(568)	(553)	(526)	0.02	0.72	0.03

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. All means are adjusted for baseline differences between groups. Sample sizes (which are found below their corresponding average scores in parentheses) refer to the number of students in each subgroup. P-values are shown below their corresponding coefficient.

*Difference in group means is statistically significant at the .05 level.

TABLE R.6. READING PERFORMANCE SCORES FOR GRADE 2 STUDENTS BY GENDER AT MIDLINE

	Intervention Group		Comparison Group		Boys in Intervention vs. Comparison	Girls in Intervention vs. Comparison	Difference in RAMP Impacts by Gender (Coefficient and P-Value)
Variable (Total # of Items)	Boys (A)	Girls (B)	Boys (C)	Girls (D)	A minus C	B minus D	

Syllable Segmentation (Max 10 Items)	6.2	6.6	4.9	5.7	1.24*	0.91*	-0.33
	(427)	(503)	(413)	(399)	0.00	0.00	0.39
Letter Sound Knowledge (Max 100 Items)	35.7	35.8	35.7	39.5	-0.05	-3.66	-3.61
	(427)	(503)	(413)	(399)	0.97	0.07	0.10
Non-Word Decoding (Max 50 Items)	9.1	10.3	8.0	10.3	1.09	-0.03	-1.12
	(427)	(503)	(413)	(399)	0.06	0.96	0.12
Reading Vocabulary (Max 10 Items)	8.0	8.6	8.1	8.8	-0.07	-0.15	-0.08
	(427)	(503)	(413)	(399)	0.74	0.29	0.74
Passage Reading (Prorated Score)	17.3	21.0	17.1	21.1	0.13	-0.06	-0.19
	(427)	(503)	(413)	(399)	0.85	0.93	0.83
Reading Comprehension (Max 6 Items)	1.9	2.4	1.8	2.5	0.06	-0.07	-0.12
	(427)	(503)	(413)	(399)	0.63	0.63	0.49

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. All means are adjusted for baseline differences between groups. Sample sizes, which are found below their corresponding average scores in parentheses) refer to the number of students in each subgroup. P-values are shown below their corresponding coefficient.

*Difference in group means is statistically significant at the .05 level.

TABLE R.7. MATH PERFORMANCE SCORES FOR GRADE 1 STUDENTS BY GENDER AT MIDLINE

	Intervention Group		Comparison Group		Boys in Intervention vs. Comparison	Girls in Intervention vs. Comparison	Difference in RAMP Impacts by Gender (Coefficient and P-Value)
Variable (Total # of Items)	Boys (A)	Girls (B)	Boys (C)	Girls (D)	A minus C	B minus D	
Counting Numbers (Out of 60 Items)	30.5	32.7	32.0	33.7	-1.50	-1.03	0.47
	(525)	(568)	(553)	(526)	0.08	0.12	0.66
Counting Objects (Out of 10 Items)	9.6	9.7	9.7	9.8	-0.02	-0.03	-0.02
	(525)	(568)	(553)	(526)	0.80	0.57	0.86
Number Identification (Out of 20 Items)	22.7	23.6	22.9	23.1	-0.18	0.50	0.68
	(525)	(568)	(553)	(526)	0.76	0.40	0.33

	Intervention Group		Comparison Group		Boys in Intervention vs. Comparison	Girls in Intervention vs. Comparison	Difference in RAMP Impacts by Gender (Coefficient and P-Value)
Variable (Total # of Items)	Boys (A)	Girls (B)	Boys (C)	Girls (D)	A minus C	B minus D	
Number Discrimination (Out of 10 Items)	8.4	8.3	8.4	8.5	0.01	-0.19	-0.20
	(525)	(568)	(553)	(526)	0.96	0.25	0.41
Missing Numbers (Out of 10 Items)	4.4	4.7	4.1	4.6	0.33	0.06	-0.27
	(525)	(568)	(553)	(526)	0.14	0.79	0.32
Addition Facts - L1 (Out of 20 Items)	5.8	4.9	5.4	4.0	0.37	0.86*	0.50
	(525)	(568)	(553)	(526)	0.53	0.05	0.43

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. All means are adjusted for baseline differences between groups. Sample sizes (which are found below their corresponding average scores in parentheses) refer to the number of students in each subgroup. P-values are shown below their corresponding coefficient.

*Difference in group means is statistically significant at the .05 level.

TABLE R.8. MATH PERFORMANCE SCORES FOR GRADE 2 STUDENTS BY GENDER AT MIDLINE

Variable (Total # of Items)	Intervention Group		Comparison Group		Boys in Intervention vs. Comparison	Girls in Intervention vs. Comparison	Difference in RAMP Impacts by Gender (Coefficient and P-Value)
	Boys (A)	Girls (B)	Boys (C)	Girls (D)	A minus C	B minus D	
Number Identification (Prorated Score)	33.1	32.0	30.9	32.8	2.25*	-0.82	-3.07*
	(427)	(503)	(413)	(399)	0.02	0.32	0.01
Number Discrimination (Max 10)	8.1	8.1	8.2	8.2	-0.09	-0.02	0.07
	(427	(503)	(413)	(399)	0.53	0.86	0.69
Missing Numbers (Prorated Score)	7.4	7.8	7.6	7.8	-0.15	-0.01	0.14
	(427	(503)	(413)	(399)	0.31	0.92	0.50
Addition Facts - L1 (Prorated Score)	11.5	11.0	11.6	11.6	-0.09	-0.63	-0.53
	(427	(503)	(413)	(399)	0.82	0.08	0.32
Addition Facts - L2 (Max 5 Items)	2.7	3.0	2.7	2.9	0.01	0.06	0.05
	(404	(480)	(376)	(376)	0.94	0.68	0.78
Subtraction Facts - L1 (Prorated Score)	5.8	5.0	6.0	5.6	-0.22	-0.53	-0.31
	(427	(503)	(413)	(399)	0.51	0.08	0.44
Subtraction Facts - L2 (Max 5 Items)	2.5	2.6	2.5	2.7	-0.01	-0.01	0.00
	(375	(463)	(360)	(355)	0.94	0.95	0.99

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. All means are adjusted for baseline differences between groups. Sample sizes (which are found below their corresponding average scores in parentheses) refer to the number of students in each subgroup. P-values are shown below their corresponding coefficient.

*Difference in group means is statistically significant at the .05 level.

RESULTS BY NUMBER OF SCHOOL SHIFTS

There were significant differences in RAMP impacts between students in single- versus double-shift schools for G1 reading vocabulary (see Table R.9) and G2 invented word decoding and reading fluency (see Table R.10). However, the results did not suggest a clear pattern indicating an advantage for students in single- or double-shift schools. These differ from endline results,

where RAMP had a tendency to have a negative impact on students in single-shift schools and a positive impact on students in double-shift schools.

TABLE R.9. READING PERFORMANCE SCORES FOR GRADE 1 STUDENTS BY NUMBER OF SCHOOL SHIFTS AT MIDLINE

	Intervention Group		Comparison Group		Single-Shift in Intervention vs. Comparison	Double-Shift in Intervention vs. Comparison	Difference in RAMP Impacts by Number of Shifts (Coefficient and P-Value)
Variable (Total # of Items)	Single (A)	Double (B)	Single (C)	Double (D)	A minus C	B minus D	
Orientation to Print (Max 5 Items)	2.8	3.1	2.9	3.0	-0.09	0.03	0.11
	(884)	(209)	(938)	(141)	0.32	0.84	0.49
Phoneme Isolation (Max 10 Items)	4.2	4.6	4.1	4.1	0.04	0.45	0.42
	(884)	(209)	(938)	(141)	0.81	0.23	0.31
Syllable Segmentation (Max 10 Items)	5.6	6.0	5.0	5.3	0.66*	0.70	0.04
	(884)	(209)	(938)	(141)	0.03	0.22	0.95
Letter Sound Knowledge (Max 100 Items)	30.9	28.7	31.5	25.7	-0.62	3.03	3.65
	(884)	(209)	(938)	(141)	0.59	0.27	0.22
Reading Vocabulary (Max 10 Items)	6.7	6.8	6.6	5.9	0.05	0.89*	0.84*
	(884)	(209)	(938)	(141)	0.77	0.02	0.05
Passage Reading (Prorated Score)	6.4	5.9	5.4	3.6	0.93	2.32*	1.39
	(884)	(209)	(938)	(141)	0.10	0.00	0.13
Reading Comprehension (Max 6 Items)	0.9	0.9	0.8	0.7	0.07	0.17	0.10
	(884)	(209)	(938)	(141)	0.38	0.17	0.50

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. All means are adjusted for baseline differences between groups. . Sample sizes (which are found below their corresponding average scores in parentheses) refer to the number of students in each subgroup. P-values are shown below their corresponding coefficient.

*Difference in group means is statistically significant at the .05 level.

TABLE R.10. READING PERFORMANCE SCORES FOR GRADE 2 STUDENTS BY NUMBER OF SCHOOL SHIFTS AT MIDLINE

	Intervention Group		Comparison Group		Single-Shift in Intervention vs. Comparison	Double-Shift in Intervention vs. Comparison	Difference in Ramp Impacts by Number of Shifts (Coefficient and P-Value)
Variable (Total # of Items)	Single (A)	Double (B)	Single (C)	Double (D)	A minus C	B minus D	
Syllable Segmentation (Max 10 Items)	6.3	6.7	5.3	5.5	0.98*	1.13*	0.15
	(748)	(182)	(714)	(98)	0.00	0.02	0.78
Letter Sound Knowledge (Max 100 Items)	36.9	31.0	38.2	34.8	-1.31	-3.82	-2.51
	(748)	(182)	(714)	(98)	0.39	0.25	0.49
Non-Word Decoding (Max 50 Items)	10.1	8.1	9.1	9.7	0.95	-1.55	-2.50*
	(748)	(182)	(714)	(98)	0.09	0.12	0.03
Reading Vocabulary (Max 10 Items)	8.4	8.1	8.5	8.4	-0.09	-0.28	-0.19
	(748)	(182)	(714)	(98)	0.46	0.29	0.53
Passage Reading (Prorated Score)	19.6	17.2	19.1	19.6	0.54	-2.42*	-2.96*
	(748)	(182)	(714)	(98)	0.37	0.00	0.00
Reading Comprehension (Max 6 Items)	2.1	2.1	2.1	2.2	-0.01	-0.12	-0.11
	(748)	(182)	(714)	(98)	0.96	0.49	0.58

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. All means are adjusted for baseline differences between groups. Sample sizes (which are found below their corresponding average scores in parentheses) refer to the number of students in each subgroup. P-values are shown below their corresponding coefficient.

*Difference in group means is statistically significant at the .05 level.

TABLE R.11. MATH PERFORMANCE SCORES FOR GRADE 1 STUDENTS BY NUMBER OF SCHOOL SHIFTS AT MIDLINE

	Intervention Group		Comparison Group		Single-Shift in Intervention vs. Comparison	Double-Shift in Intervention vs. Comparison	Difference in RAMP Impacts by Number of Shifts (Coefficient and P-Value)
Variable (Total # of Items)	Single (A)	Double (B)	Single (C)	Double (D)	A minus C	B minus D	
Counting Numbers (Out of 60 Items)	31.3	32.2	33.0	32.2	-1.62*	-0.06	1.56
	(884)	(209)	(938)	(141)	0.01	0.96	0.20
Counting Objects (Out of 10 Items)	9.7	9.7	9.7	9.8	-0.01	-0.11*	-0.11
	(884)	(209)	(938)	(141)	0.91	0.03	0.13
Number Identification (Out of 20 Items)	23.1	23.3	23.0	22.9	0.09	0.37	0.28
	(884)	(209)	(938)	(141)	0.87	0.67	0.78
Number Discrimination (Out of 10 Items)	8.3	8.2	8.4	8.4	-0.08	-0.18	-0.09
	(884)	(209)	(938)	(141)	0.52	0.54	0.77
Missing Numbers (Out of 10 Items)	4.5	4.8	4.3	4.5	0.11	0.32	0.21
	(884)	(209)	(938)	(141)	0.54	0.30	0.57
Addition Facts - L1 (Out of 20 Items)	5.1	6.0	4.5	5.7	0.59	0.28	-0.31
	(884)	(209)	(938)	(141)	0.11	0.77	0.77

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. All means are adjusted for baseline differences between groups. . Sample sizes (which are found below their corresponding average scores in parentheses) refer to the number of students in each subgroup. P-values are shown below their corresponding coefficient.

*Difference in group means is statistically significant at the .05 level.

TABLE R.12. MATH PERFORMANCE SCORES FOR GRADE 2 STUDENTS BY NUMBER OF SCHOOL SHIFTS AT MIDLINE

	Intervention Group		Comparison Group		Single-Shift in Intervention vs. Comparison	Double-Shift in Intervention vs. Comparison	Difference in RAMP Impacts by Number of Shifts (Coefficient and P-Value)
Variable (Total # of Items)	Single (A)	Double (B)	Single (C)	Double (D)	A minus C	B minus D	
Number Identification (Prorated Score)	32.1	33.1	31.7	32.8	0.38	0.30	-0.08
	(748)	(182)	(714)	(98)	0.57	0.84	0.96
Number Discrimination (Max 10 Items)	8.1	8.1	8.2	8.0	-0.08	0.07	0.14
	(748)	(182)	(714)	(98)	0.46	0.79	0.59
Missing Numbers (Prorated Score)	7.5	7.8	7.7	7.6	-0.17	0.18	0.36
	(748)	(182)	(714)	(98)	0.11	0.52	0.24
Addition Facts - L1 (Prorated Score)	11.1	11.6	11.6	11.5	-0.55	0.05	0.60
	(748)	(182)	(714)	(98)	0.06	0.94	0.37
Addition Facts - L2 (Max 5 Items)	2.8	2.9	2.8	2.8	0.01	0.07	0.06
	(709)	(175)	(666)	(86)	0.90	0.80	0.86
Subtraction Facts - L1 (Prorated Score)	5.3	5.7	5.7	6.2	-0.44	-0.54	-0.10
	(748)	(182)	(714)	(98)	0.09	0.39	0.88
Subtraction Facts - L2 (Max 5 Items)	2.5	2.5	2.5	2.6	0.00	-0.10	-0.09
	(667)	(171	(634)	(81)	0.97	0.75	0.77

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. All means are adjusted for baseline differences between groups. . Sample sizes (which are found below their corresponding average scores in parentheses) refer to the number of students in each subgroup. P-values are shown below their corresponding coefficient.

*Difference in group means is statistically significant at the .05 level.

MIDLINE ZERO-SCORES

This section presents key results from the midline analysis of zero-scores in reading and math subtasks. A full discussion of the results is provided in the RAMP Impact Evaluation Midline Report (MESP, 2017).

There was a small significant difference between the intervention and comparison groups in the percent of G1 and G2 students who obtained a score equal to zero in the syllable segmentation subtask at midline (see Table R.13). While 29 percent of G1 students in RAMP schools obtained zero scores in this task, 34 percent of students in comparison schools did. Similarly, while 14 percent of G2 intervention students scored zero in syllable segmentation, 19 percent of students in the comparison group did.

TABLE R.13. READING PERFORMANCE ZERO SCORES FOR GRADE 1 STUDENTS AT MIDLINE

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C*)	Effect Size	Number of Students
Orientation to Print Zero Score	3.8	3.3	0.45	0.03	2172
			0.58		
Phoneme Isolation Zero Score	13.0	10.7	2.31	0.07	2172
			0.11		
Syllable Segmentation Zero Score	28.6	34.5	-5.85*	-0.13	2172
			0.05		
Letter Sound Knowledge Zero Score	11.7	10.5	1.24	0.04	2172
			0.52		
Reading Vocabulary Zero Score	7.3	9.7	-2.48	-0.10	2172
			0.11		
Passage Reading Zero Score	45.0	47.1	-2.11	-0.04	2172
			0.47		
Reading Comprehension Zero Score	52.6	55.4	-2.84	-0.06	2172
			0.37		

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. Columns T and C present group means for first grade students in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown below the corresponding coefficient. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. All means are adjusted for baseline differences between groups. Sample size refers to the analytical sample used for difference testing. All models used equated scores.

*Difference in group means is statistically significant at the .05 level.

TABLE R.14. READING PERFORMANCE ZERO SCORES FOR GRADE 2 STUDENTS AT MIDLINE

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C)*	Effect Size	Number of Students
Syllable Segmentation Zero Score	13.7	18.5	-4.77*	-0.13	1741
			0.04		
Letter Sound Knowledge Zero Score	10.1	9.1	1.01	0.03	1741
			0.62		
Non-Word Decoding Zero Score	19.1	19.9	-0.89	-0.02	1741
			0.71		
Reading Vocabulary Zero Score	2.8	2.5	0.25	0.01	1741
			0.79		
Passage Reading Zero Score	7.6	8.0	-0.41	-0.01	1741
			0.79		
Reading Comprehension Zero Score	36.1	34.0	2.09	0.04	1741
			0.42		

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. Columns T and C present group means for second grade students in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown below the corresponding coefficient. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. All means are adjusted for baseline differences between groups. Sample size refers to the analytical sample used for difference testing. All models used equated scores.

*Difference in group means is statistically significant at the .05 level.

TABLE R.15. MATH PERFORMANCE ZERO SCORES FOR GRADE 1 STUDENTS AT MIDLINE

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C)*	Effect Size	Number of Students
Number Identification Zero Score	0.0	0.1	-0.06	-0.04	2172
			0.23		
Number Discrimination Zero Score	5.5	4.2	1.27	0.07	2172
			0.15		
Missing Numbers Zero Score	14.7	14.2	0.41	0.01	2172
			0.82		
Addition Facts - L1 Zero Score	35.8	36.1	-0.28	-0.01	2172
			0.91		

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. Columns T and C present group means for first grade students in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown below the corresponding coefficient. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. All means are adjusted for baseline differences between groups. Sample size refers to the analytical sample used for difference testing. All models used equated scores.

*Difference in group means is statistically significant at the .05 level.

TABLE R.16. MATH PERFORMANCE ZERO SCORES FOR GRADE 2 STUDENTS AT MIDLINE

Variable (Total # of Items)	Intervention (T)	Comparison (C)	Impact (T-C)*	Effect Size	Number of Students
Number Discrimination Zero Score	1.2	0.9	0.30	0.02	1741
			0.62		
Missing Numbers Zero Score	4.5	3.1	1.41	0.06	1741
			0.12		
Addition Facts - L1 Zero Score	6.7	5.1	1.58	0.05	1741
			0.19		
Addition Facts - L2 Zero Score	13.8	12.0	1.85	0.05	1635
			0.32		
Subtraction Facts - L1 Zero Score	17.5	13.7	3.77	0.09	1741
			0.06		
Subtraction Facts - L2 Zero Score	18.9	16.5	2.38	0.06	1553
			0.27		

Source: RAMP Impact Study - Midline Data 2017 Student Assessments

Note: The table presents weighted ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. Columns T and C present group means for second grade students in the intervention group (Cohort 2) and the comparison group (Cohort 3), respectively. T minus C is the difference in the means between the intervention and comparison groups at midline. P-values are from tests of differences between group means and are shown below the corresponding coefficient. Effect sizes are the standardized mean difference between the intervention and comparison groups. The numerator is the unstandardized regression coefficient and the denominator is the pooled standard deviation. All means are adjusted for baseline differences between groups. Sample size refers to the analytical sample used for difference testing. All models used equated scores.

*Difference in group means is statistically significant at the .05 level.

ANNEX S. MEMORANDUM: "COMPARISON BRIEF OF STUDIES: RAMP IMPLEMENTATION VS. RAMP IMPACT EVALUATION"

Date: December 2017 (accepted August 2018)
Prepared by: MESP RAMP Impact Evaluation Team
Prepared for: USAID Jordan
Purpose: The RAMP impact evaluation team prepared this document to compare findings from the IE with RTI's midline survey results per the request of the USAID Jordan Mission.

TABLE S.1. COMPARISON OF RTI AND MSI RESULTS

RAMP Implementation (RTI)	RAMP Impact Evaluation (MSI)
Study Design	
• Performance Evaluation • No counterfactual • Cross-Sectional Studies	• Quasi-Experimental Impact Evaluation • Intervention and Comparison Groups • Counterfactual allows attribution to investigate change over time
Sampling	
• **Different samples for the two studies** • 2014: from **156** schools (national representative), purposively selected 43 schools for pilot • Unclear if remaining 110 schools nationally representative • 2016: **240** schools • Small schools excluded Cannot draw conclusions from different samples	• **Following same schools over time** • Randomly selected schools based on RAMP rollout (120 intervention, 120 comparison schools for total of **240** schools) • Statistically representative of Cohort 2 and Cohort 3 • Propensity Score Matching at 2 levels (schools and students) to ensure baseline equivalency • Following same students, schools, teachers over time (G3) Thus, differences can confidently be attributed to intervention
Tools/Instruments	
• Used same tool for G2 and G3 students**, which doesn't acknowledge differences in curriculum and grade level standards** • **2014 study: based on 2012 curriculum; 2016/2017 study: revised, unclear if based on 2016 curriculum**	• **Separate tools for G2 and G3** • Both baseline and midline tools based on 2016 curriculum for each grade • Tools calibrated to standards at beginning and end of school year • Developed in a collaborative process with MoE
Inferences: What does the approach tell us?	
• Descriptive info on students' outcomes • Identification of foundational skills students struggle with; Determine focal areas to help students meet benchmarks • If the sample of 110 schools is nationally representative, study could provide data on trends, but not causal changes	• Descriptive info on students in Cohorts 2 & 3 • Allow us to draw causal conclusions about impacts and, with some confidence, attribute changes to intervention • Student learning in reading and math • Teacher practices in G3 teachers

• National level results include students who have not been exposed to RAMP	

COMPARISON BRIEF OF RAMP STUDIES:

RAMP IMPLEMENTATION VS. RAMP IMPACT EVALUATION

In the second half of 2017, RTI and MSI submitted reports aimed at establishing whether the early grade reading and math program (RAMP) had an *impact* on students learning outcomes. Per the request of USAID, to clarify why findings and conclusions differ between the two studies, this document (1) describes key requirements to estimate causal impacts of an intervention like RAMP, (2) identifies key differences in the study designs that RTI and MSI used, (3) notes differences in instrumentation, and (4) explains what inferences can (and cannot) be drawn from each study.

WHAT IS REQUIRED TO ESTABLISH WHETHER RAMP *HAD AN IMPACT* ON STUDENT OUTCOMES?

Both RTI's and MSI's studies claim to establish whether RAMP had an *impact* on students' reading and math outcomes. Assessing the differences between the two studies necessitates understanding the requirements of an *impact evaluation*; that is, an evaluation that allows answering causal questions about an intervention's impacts.

USAID's Evaluation Policy (2011, updated in 2016) defines impact evaluations as those that "measure change in a development outcome that is attributable to a defined intervention" (p. 3). To attribute changes in student outcomes to RAMP or to any other intervention, impact evaluations need a *counterfactual*, that is, a measure of "what would have happened in the absence of RAMP" (USAID, 2013). "Impact evaluations in which comparisons are made between beneficiaries that are randomly assigned to either a treatment or control group provide the strongest evidence of a relationship between the intervention and the outcome measured" (USAID 2011, p. 3).

The key challenge of any impact evaluation is constructing a counterfactual that allows a comparison of students with and without an intervention so that differences in outcomes can be compared. However, we cannot observe the same students (or group of students) simultaneously in the two conditions (with and without RAMP). Thus, a credible impact evaluation rests on its ability to find a plausible approximation for comparison. The comparison group must be rigorously assessed to ensure that it is an appropriate counterfactual or comparison to the intervention group. Without a credible counterfactual, it is not appropriate to attribute changes in outcomes to the program.

According to USAID's Evaluation Policy (2011), *performance evaluations* differ from impact evaluations in that they "often incorporate before-after comparisons but lack a rigorously defined counterfactual" (p. 3). As such, *they can only demonstrate whether change has occurred, but cannot establish what actually caused the observed change* (USAID, 2013; emphasis added). Without a rigorous approximation to the counterfactual, there are a myriad of alternative explanations for the observed differences. Therefore, performance evaluations cannot claim that the intervention caused observed changes.

Based on the above definitions, *RTI's study is a performance evaluation*, whereas *MSI's study is an impact evaluation*. The remaining of this document centers on describing the counterfactual of each study and its implications for the interpretation of results.

RTI'S "EARLY GRADE READING AND MATHEMATICS INITIATIVE: MIDLINE SURVEY REPORT"

RTI aims to estimate the counterfactual with data collected from G2 and G3 students in 110 schools in 2014, as part of the Intervention Pilot Research Activity implemented in 2013-2014. RTI selected 110 schools out of a nationally representative sample of 156 schools. From this sample, 43 schools were *purposely* selected to receive the Pilot treatment. The report does not specify whether the remaining schools were still a nationally representative sample once the 43 pilot schools were excluded. The authors refer to data collected from those 110 schools as "baseline" data.

Next, in 2016/2017, RTI collected "midline" data from G2 and G3 students in a sample of 240 nationally representative schools, between the end of 2016 and the beginning of 2017. The sample excluded schools without at least 20 students in G2 and G3 (combined), 240 schools sampled for MSI's evaluation, and 99 recently established schools.

RTI's main analysis compares outcomes for students in the "baseline" and "midline" samples. However, this comparison cannot be used to draw causal conclusions about the impacts of RAMP because it is comparing outcomes of two samples at two different points in time. While it is possible to estimate differences in indicators between the two samples across the two periods, the difference cannot be attributed to the intervention given that there may be other causes for the changes that cannot be ruled out. Alternative explanations (or confounding factors) for the differences observed can include "interventions run by other donors (or the Ministry of Education), natural events (e.g., rainfall, drought, earthquake, etc.), government policy changes, or natural changes that happen in an individual or community over time" (USAID 2013, p. 2).

Two additional issues threaten the validity of RTI's conclusions: (1) It is unclear whether the sample of 110 schools is nationally representative and (2) the midline sample includes students with varying levels of exposure to RAMP, including children that have not been exposed to RAMP. We elaborate on these issues below.

Again, it is unclear whether the 110 schools in the baseline sample constitute a nationally representative sample. Even if we assumed that no other factors could have led to differences in students' outcomes between 2014 and 2016/17, the 110 schools could have differed from schools in the midline sample before the introduction of RAMP. Comparing two samples from dissimilar populations is problematic because differences could be related to the sample composition, rather than to the intervention. Importantly, impact evaluations follow rigorous procedures to ensure that samples are similar for all or most measurable characteristics so that the only difference will be exposure to the intervention. For example, the school selection used for the Pilot activity could have produced a sample of 110 schools that, on average, were poorer or had a larger share of underperforming students than schools in the nationally representative sample. If that were the case, it would be reasonable to expect students in the 2016/2017 sample to perform better than students in the 2014 sample. Further, the report indicates that 43 schools were *purposively* selected (as opposed to *randomly* selected) to receive the Pilot activity in 2014. The rationale for the purposive sampling was not provided but there could be something about these schools that makes them non-representative of the typical Jordanian school. There is no analysis of the pilot group or the larger sample to determine whether they are representative of all Jordanian schools.

The report also states that approximately 50% of the students in the study had not been exposed to RAMP by the time of the midline assessment (p. 4). This means that the results presented in Tables ES3 and ES4—even *if* attribution were possible from a performance evaluation—RAMP only reached half the sample. The report further states that, due to the variation in exposure, significant differences observed in G2 are encouraging and null findings in G3 are not reason for concern. Yet, the overall positive trends that RTI attributes to RAMP include significant differences for untreated students (see Tables 9 and 14 for example), which *indicates that factors other than RAMP* underlie the findings.

Despite not being an impact evaluation, *if* RTI were able to provide evidence that the 110 schools are indeed a nationally representative sample of Jordanian schools, then the study could make important contributions to understanding students' early reading and mathematics. The study, better characterized as a *performance evaluation*, provides useful descriptive information about national trends in G2-G3 students' outcomes from 2014 to 2017. Moreover, it identifies discrete foundational reading and math skills where

students need additional support to meet performance benchmarks. Finally, it offers descriptive information about Syrian students, which is not available from other sources and can be useful for education authorities in the region.

MSI'S "RAMP IMPACT EVALUATION: ESTIMATING IMPACTS OF EARLY-GRADE READING AND MATH PROJECT (RAMP) IN JORDAN": MIDLINE REPORT.

In the most rigorous impact evaluation design—a randomized control trial—schools would be randomly assigned to the intervention or comparison groups before implementation (USAID, 2011). Random assignment is the best method to create a credible approximation to the counterfactual, because it increases the likelihood that that the groups being compared are similar before introduction of the intervention. MSI was unable to randomize schools for this impact evaluation, because RTI planned to implement RAMP at the governorate level and had already determined that cohort 2 governorates would receive the intervention in 2017.

The evaluation used a quasi-experimental design (QED) to approximate the counterfactual, using *propensity score matching* to create groups that were as similar as possible at baseline. USAID (2013) notes that matching is "the most common means for selecting a comparison group, wherein the evaluator selects a group of similar units based on observable characteristics that are thought to influence the outcome" (p. 5). MSI's evaluation used a two-step propensity score matching process with school-level characteristics and later with student data. The midline report shows that, after matching, impact inferences are based on groups of students that are statistically equivalent at baseline, according to strict criteria set by the Institute of Education Science's What Works Clearinghouse.[13]. Having equivalent groups minimizes the risk that between-group differences observed at midline (or endline) are due to factors other than RAMP.

MSI's evaluation is not without limitations. A key limitation is that matching cannot account for differences in unobservable (or unmeasured) characteristics. There is a risk that statistically equivalent groups may differ in variables not included in the matching process, which could lead to erroneous conclusions about the intervention's impacts. Most relevant to this document is that while RTI's 2016/2017 are generalizable across the country, results from MSI's evaluation are generalizable to cohort 2 schools only (although there were differences in the EGRA/EGMA instrumentation for the two studies.) The reasons are that (1) cohort 1 governorates were not included in the study

[13] The What Works Clearinghouse (WWC) is an initiative of the U.S. Institute of Education Sciences to evaluate studies on the effectiveness of programs, policies and practices. WWC Standards Briefs lay out rules to assess the quality of studies and are highly regarded in the field of program evaluations.

because the implementation of RAMP was already underway in those governorates, and (2) propensity score matching created a sample of cohort 3 schools and students that was as similar as possible to cohort 2 schools to credibly approximate the counterfactual.

INSTRUMENTATION

There were also differences in the EGRA/EGMA instrumentation used to measure students' early grade reading and math outcomes. First, RTI used the same tool for grade 2 and grade 3 students. In contrast, the MSI evaluation used separate EGRA/EGMA tools for students in grades 2 and 3. These tools were developed in a collaborative process with the Ministry of Education. The MSI tools were matched to the Ministry's most recent 2016 curriculum.

CONCLUSION

The reports prepared by RTI and MSI summarize different studies aimed at estimating students' reading and mathematics outcomes. However, as per USAID's Evaluation Policy each RTI and MSI differ in their studies. RTI's study can be characterized as a performance evaluation and cannot be used to make inferences about the impact of RAMP on students' outcomes directly. MSI's study, in contrast, is an impact evaluation draws causal conclusions about the impacts of RAMP (or lack thereof). Differences in the designs of the two studies, and differences in instrumentation, explain the lack of convergence in the results and conclusions.

REFERENCES

USAID (United States Agency for International Development), 2011. *USAID Evaluation Policy*. U.S. Agency for International Development, Washington D.C. Retrieved on December 5, 2017, from https://www.usaid.gov/sites/default/files/documents/1870/USAIDEvaluationPolicy.pdf

USAID (United States Agency for International Development), 2013. Impact Evaluations Technical Note. *Monitoring and Evaluation Series*. Bureau for Policy, Planning and Learning: Retrieved on December 5, 2017, from https://usaidlearninglab.org/sites/default/files/resource/files/ie_technical_note_2013_0903_final_2.pdf

ANNEX T. MESP RAMP IMPACT EVALUATION ENDLINE PRESENTATION TO USAID/JORDAN (AUGUST 2018)

Please see slides on the following pages.

ESTIMATING THE IMPACT OF RAMP ON TEACHER PRACTICES AND STUDENT EARLY GRADE LEARNING IN JORDAN

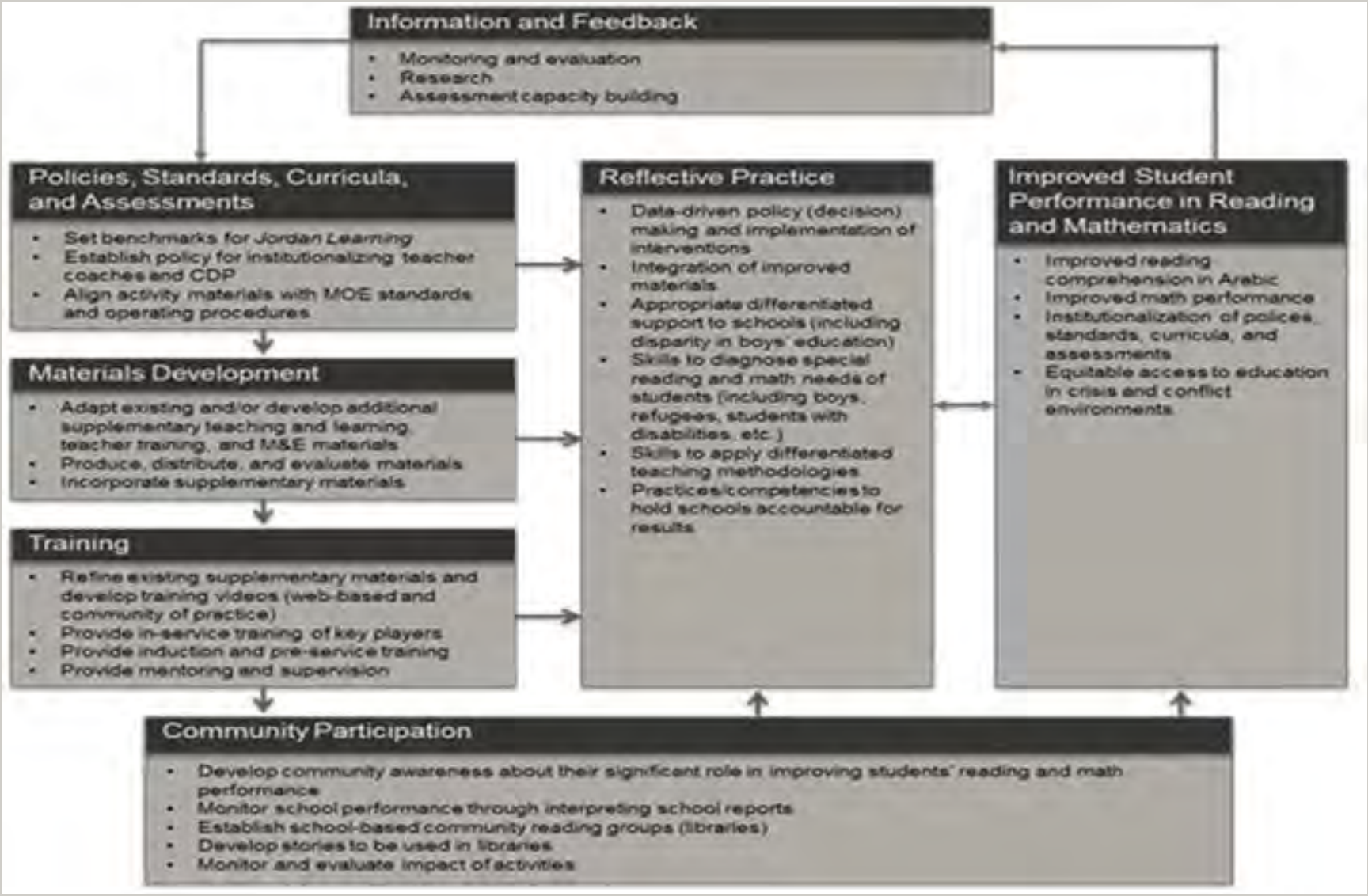
Preliminary Impact Evaluation Endline Results – August 2018

Jordan Monitoring and Evaluation Support Project (MESP)

RAMP Impact Evaluation Questions

1. What are the **impacts of RAMP on teachers** delivering effective reading and math instruction?
2. What are the **impacts of RAMP on G1, G2, and G3 students'** proficiency in reading and math?
3. Did the **RAMP impacts vary by gender,** location (urban or rural), nationality (Jordanian or other), session (morning or afternoon), school type (boys, girls, or mixed), **number of shifts** (single or double shift), or whether the school received infrastructure support from USAID?
4. How **cost-effective** is the RAMP intervention for impacts on reading and math outcomes? (cancelled)

Theory of Change



Quasi-Experimental Longitudinal Study

Study	Instruments	Goal
Impact study of teachers' practices	• Classroom Observations of Teaching Instruction (COTI) tool • Teacher survey	Measure the impacts of RAMP on G3 teachers' practices
Impact study of students' learning	• Early Grade Reading Assessment (EGRA) and Early Grade Math Assessment (EGMA) • Student survey • Teacher survey • Principal survey	Measure the impacts of RAMP on students' learning in G1, G2, and G3
Qualitative study	Observation data of mentoring, and interviews with teachers, principals, mentors/coaches, RTI, QRTA, Dijani, MoE	Assess fidelity of implementation and perceptions of RAMP

Teacher Study Design

Descriptive study of teacher practices

(G1 & G2)

RAMP training held:
- RAMP: July/Aug 16
- Comparison: July/Aug 17

Baseline
November 2016

Explore early differences between RAMP and Comparison (N=40)

Midline
April/May 2017

Explore differences between RAMP and Comparison using larger sample (N=80)

Endline
April 2018

Not implemented

Impact study of teacher practices

(G3; N=200)

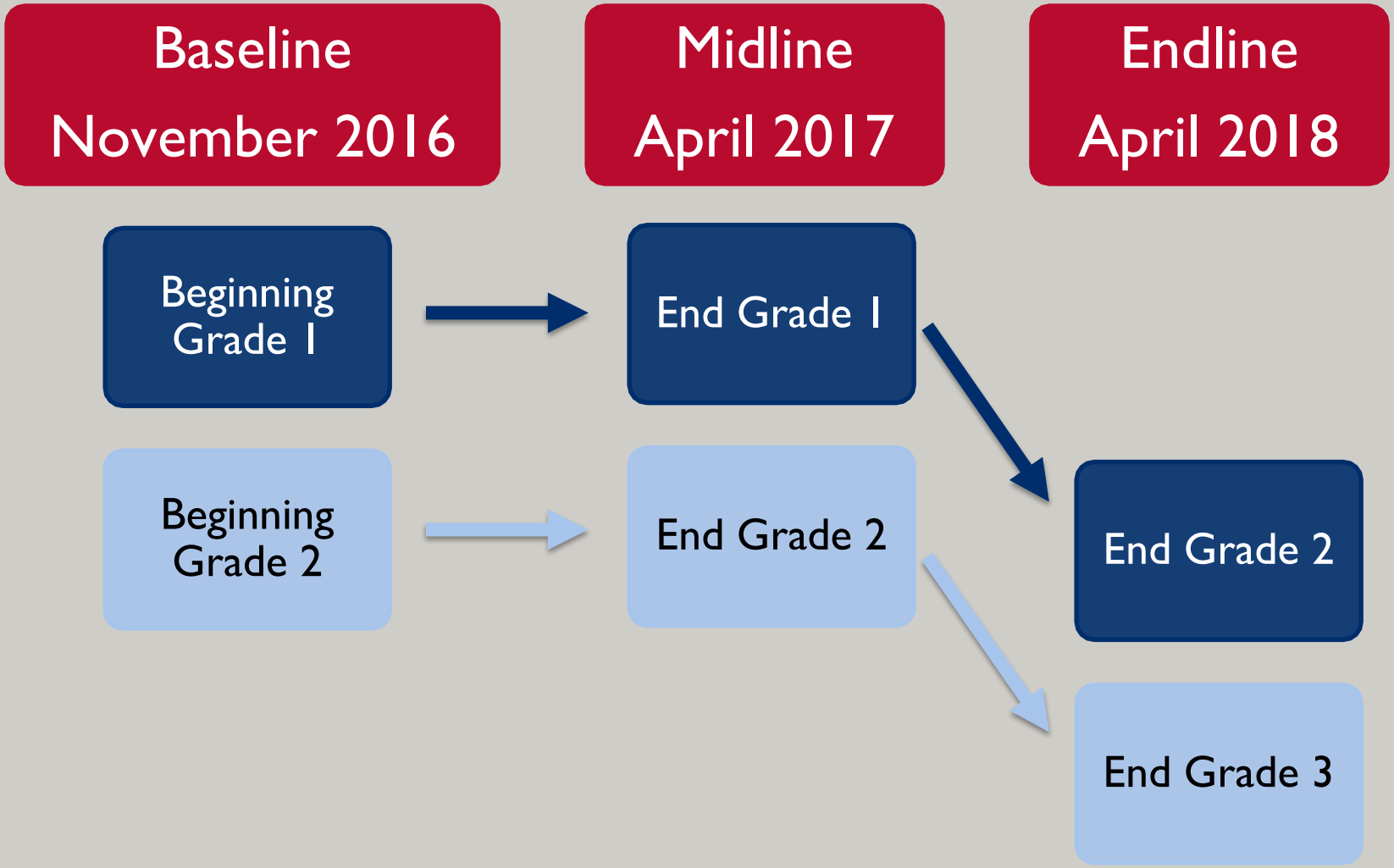
RAMP training held
- RAMP: July/Aug 17
- Comparison: July/Aug 18

Teacher impact Baseline
April/May 2017

Establish baseline equivalence between the groups before RAMP for sample of 200 teachers

Estimate impacts of RAMP

Student Assessment Design



STUDY OF TEACHER PRACTICES

USAID
FROM THE AMERICAN PEOPLE

Results

EQ1 - What are the **impacts of RAMP on teachers** delivering effective reading and math instruction?

RAMP Implementation in the Classroom

RAMP instructional strategies for G3 included asking reading comprehension questions for each paragraph in a text or demonstrating fractions using physical objects.

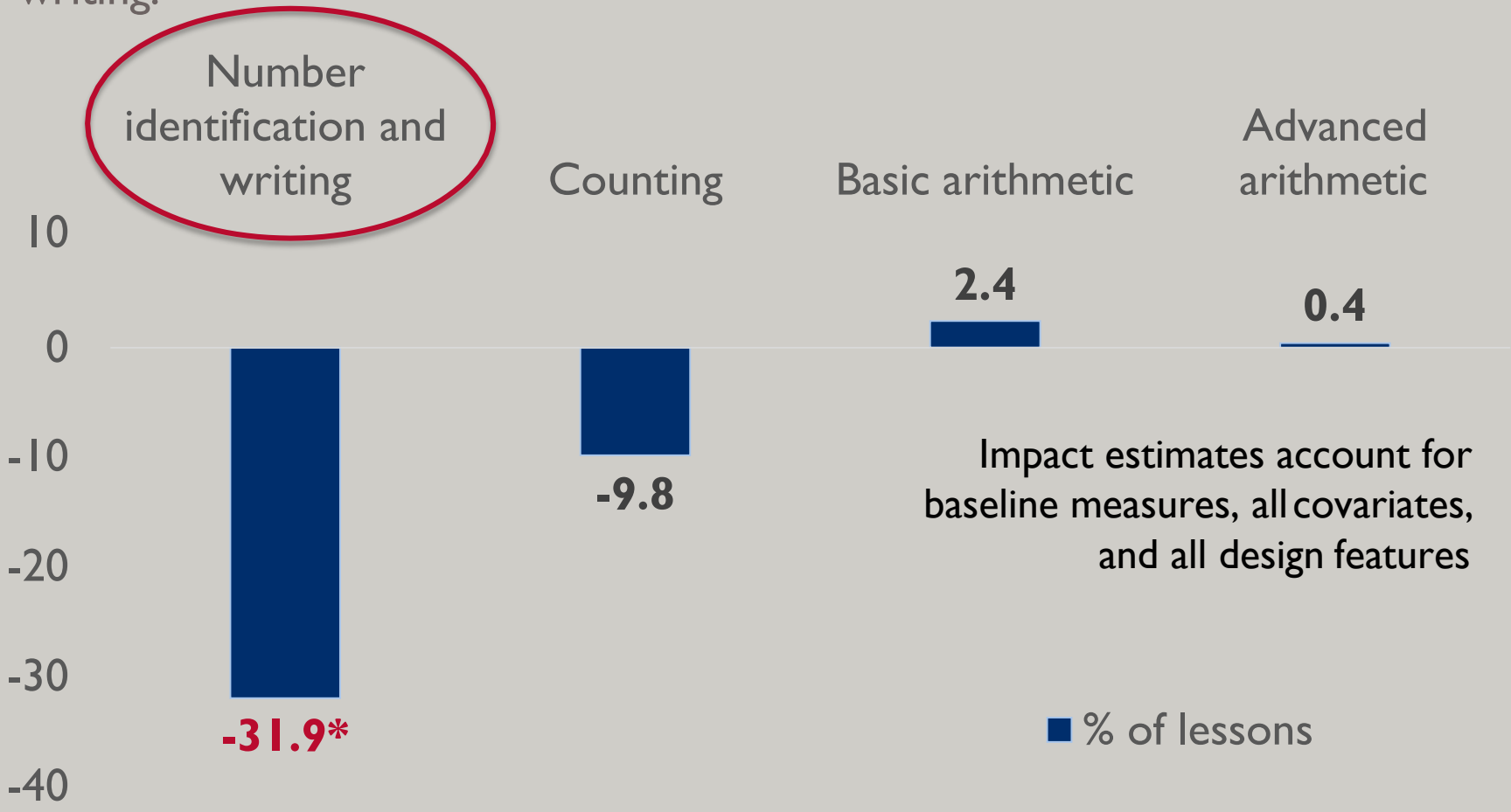
Subject	Group	Midline (2017)	Endline (2018)	Change	RAMP Impact
Math lessons	RAMP	0.2	1.8	1.6	0.6*
	Comparison	0.2	1.2	1.0	
Reading lessons	RAMP	0.3	1.7	1.4	0.4*
	Comparison	0.2	1.1	0.4	

Teachers integrated practices by adding RAMP strategies to a lesson from the MOE curriculum.

Subject	Group	Midline (2017)	Endline (2018)	Change	RAMP Impact
Math lessons	RAMP	0.2	1.8	1.6	0.6*
	Comparison	0.2	1.2	1.0	
Reading lessons	RAMP	0.3	1.7	1.4	0.4*
	Comparison	0.2	1.1	1.0	

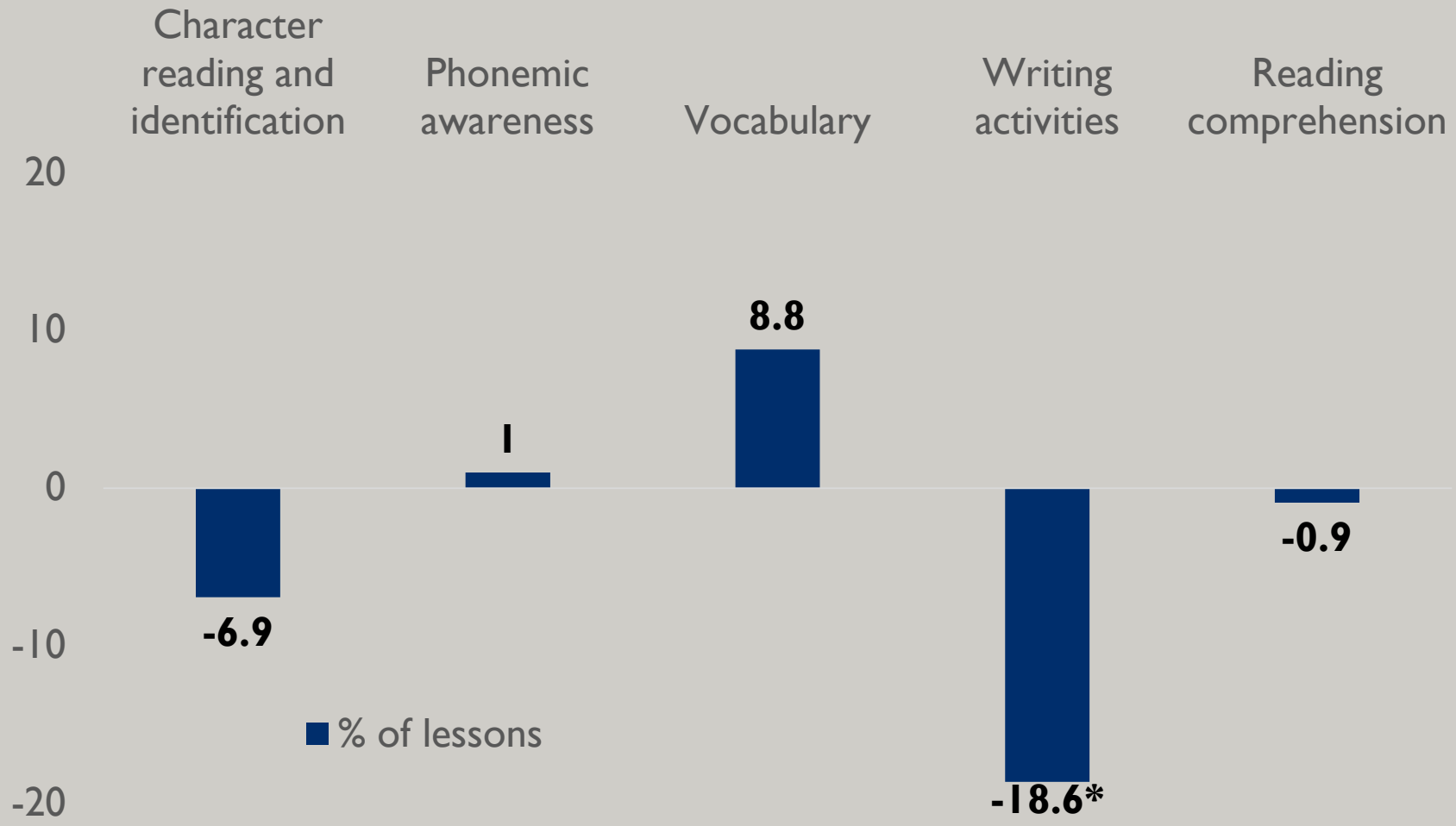
RAMP Impacts on Mathematics Lesson Content

RAMP reduced the amount of time spent on number identification and writing.



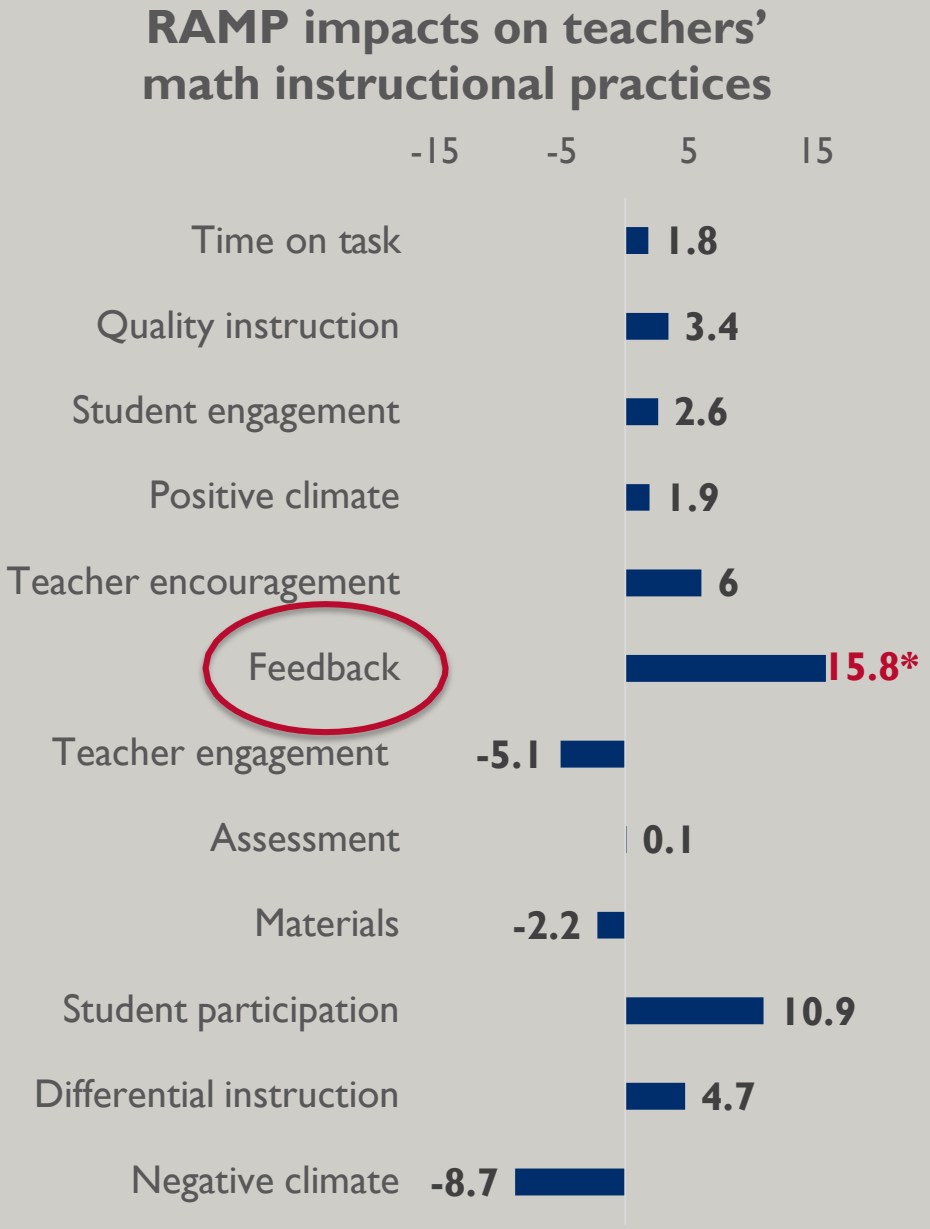
RAMP Impacts on Reading Lesson Content

RAMP reduced the amount of time spent on identifying characters and writing activities, and increased time spent on vocabulary.



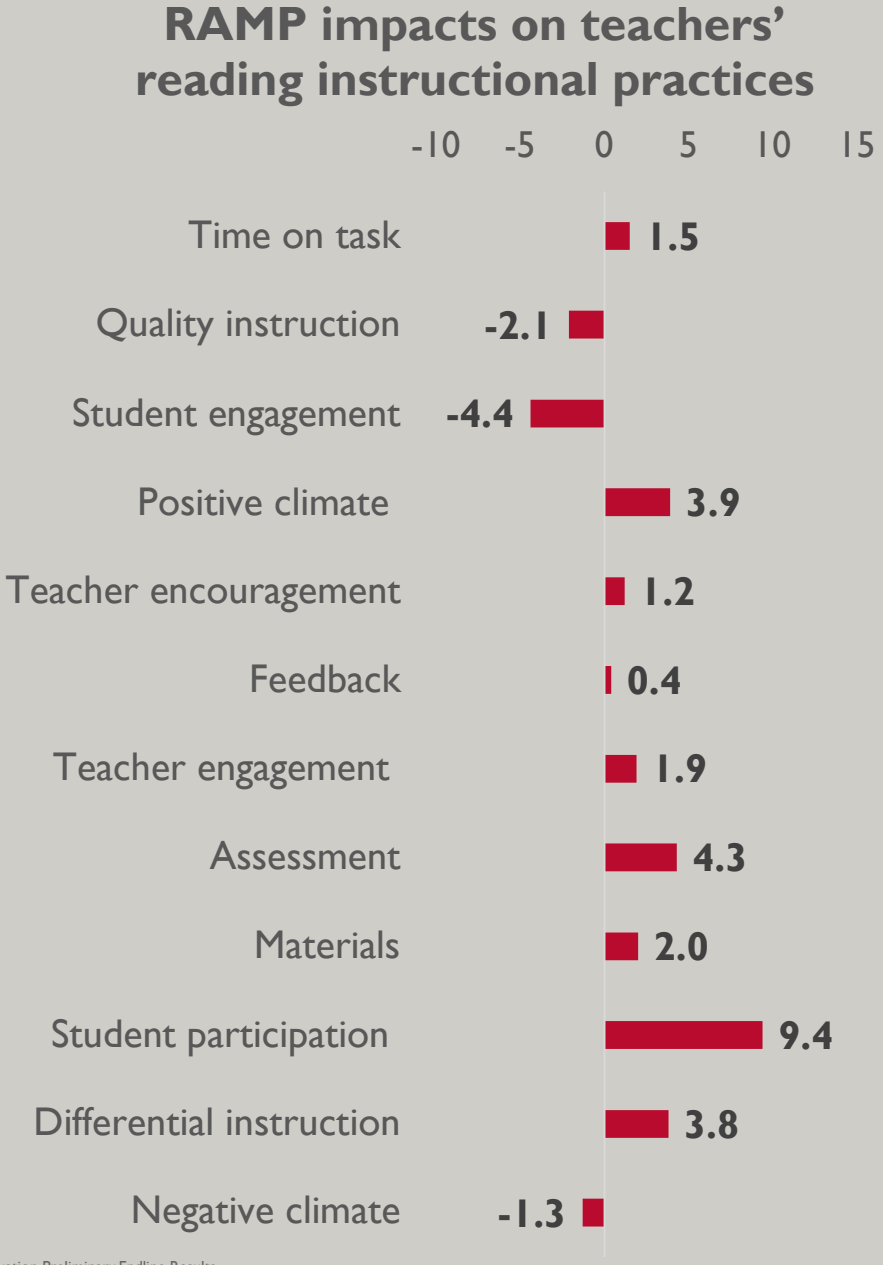
Instructional Practices: Math Lessons

- RAMP had significant positive impacts on teachers' **feedback** on class participation and written work.
- RAMP teachers were more likely to give **specific and strategic feedback**
  - "You used those counters correctly to solve the problem!"
  - "How did you figure out that the answer was 6?"



Instructional Practices: Reading Lessons

Although some outcomes, such as student participation showed positive trends, there were no significant impacts on teachers' instructional practices in reading.



IMPACT STUDY OF STUDENT LEARNING

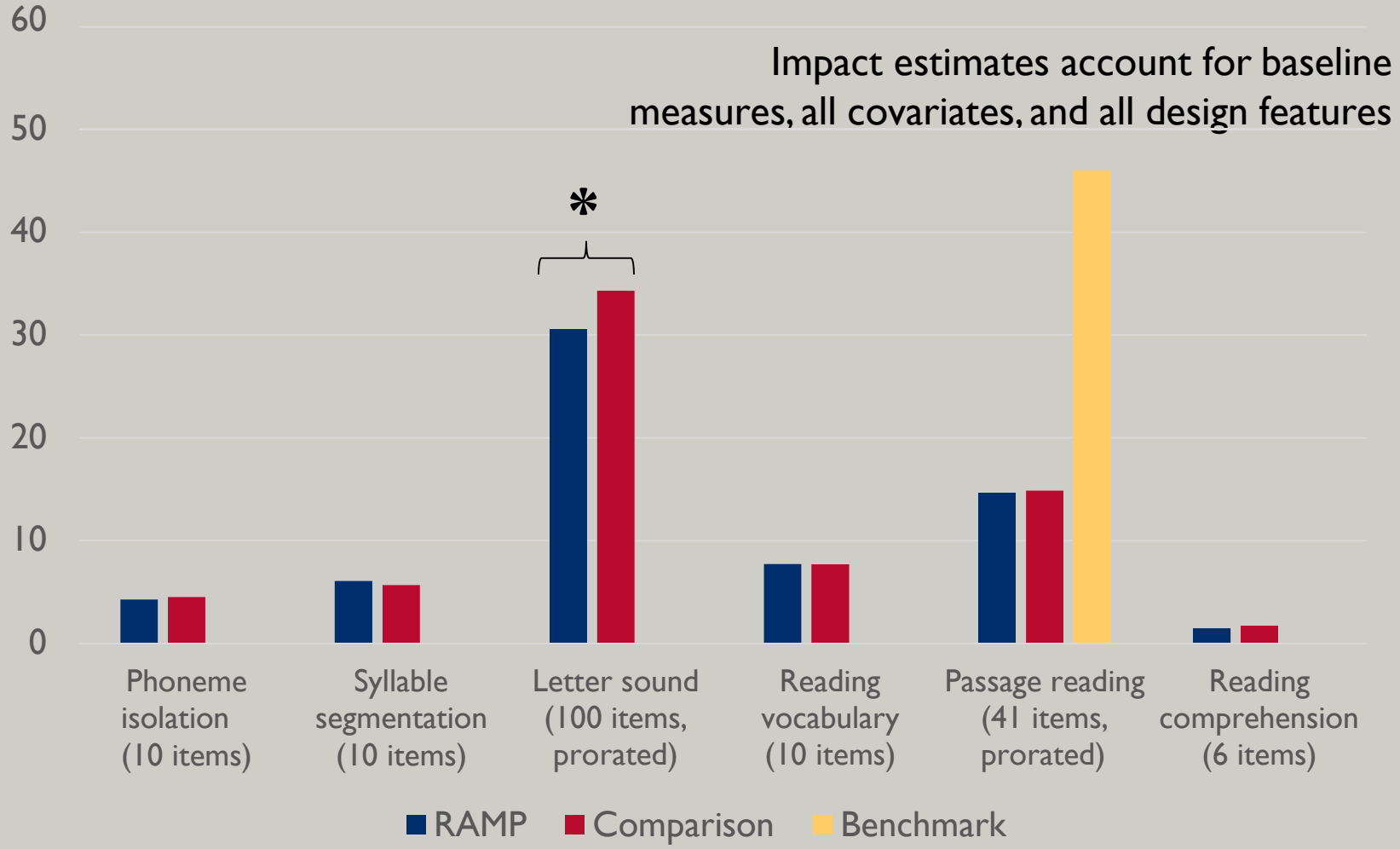
USAID
FROM THE AMERICAN PEOPLE

Results

EQ2 - What are the **impacts of RAMP on G1, G2, and G3 students'** proficiency in reading and math?

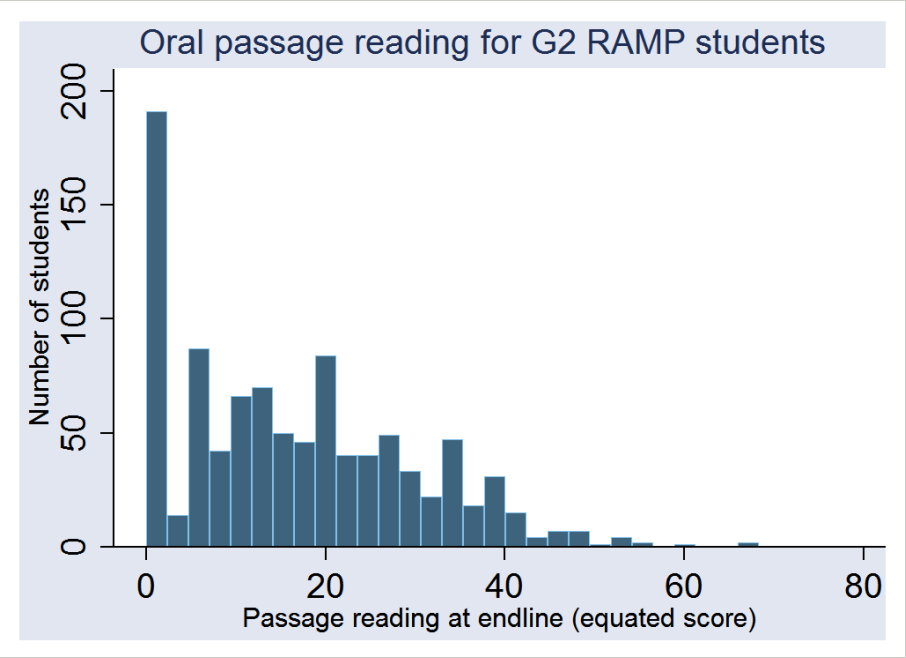
Impacts on Reading Scores for G1 Students

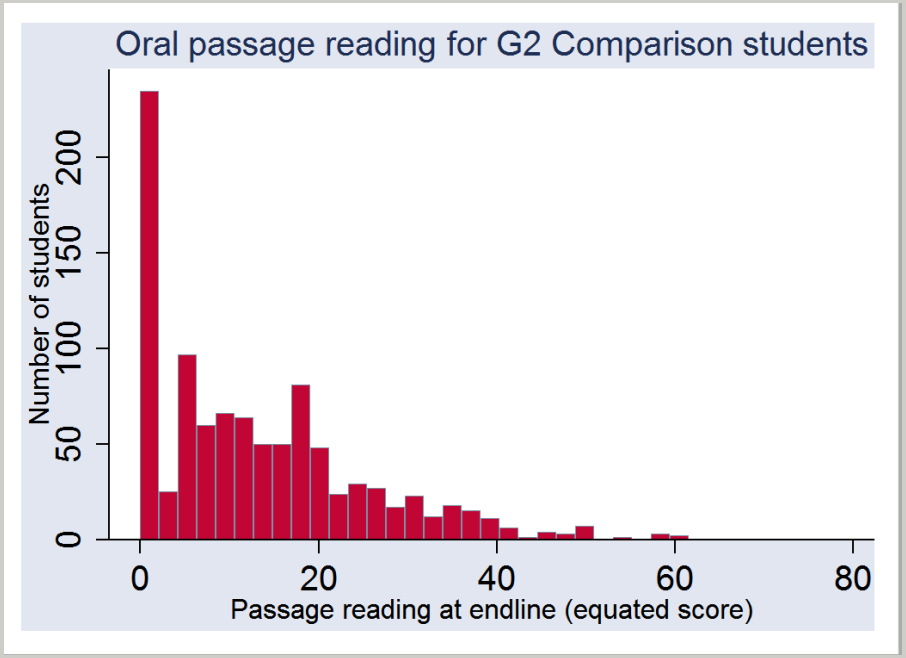
RAMP had no positive impact on G1 students' reading ability and had a negative impact on knowledge of letter sounds.



Impacts on Reading Scores for G1 Students

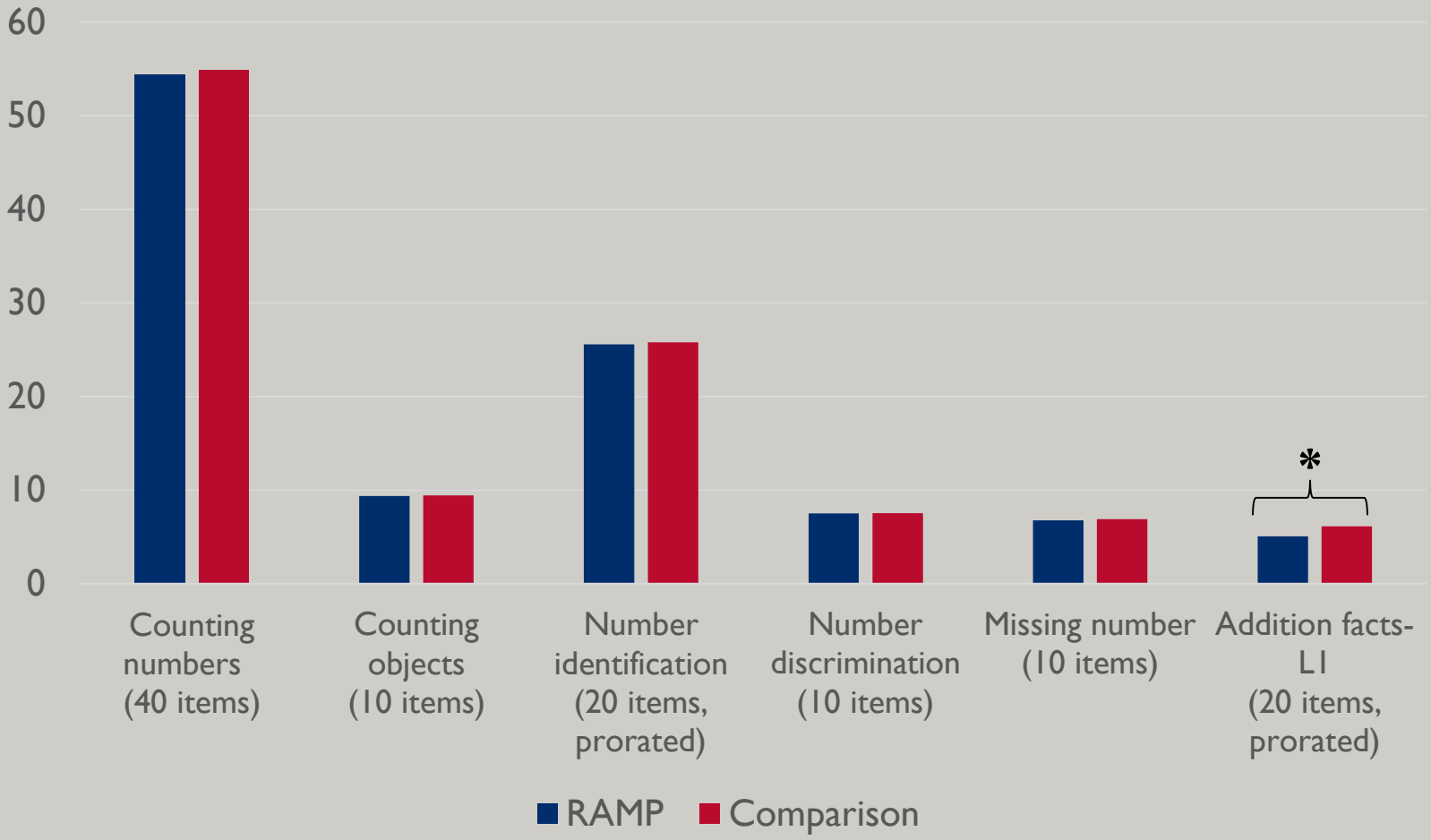
- A large number of students in both groups were unable to read a single word correctly (18% RAMP and 22% comparison)
- Only 2-3% of students in both groups were able to read 46 or more words per minute.





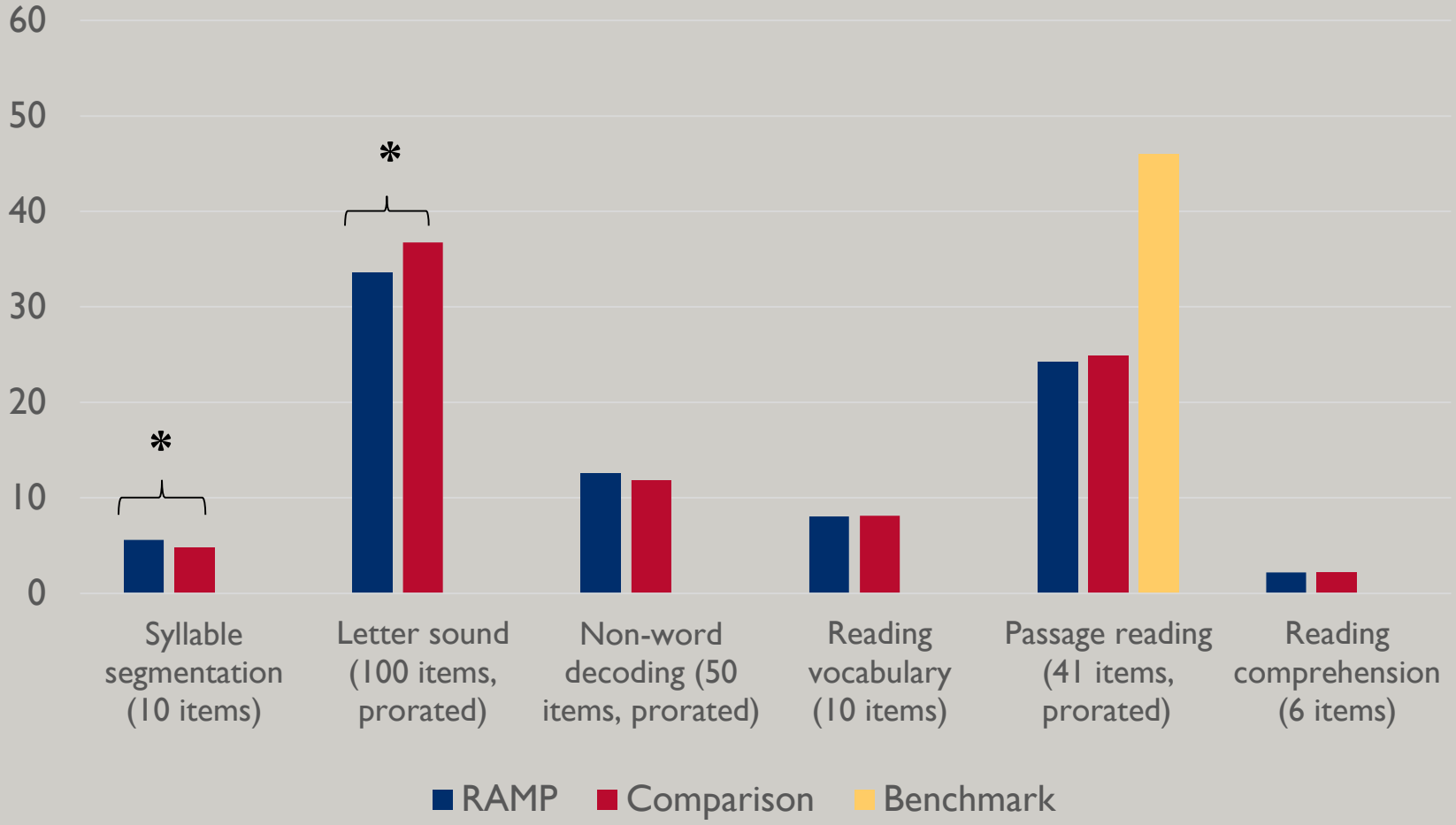
Impacts on Math Scores for G1 Students

RAMP had no positive impact on G1 students' math outcomes and had a small negative impact on addition.



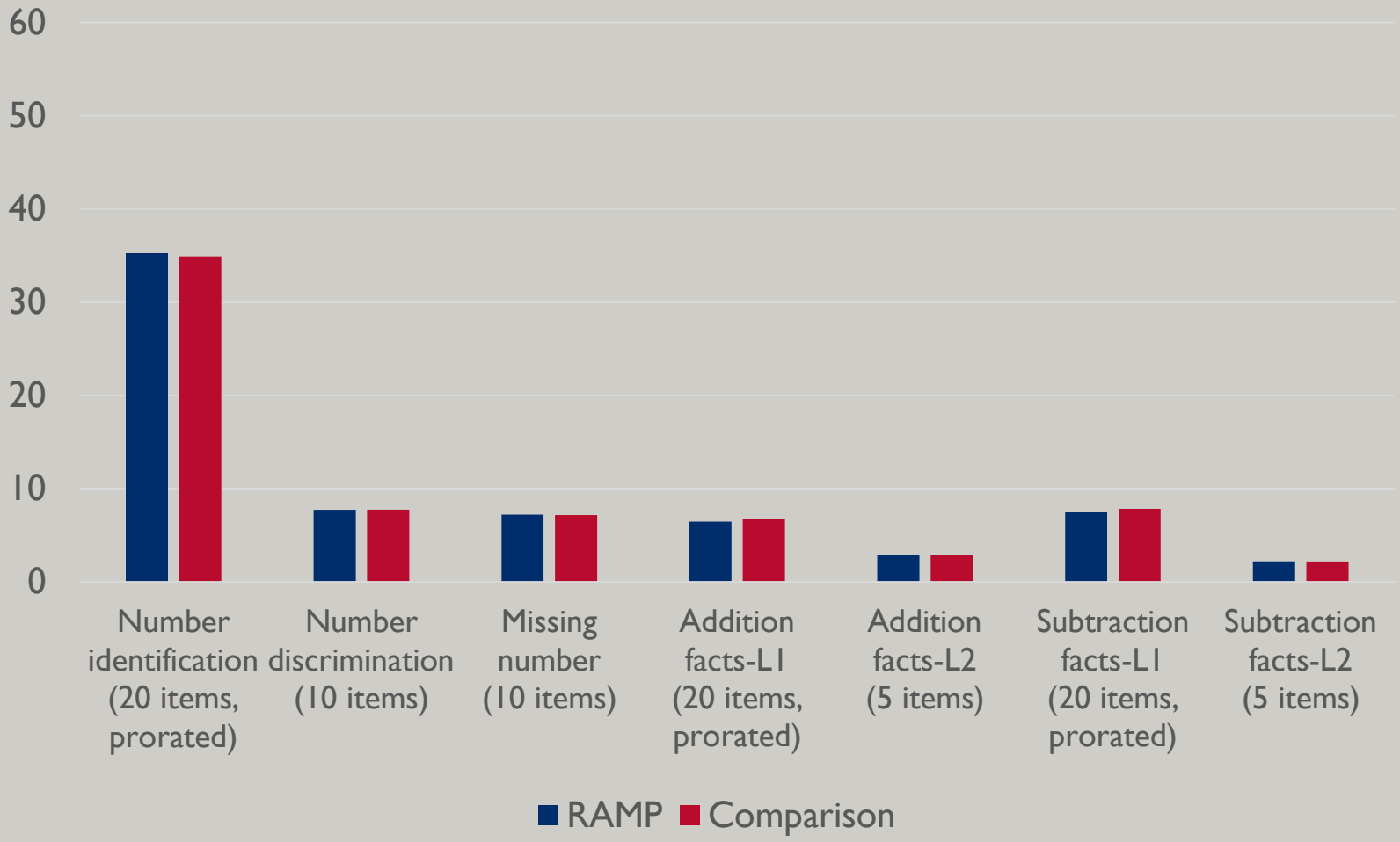
Impacts on Reading Scores for G2 Students

RAMP had a small positive impact on G2 students' ability to segment words into syllables and a small negative impact on knowledge of letter sounds.



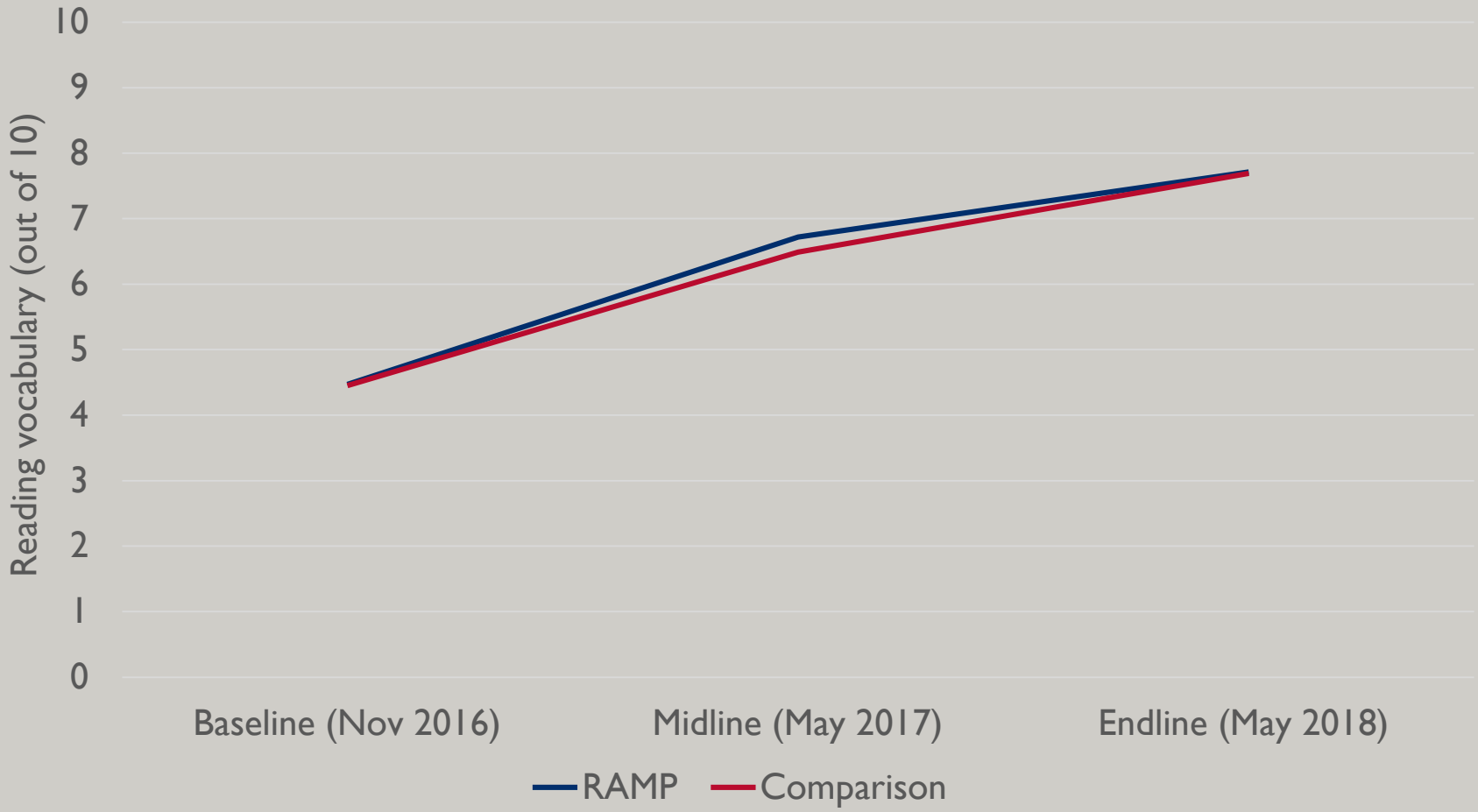
Impacts on Math Scores for G2 Students

RAMP had no positive or negative impacts on G2 students' math proficiency.



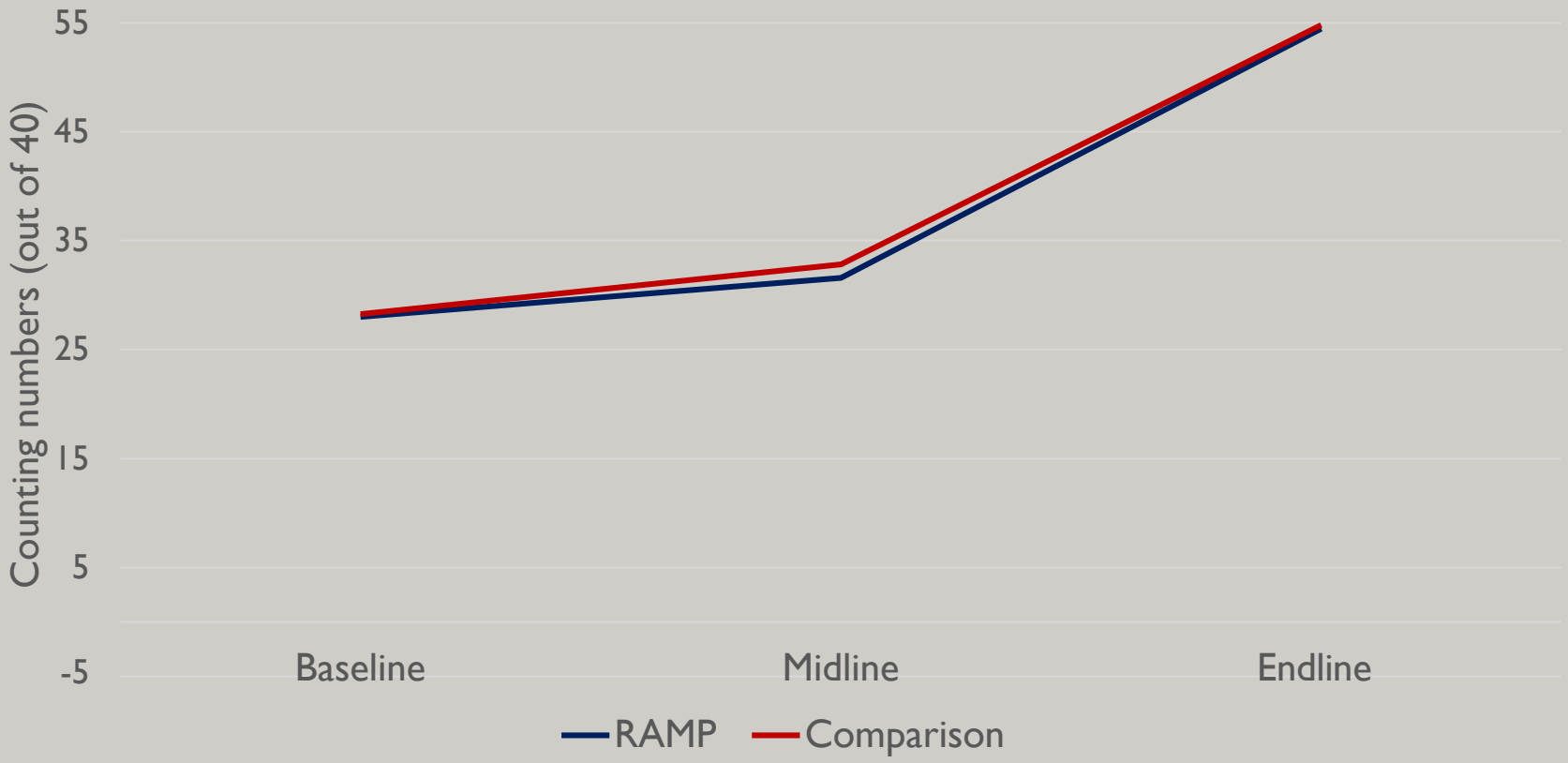
Growth in Reading Scores for G1 students

Example: **Vocabulary** improved at similar rates in both groups over the course of G1 and G2



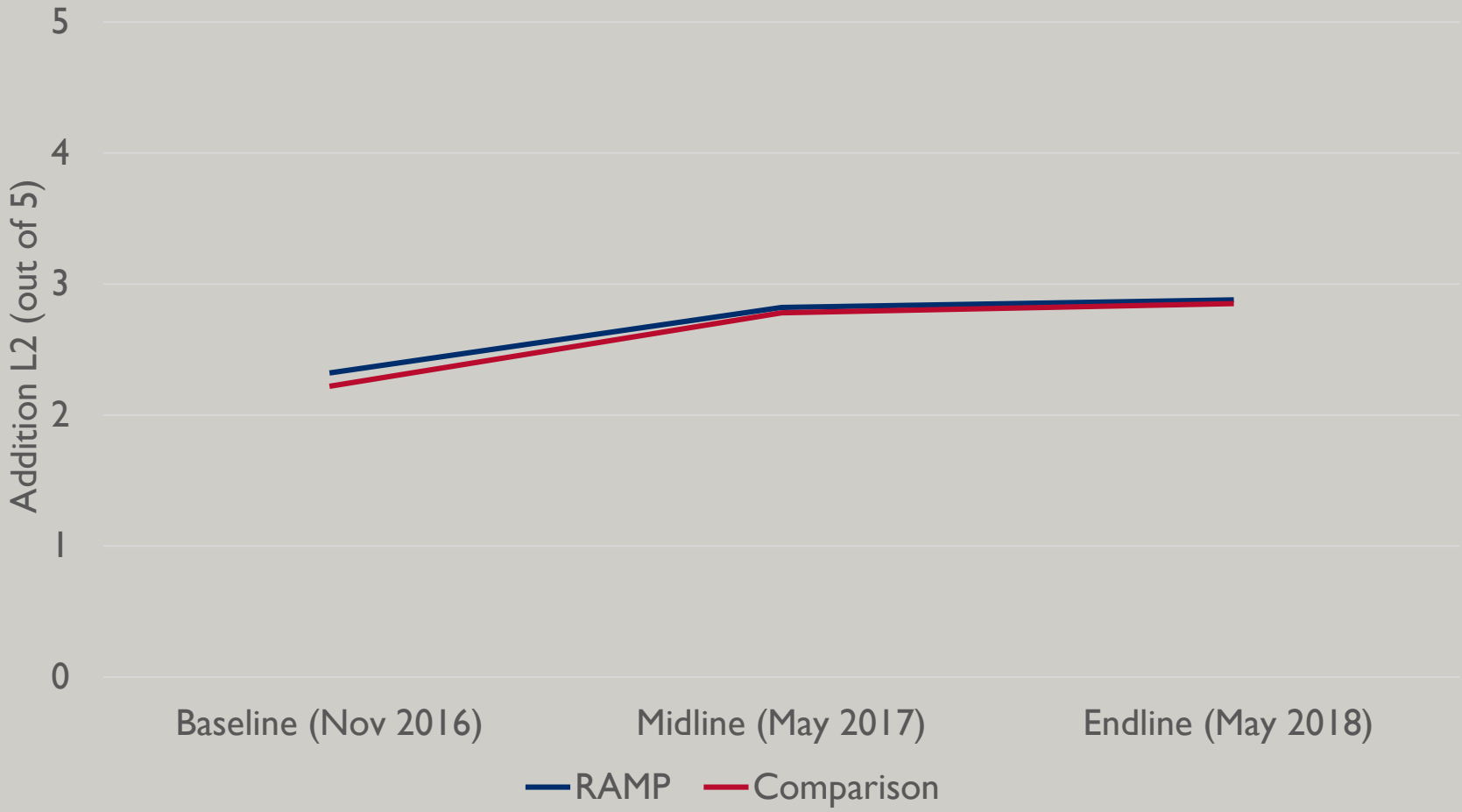
Growth in Math Scores for G1 students

Example: Students in both groups improved their ability to count numbers over the course of G2



Growth in Math Scores for G2 students

Example: Students' ability to solve basic addition problems remained stable over the course of G2 and G3 in both groups

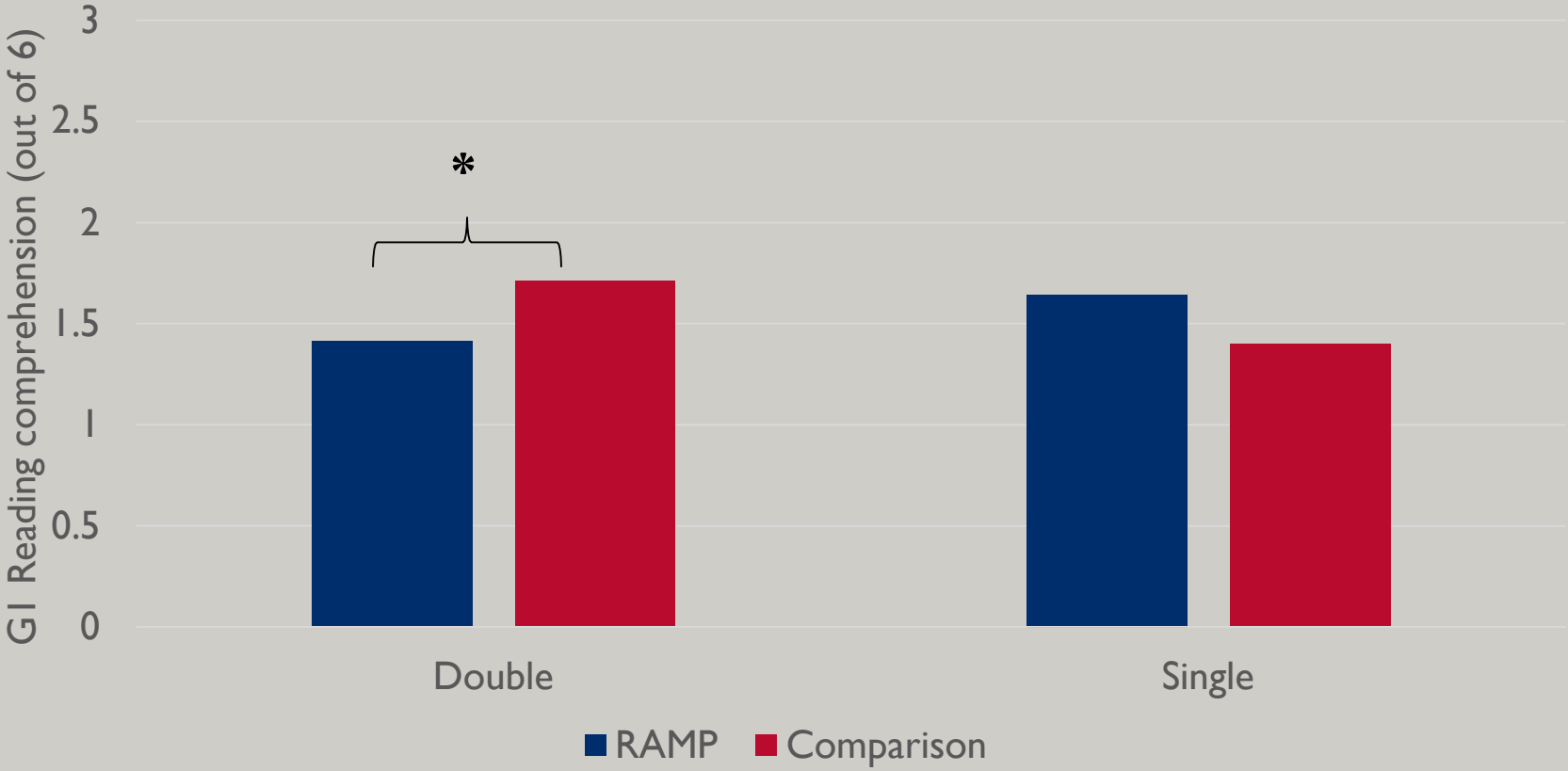


Results

EQ3 - Did the **RAMP impacts vary by gender,** location (urban or rural), nationality (Jordanian or other), session (morning or afternoon), **school type** (boys, girls, or mixed; single or double shift), or whether the school received infrastructure support from USAID?

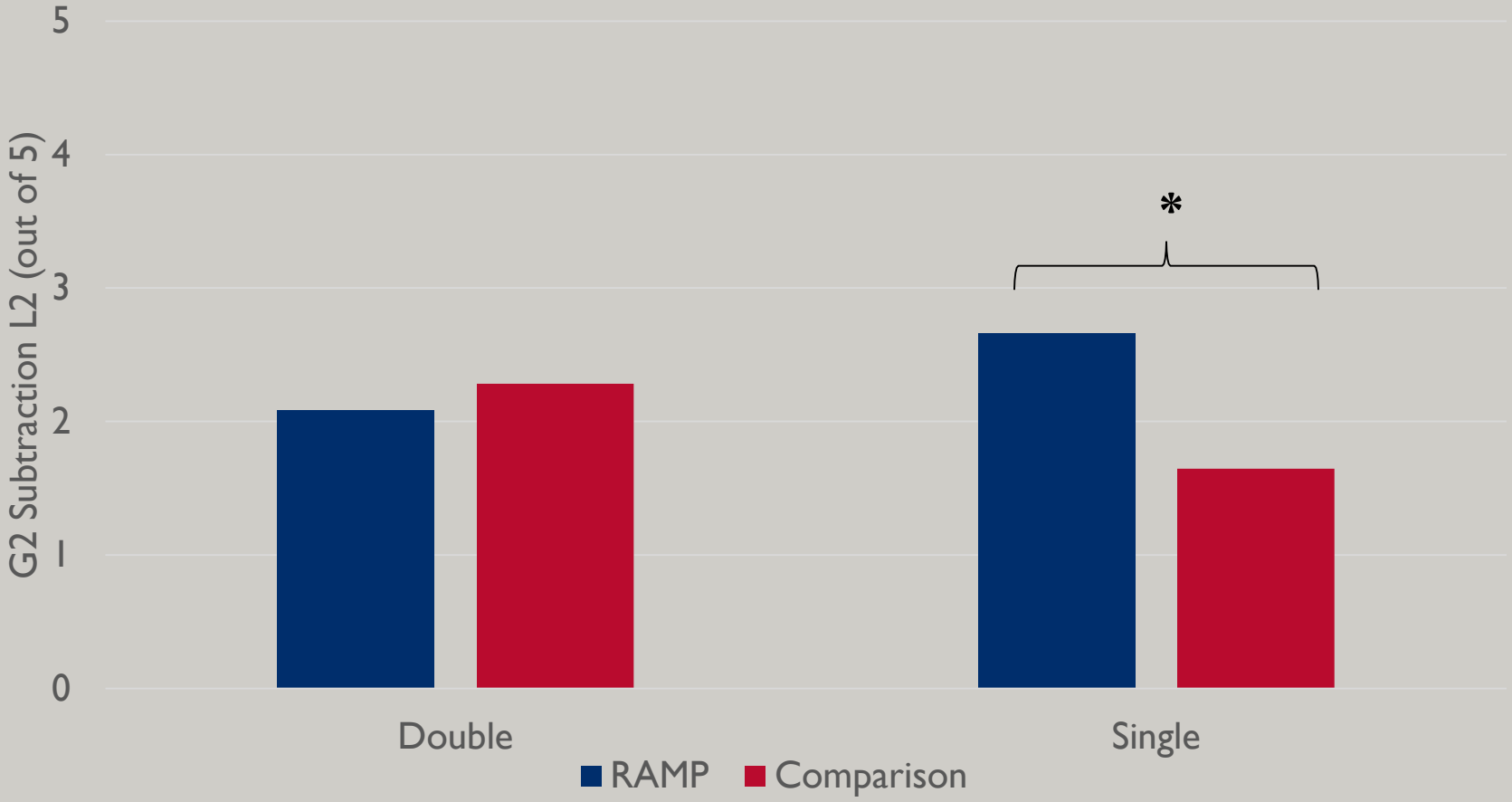
RAMP Impacts by Number of School Shifts G2

In double-shift schools, G2 RAMP students scored lower in reading comprehension than comparison students. In single-shift schools, the difference was not statistically significant and it was in the opposite direction.



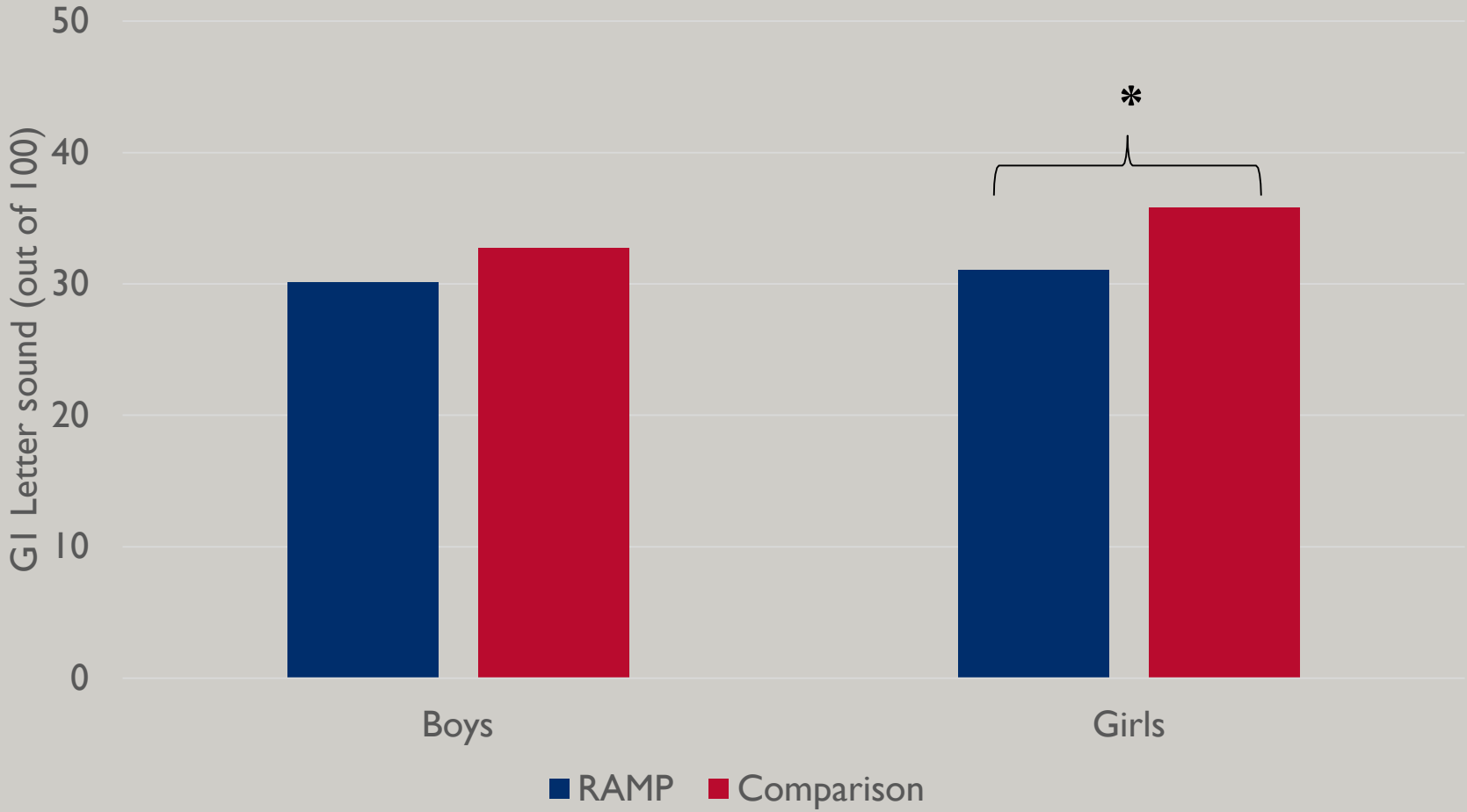
RAMP Impacts by Number of School Shifts G3

In double-shift schools, students in both groups performed similarly in the subtraction subtask. In single-shift schools, G3 RAMP students outscored students in comparison schools.



RAMP Impacts by Gender on Letter Sounds G1

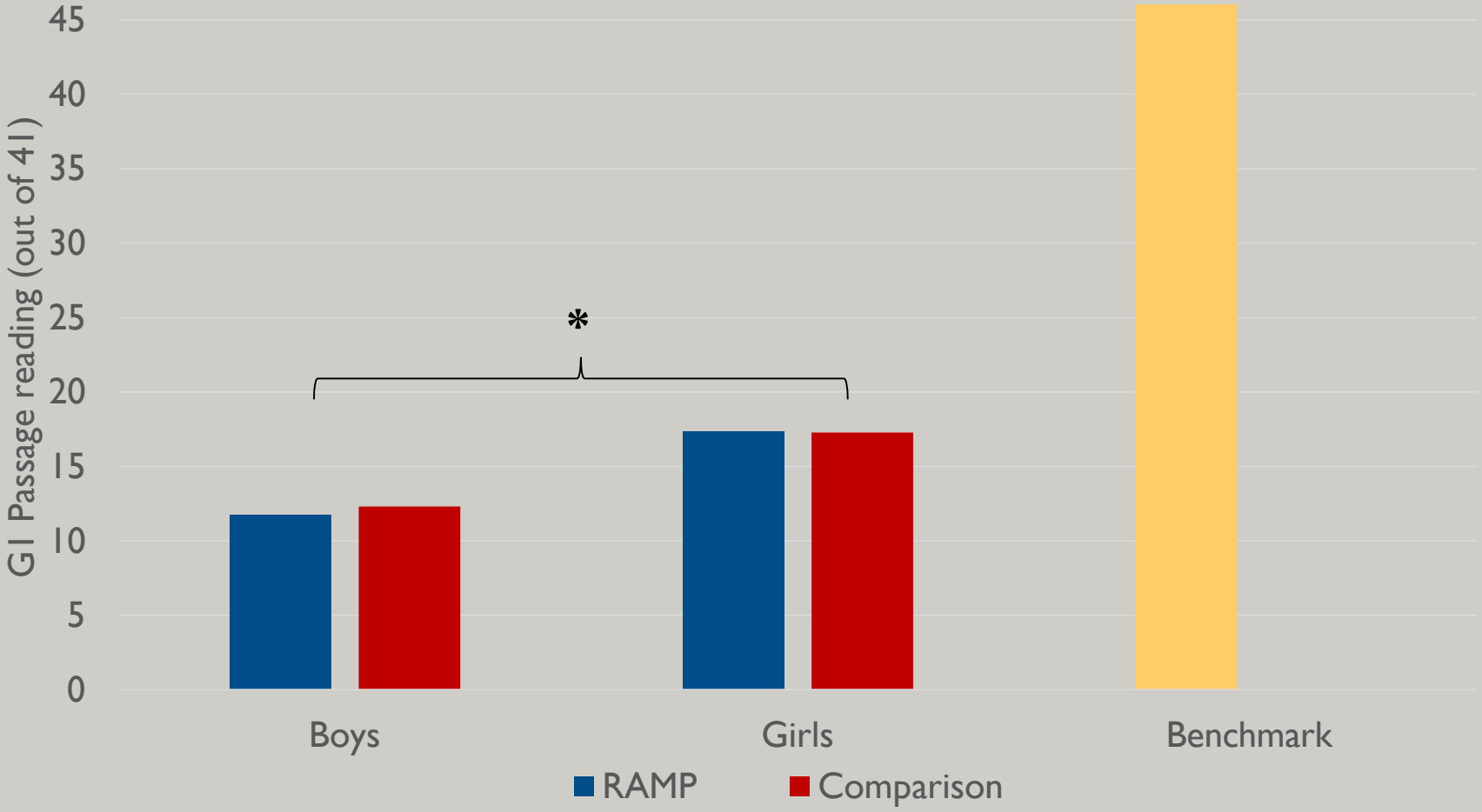
The impacts on reading (in both grades) did not vary by student gender.



Reading Performance by Gender for Passage Reading G1

G2 girls and boys had similar performance in most reading tasks.

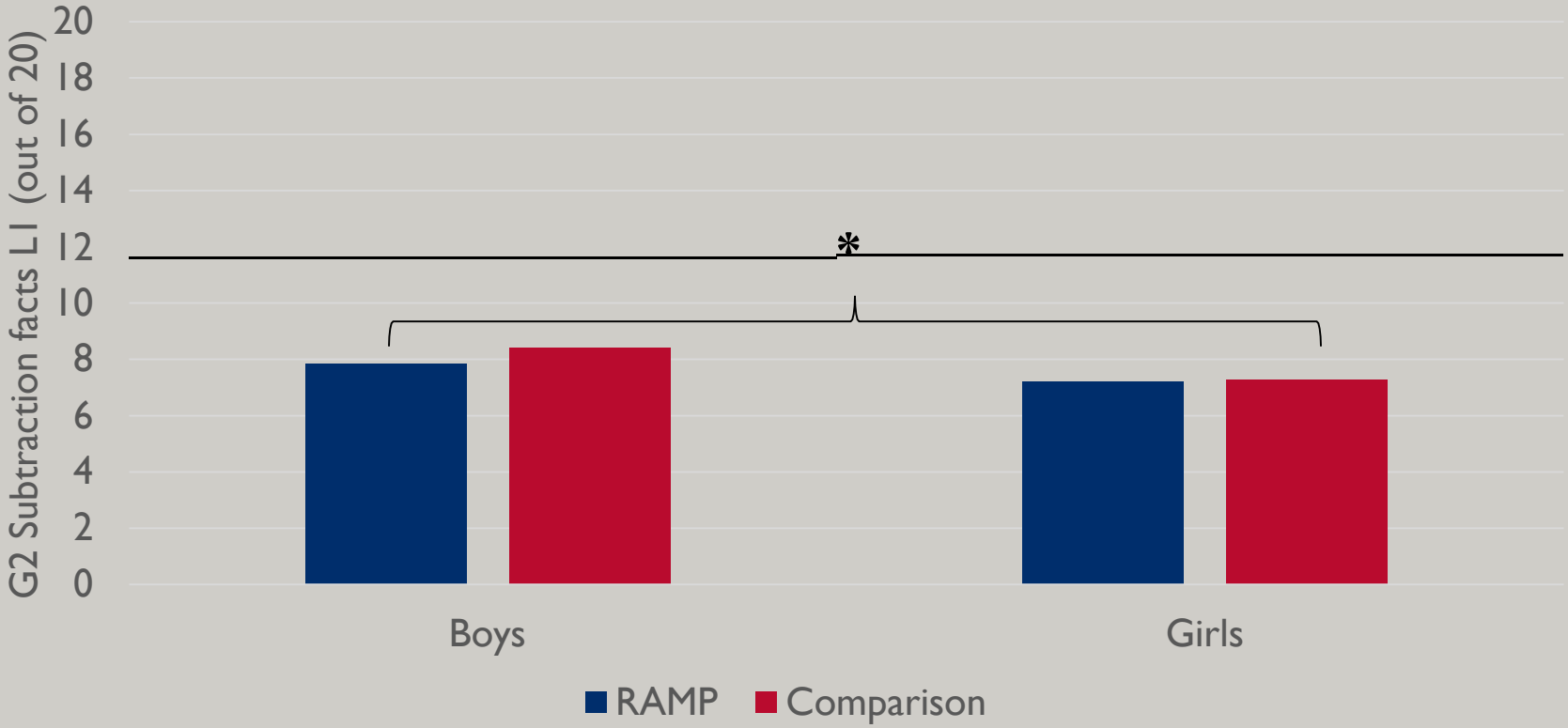
G2 girls outscored boys in reading vocabulary and oral passage reading.



Math Performance by Gender for Subtraction G2

G2 and G3 boys and girls had similar performance in timed math sub-tasks.

G2 boys outscored girls in addition and G3 boys outscored girls in subtraction.



QUALITATIVE STUDY

Factors to Consider – Understanding of RAMP

- **Scaling-up**
  - Challenges associated with scaling up a successful pilot project nationally.
    - Variation in context and conditions
      - "*The classes for the intervention schools were purposefully selected to represent, as far as possible, those classes in which the intervention conditions could be as ideal as possible. This was to ensure that the endline survey measured what could be achieved if the intervention were implemented under the best possible conditions.*"

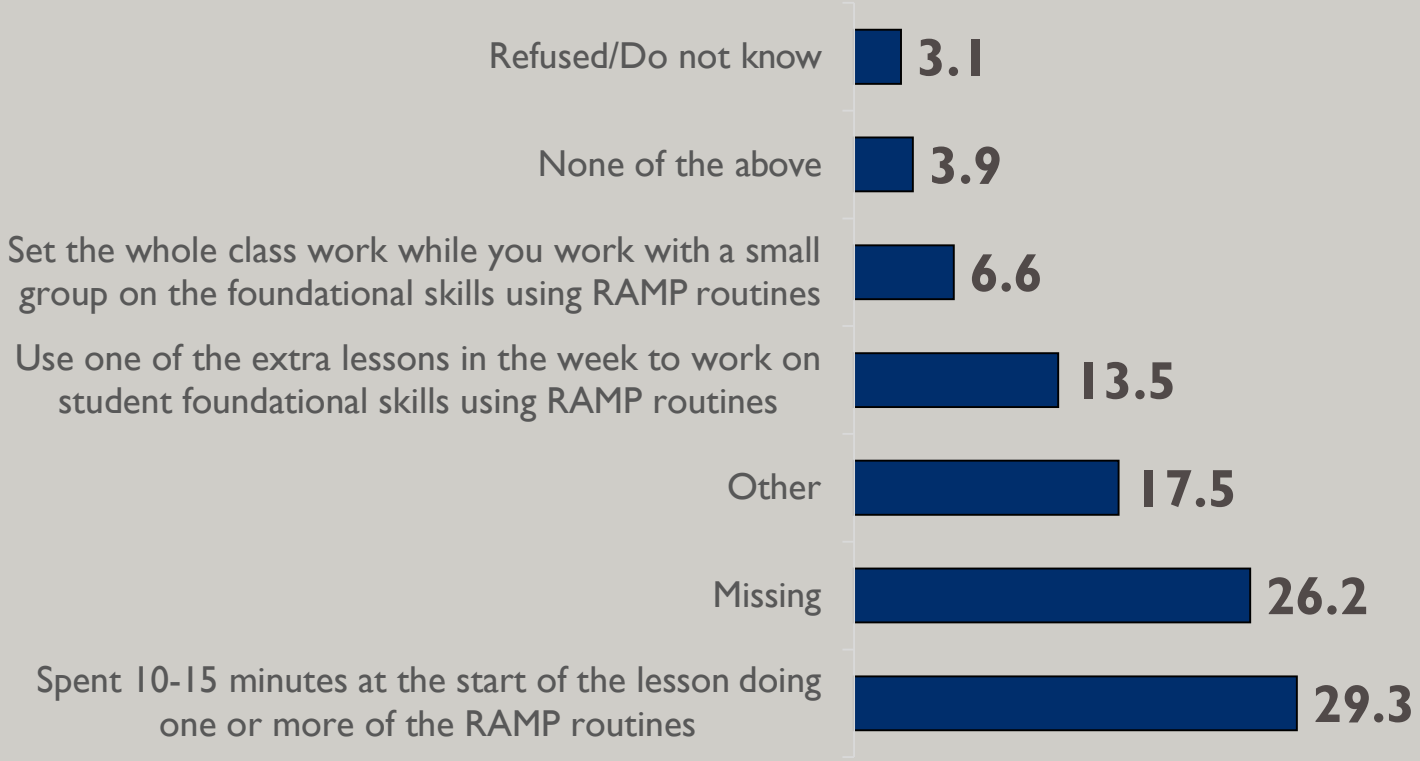
RTI Impact Analysis Report 2014, p33

Factors to Consider – Understanding of RAMP

- **Differences in understanding of RAMP methods/routines among key stakeholders and how to apply them**
  - Teachers
    - Variation at the training level
    - Variation in feedback received
  - Coaches/Mentors
  - Principals

Approach Used to Apply RAMP Routines

Teacher Percent Response by Type of RAMP Approach Used in the Classroom



Factors to Consider – Theory of Change

- Intervention primarily focuses on teachers. Impact at the student level results may take a longer time.
- Some elements of the theory of change are supported by existing literature and evidence (e.g. coaching and mentoring) though there is a need to consider the evidence base around the combination of inputs that constitute RAMP having an impact at the student level.

Coaching/Mentoring Model

- The mentoring element of the RAMP intervention has evolved into a three phase approach. A shift from one phase to the next is not based on satisfaction on any specific criteria, but rather based on maturity of the cohort (i.e. the passage of time).
  - **Phase 1:** Coach/mentor visits, and frequency of visits are the most important aspect of this phase to prompt the teachers to implement RAMP methods. This phase is not evaluative, as it does not aim to evaluate how effectively the teachers are applying the RAMP methods.
  - **Phase 2:** During this phase the coaches can have feedback for the teachers. As of spring 2018, this phase has primarily started in cohort 1 schools.
  - **Phase 3:** During this phase schools will be receiving differentiated support based on student performance. RAMP has been able to classify schools based on student performance using their Lot Quality Sampling (LQS) study. The classification includes A (teacher level support), B (community of practice level support), C (cluster-level support) categories.

Understanding of the Coaching/Mentoring Model

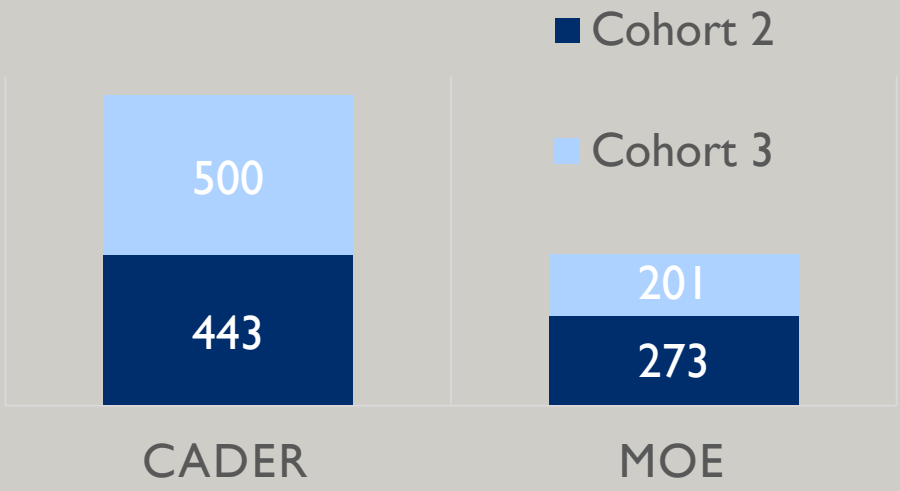
Awareness of 3-phased mentoring/coaching approach

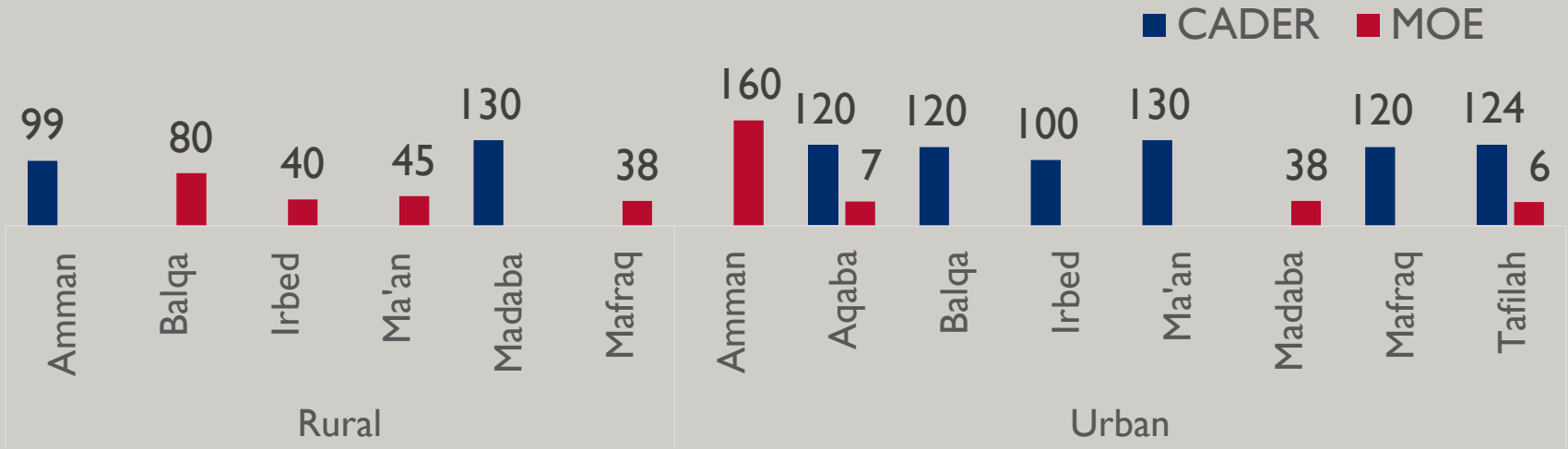
- **Teachers**:
  - 100% responses did NOT know
- **Mentors/Coaches**
  - 100% responses did NOT know
  - Some responded "yes" but then had incorrect explanations

Workload

- 1,417 teachers for 16 mentors/coaches
- Larger number of teachers for RAMP Coaches (their primary task)
- Total of 6,080 mentoring/coaching visits

Number of teachers covered by mentor type per cohort 2 and 3





Factors to Consider – Monitoring and Evaluation

- Lack of inter-rater reliability in coaches/mentor scoring and guidance to teachers
- Little or no utilization of the coaches/mentoring data to tailor trainings and/or inform guidance to teachers
- Use of "Coarse Grain" tools has not been systematically reviewed and followed-up on

Factors to Consider – Miscellaneous

- **Teacher Level**
  - Teacher workload
  - Teacher turnover (Teacher Observation Attrition)
  - Delays in teacher onboarding
- **Availability of Material (e.g. worksheets, photo copying etc.)**
- **Dosage – 10 Day Training**
- **Over-crowding**
  - Ability to apply RAMP approaches
  - Resources and time required

NEXT STEPS

Next steps *(in progress)*

- Analyses
  - Quantitative
    - Quality assurance and code review
    - Robustness tests and sensitivity analysis
    - Supplementary analyses
      - E.g. estimate impacts on students' zero scores
  - Qualitative
- Report writing

THANK YOU!

Jordan MESP

USAID
FROM THE AMERICAN PEOPLE

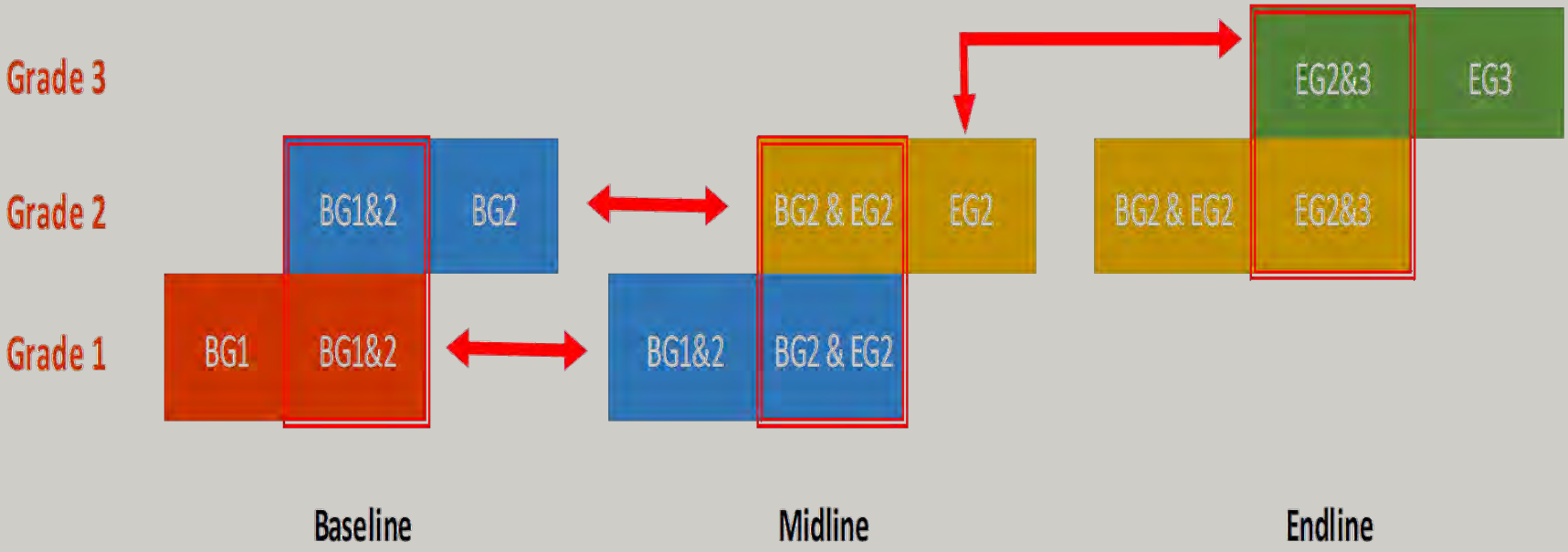
Why Test Equating?

- Grade appropriate and curriculum aligned instruments are developed for EGRA and EGMA
- Different instruments are used at baseline, midline, and endline; each with grade-appropriate, thus different, difficulty levels
- To estimate true change in student learning outcomes due to RAMP, difference in difficulty levels of the instruments at baseline-midline-endline should be taken into account to obtain equivalence of test forms through test equating
- "Equivalent tests forms refers to tests that are intended to be of equal difficulty (and thus directly substitutable for one another)." - EGRA Toolkit, 2016

Why Test Equating (continued)?

- In equating, the scores on one test instrument are matched to or paired with scores on another test instrument.
- Equating is a process conducted to establish comparable scores, with equivalent meaning, on different versions of test instruments; it allows them to be used interchangeably.
- "Equated test forms, therefore, refers to forms that have been adjusted by a statistical process following test administration to make scores comparable." - EGRA Toolkit, 2016

Assessment Design for Test Equating



Comparison of Scores (1): Syllable Segmentation

- This subtask has three anchor (common) items between Grades 2 and 3
- These anchor items help endline scores converted into midline comparable scores through test equating
- This subtask also had three anchor items between Grades 1 (BG2) and 2, and helped converting midline scores into baseline comparable scores
- All three sets of scores are on the same scale (i.e., vertically scaled), and can be used interchangeably

Endline G3	Midline G2	Baseline G1 (BG2)
0	0	0
1	2	3
2	3	4
3	4	5
4	5	6
5	6	7
6	7	8
7	8	8
8	9	9
9	9	9
10	10	10

Equating Method for Oral Reading Fluency and Comprehension (1)

- For oral reading fluency (ORF) and comprehension, a common person equating design is used
- To estimate true change in ORF and comprehension due to the RAMP, a two step process is implemented:

1. Estimate relative difference in difficulty levels of the oral reading passages and **comprehension items used at baseline, midline,** and endline

   At endline, both midline (41 words for Grade 1 and 52 words for Grade 2) and endline (52 words for Grade 2 and 60 words for Grade 3) oral reading passages and comprehension items are administered to additional samples of around 1000 grades 2 and 3 students

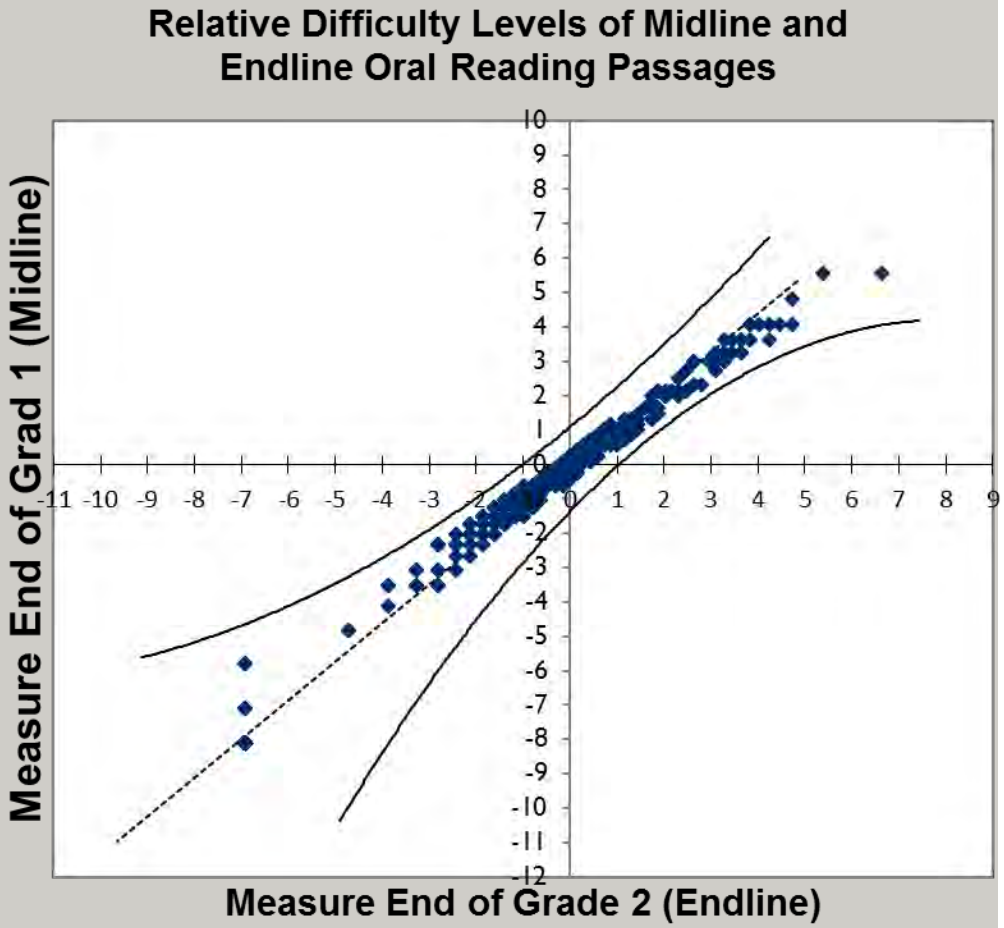
   We analyze the students' paired data using the IRT-based calibrations to create pairwise conversion tables for each oral reading fluency and comprehension scores

2. Difference in difficulty levels of the oral reading passages is accounted in the estimation of change using the conversion scores

Equating Method for Oral Reading Fluency and Comprehension (2)

Accuracy of conversion tables:

- Excellent model fit to the pairwise ORF scores
- The reliability coefficients for midline and endline scales are 0.94 and 0.92, respectively
- The correlation between midline and endline ORF scores is 0.995
- The endline passage was slightly more difficult than the midline passage for students with low reading ability



Equating Method for Oral Reading Fluency and Comprehension (3)

A pairwise conversion table for the ORF scores between endline (grade 2) and midline (grade 1)

- The passage is slightly more difficult at endline for students who obtain a score of 4 or 5
- Slightly easier at endline for students who obtain a scores of 36-40
- Much easier at endline for students with 40+ scores

EG 2	EG 1	EG 2	EG 1	EG 2	EG 1	EG2	EG 1
0	0	13	13	26	26	39	37
1	1	14	14	27	27	40	38
2	2	15	15	28	28	41	38
3	3	16	16	29	29	42	39
4	5	17	17	30	30	43	39
5	6	18	18	31	31	44	40
6	6	19	19	32	32	45	40
7	7	20	20	33	33	46	40
8	8	21	21	34	34	47	40
9	9	22	22	35	35	48	40
10	10	23	23	36	35	49	41
11	11	24	24	37	36	50	41
12	12	25	25	38	37	51, 52	41

Equating: Method and Rationale

- An Item Response Theory (IRT) based Fixed Common Item Parameter (FCIP) method is used for equating of the following subtasks administered at baseline, midline, and endline
  - For EGRA, phoneme isolation, syllable segmentation, reading vocabulary, and non-sense word (invented word)
  - For EGMA, number counting, enumerating quantitates, number identification, number discrimination, missing number, addition level 1, addition level 2, subtraction level 1, and subtraction level 2
- A subset of 30% items within each subtask for the adjacent grades are used as anchor items (i.e., common items)
- The subset of anchor items brings the adjacent grades onto the same measurement scale (establishes comparable scores)

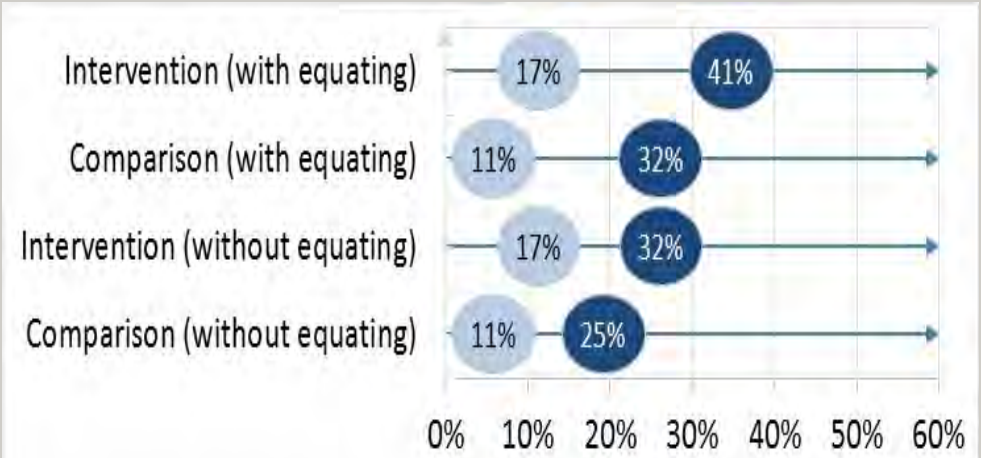
Comparison of Scores: Syllable Segmentation

- If equating were **not** conducted:
  - Measures at baseline, midline, and endline will be on different scales (not linked);
  - Students' true scores at endline will be under-estimated, particularly for students who obtain a score between 1 and 8
  - Difficult to interpret the results – change due to RAMP? or due to differential item difficulty?

	Comparison (without equating)	Intervention (without equating)	Comparison (with equating)	Intervention (with equating)
Endline	4.63	6.21	5.86	7.23
Baseline	3.11	4.91	3.11	4.91

Equating Method for Oral Reading Fluency and Comprehension

- If equating were **not** conducted:
  - Measures at baseline, midline, and endline will be on different scales (midline and endline %s have different denominators);
  - Students' true scores at endline will be under-estimated, particularly for students who obtain a score between 4 and 5
  - Difficult to interpret the results – change due to RAMP? or due to differential passage difficulty?

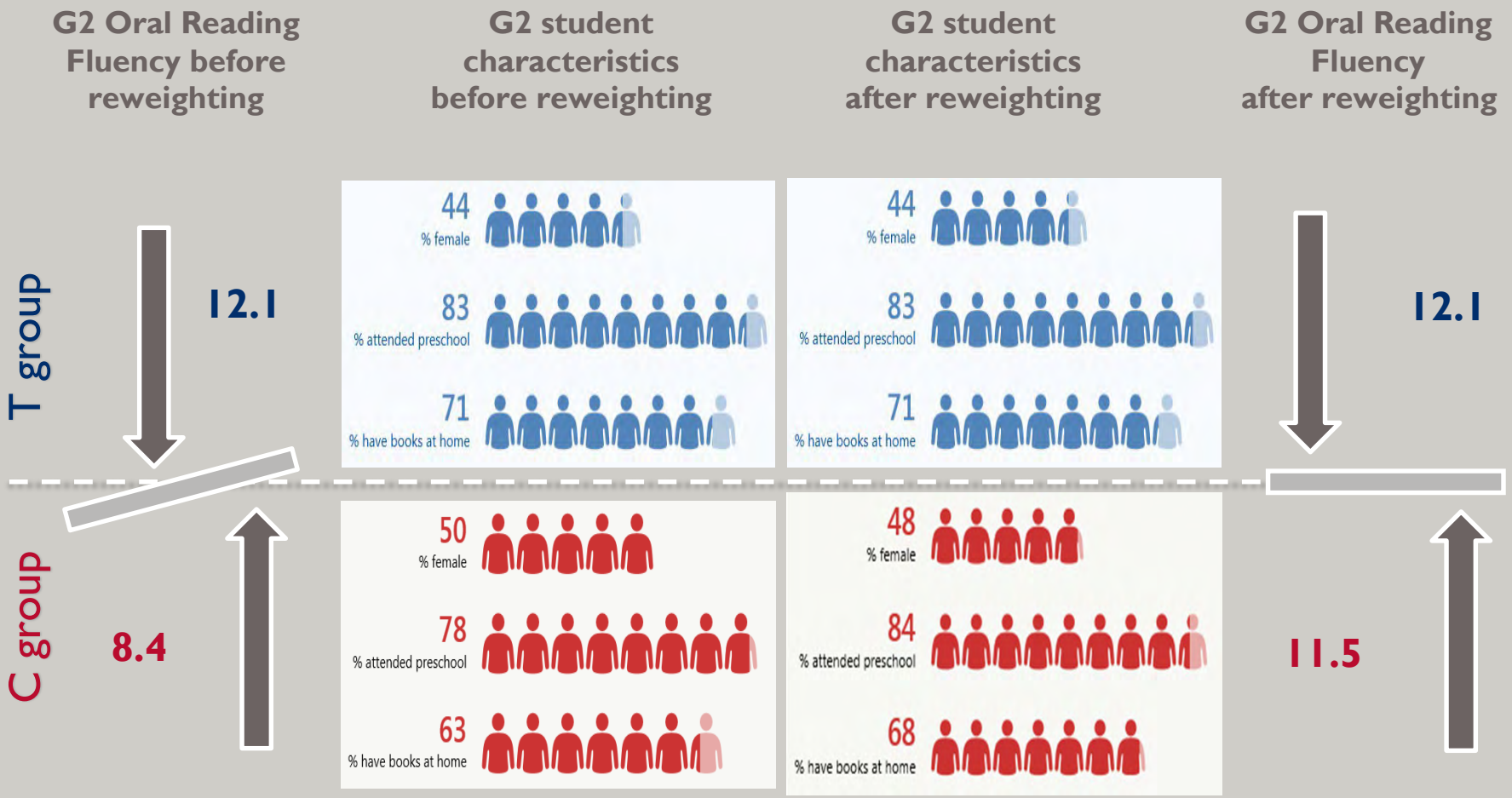


	Comparison (without equating)	Intervention (without equating)	Comparison (with equating)	Intervention (with equating)
Endline	25%	32%	32%	41%
Midline	11%	17%	11%	17%

Baseline Equivalency: Estimating Unbiased Impacts of RAMP

- Baseline Equivalency means that intervention and comparison groups were similar *or equivalent* at baseline
- This allows the attribution of differences in teacher and student outcomes—between the intervention and comparison groups—at midline or endline to RAMP
- Used **propensity score matching** and student **propensity score weights**
- Any differences in student outcomes at midline or endline can more confidently be attributed to RAMP, instead of pre-existing differences.

Weighting – Balancing the Groups



ANNEX U. JORDAN MINISTRY OF EDUCATION CONSENT LETTER

وزارة التربية والتعليم

الرقم ٤٢٠٣٩ \ ١٨٠/١٤
التاريخ ٢٩ ذو القعدة ١٤٣٧
الموافق ٢٠١٦/٠٩/٠١

السيد مدير التربية والتعليم

الموضوع: مشروع مبادرة تحسين القراءة والحساب للصفوف المبكرة (RAMP)/ دراسة تقييم أثر التدخلات

السلام عليكم ورحمة الله وبركاته، وبعد؛

فأرجو العمل على تسهيل مهمة فريق البحث المكلف من الوكالة الأمريكية للتنمية الدولية بجمع البيانات المتعلقة بدراسة تقييم أثر التدخلات الخاصة بمشروع مبادرة تحسين القراءة والحساب للصفوف المبكرة (RAMP)، وذلك بهدف توفير تقييم دقيق وموضوعي عن مدى مساهمة هذه التدخلات في تحسين مخرجات التعلم للطلبة، وممارسات المعلمين التدريسية في مرحلة الصفوف المبكرة التي استهدفها المشروع، علماً بأن فريق التقييم سيقوم بزيارة المدارس التابعة لمديريتكم الواردة في الكشف المرفق أثناء الفصل الدراسي الأول من العام الدراسي (2017/2016)، وسيتم خلال هذه الزيارات ما يأتي:

1. تطبيق أدوات تقييم مهارات القراءة والحساب(EGRA)، (EGMA) على عينة من الصفوف(1-3).
2. تطبيق الملاحظة الصفية على أداء المعلمين خلال حصص(القراءة والحساب) للصفوف(1-3).
3. إجراء مقابلات مع كادر المدرسة (مديري المدارس، المعلمين، إلخ)

واقبلوا الاحترام

وزير التربية والتعليم

نسخة: للمدير التنفيذي لوحدة التنسيق التنموي

محمد جمعة العكور
الأمين العام للشؤون التعليمية

المملكة الأردنية الهاشمية
هاتف: ٥٦٠٧١٨١ ٦ ٩٦٢+ فاكس: ٥٦٦٦٠١٩ ٦ ٩٦٢+ ص.ب: ١٦٤٦ عمان ١١١١٨ الأردن . الموقع الإلكتروني: www.moe.gov.jo

ANNEX V: TEACHER TRAINING AND MENTORING: A QUALITATIVE STUDY OF RAMP

Courtesy of Ms. Gaelle Simon at MSI.

TEACHER TRAINING AND MENTORING:

A QUALITATIVE STUDY OF RAMP

MAY 2019

This publication was produced for review by the United States Agency for International Development. It was prepared by Dr. Carolyn Fonseca and Ms. Afnan Al-Hadidi for Management Systems International (MSI), A Tetra Tech Company.

TEACHER TRAINING AND MENTORING

A Qualitative Study of RAMP

Contracted under AID-278-C13-00009

Jordan Monitoring and Evaluation Project

ACKNOWLEDGEMENTS

The authors of this report would like to thank the entire MESP team in Jordan for its unwavering support for this activity. We are most grateful to Mr. Bandar Al-Huniti and Mr. Amer Zidan for the many trips and long hours of qualitative data collection. In addition, the team is grateful for Ms. Idalia Rodriguez from the MSI education team, for her time and expertise reviewing the data. A special thank you to the Mathematica team lead, Dr. Candace Miller, for providing feedback and guidance on the qualitative research. Finally, the team is most grateful to Ms. Kenana Amin, the MESP Contracting Officer's Representative from USAID/Jordan, who supported this work. Her extensive knowledge and leadership ensured the RAMP impact evaluation could have additional data to help explain the findings, thus leading to more helpful recommendations for future iterations of RAMP.

ACRONYMS

ADS	Automated Directives System
CADER	Change Agent for Arab Development and Education Reform
COTI	Classroom Observation of Teacher Instruction
CWPM	Correct Words per Minute
DEC	Development Experience Clearinghouse
EA	Evaluability Assessment
EGMA	Early-Grade Math Assessment
EGRA	Early-Grade Reading Assessment
EMIS	Education Management Information Systems
EQ	Evaluation Question
FY	Fiscal Year
G1	Grade 1
G2	Grade 2
G3	Grade 3
GoJ	Government of Jordan
IE	Impact Evaluation
IP	Implementing Partner
IR	Intermediate Results
KG2	Kindergarten 2
KIIs	Key Informant Interviews
LQAS	Lot Quality Assurance Sampling
MESP	Monitoring and Evaluation Support Project
MoE	Ministry of Education
MSI	Management Systems International
M&E	Monitoring and Evaluation
PE	Performance Evaluation
PII	Personal Identifying Information
QRTA	Queen Rania Teacher Academy
RAMP	Early Grade Reading and Math Project

RTI	Research Triangle Institute
SOW	Statement of Work
USAID	United States Agency for International Development

EXECUTIVE SUMMARY

The purpose of this study was to assess the implementation fidelity of Early Grade Reading and Math Project (RAMP) in Jordan. The USAID/Jordan Mission requested additional qualitative work after the RAMP impact evaluation (IE) midline results were presented in August 2017. This was to better clarify "why" and "how" RAMP might be achieving/or not achieving the intended impact. In addition, USAID desired more information about the mentoring component of RAMP and the views by stakeholders such as teachers, mentors, and principals of this activity.

The study had three research questions, with a fourth question focused on recommendations to improve RAMP components. The evaluation team collected key informant interviews and observation data.

The evaluation team focused on two RAMP components: teacher training and teacher mentoring. Table 1 summarizes the key findings and conclusions by research question, reflective of the opinions of the respondents in the sample and observation of training modules, RAMP routines, and mentoring sessions.

TABLE V.1. SUMMARY OF FINDINGS AND CONCLUSIONS

Research Question	Intervention Areas	Key Findings and Conclusions
1: ADHERENCE ***Was the planned intervention for teacher training and mentoring implemented to the RAMP design specifications?***	Teacher Training	• Teacher training was implemented in accord with the design of 15 days, but dosage on math varied. • Length of training (intervention) was too short, according to responses from teachers and observation of training modules.
	Teachers' Application of RAMP	• Teachers found the implementation of routines difficult. • RAMP increased workload on teachers without providing additional teaching time. • No incentives were provided to teachers to carry out RAMP and increase workload. • Limited resources were available for teachers and schools to implement RAMP. • Limited class time was available to implement RAMP due to class crowding. • RAMP was not officially included in the curricula as a teaching method by the Ministry of Education (MOE).
	Mentoring	• Based on observation and key informant interview (KII) data, mentoring was not implemented according to the Research Triangle Institute (RTI) plan. Teachers received much lower dosages compared to the design of 12 visits per year (RTI had designed their intervention so that teachers were to receive 6 mentor visits per semester for a total of 12 per year). • More mentors were needed than anticipated and additional staff were added, causing variance in dosage of mentor visits.

Research Question	Intervention Areas	Key Findings and Conclusions
		• Expectations around mentoring process and impact were designed and changed half way through the implementation of RAMP by RTI ,the implementing partner (IP). From the implementation process, the IP identified a much larger need for mentors, and increased their number (which included both MOE and Change Agent for Arab Development and Education Reform (CADER) staff). This was done because of the high load per mentor, making it difficult to achieve the 6 visits per semester per teacher, as called for by the RAMP design. • According to KIIs and observations, teachers/mentors/principals were not clear about the mentoring process as per the new 3 phased approach designed by the IP. • Coaches and mentors encountered barriers to mentoring visits.
	Community Participation	• Almost no parental or community participation for RAMP was identified by teachers and principals, but this was highly desired by these two groups. • Principal involvement is key to community participation according to participants. • The IP has conducted more efforts in this component for RAMP, but as of the evaluation team's data collection these outcomes had largely not been included in RAMP's formal performance reporting.
	Monitoring of RAMP	• Tracking and monitoring processes for RAMP were not robust, and the feedback loop in the design was not fully implemented. However, the IP has made significant revisions to meet the scale of the activity. The IP has created a more comprehensive and complex monitoring and evaluation (M&E) system to reflect the needs of the intervention, with protocols for field data collection. • This study found limited use of monitoring data (e.g., sign-in sheets for training, mentoring reports by coaches and mentors) even though interviews with partners suggested the IP collects many data streams. While internal analysis may exist of these streams, this evaluation team did not review these.
2: EXPOSURE/ DOSAGE ***What, if any, were barriers in the full implementation of these two training elements (teacher training and mentoring) that could potentially affect, dilute,***	Teacher Training	• Teachers received the training dosage of three weeks per the RAMP design. • The actual dosage of math instruction, as a RAMP routine, was found to vary in trainings. Not all teachers received a full dosage of math. o At the onset of RAMP there was likely an underlying assumption about teacher math skills being similar; data from this study suggest otherwise.

Research Question	Intervention Areas	Key Findings and Conclusions
or diminish the effectiveness of RAMP on students?		• Incentives for teachers to implement RAMP are not aligned with MOE curricula and therefore some teachers may be implementing RAMP less than others.[14] • Limited access to resources and materials was constraining teachers' ability to implement RAMP in the classroom. This likely led to variations in RAMP dosage to students. • Interview data suggested use of both coarse and fine-grained tools.
	Mentoring	• Planned dosage of 12 visits per year was not fully implemented; rather, mentoring dosage was reduced. • On average, the data suggests teachers received 2-4 mentor visits per year. Some cohorts may have received more mentoring exposure than others. • Mentoring visit quality varied among teachers, likely affecting level of support and therefore mentoring dosage. • Implementation level of the mentoring stages is unclear; therefore, there is limited data on mentoring effectiveness.
3: PARTICIPANT RESPONSIVENESS ***What were the perceptions by stakeholders of these two training elements?***	Teacher Training	• Overall, teachers view the RAMP training very positively and seem to want in-service training to continue. However, for RAMP to continue, teachers noted a need for additional support in materials, teacher aids, and other in-class resources. • At the start of RAMP, incentives for teachers to implement RAMP were not aligned with MOE curricula. However, efforts by the IP with the MOE have been made to formally integrate RAMP into the teaching requirements. • Interview data showed some use of the coarse- and fine-grained tool, albeit with great variance across the schools. The IP may wish to further study this variance which could highlight further the diminished effects of RAMP, as of Spring 2018. • Principals have positive views, similar to the teachers, about RAMP. • At the onset of RAMP, there was likely an underlying assumption about teacher math skills being similar; data from this study suggest otherwise.
	Mentoring	• Perceptions by teachers regarding the mentors varied by mentor type. MOE mentors were viewed with more authority and RAMP coaches were viewed as supportive colleagues.

[14] At the start of the study, the IP was working with the MOE to incorporate RAMP officially into the curriculum. According to the interviews with the IP and other stakeholders, this process can be lengthy and complex. However, by the end of this study, the evaluation team was informed RAMP would be officially in the teaching curricula in the coming year, 2019.

Research Question	Intervention Areas	Key Findings and Conclusions
		• Implementation level of the mentoring stages is unclear, and therefore there is limited data on mentoring effectiveness. • There was an overall positive view of mentorship by teachers. • Mentors identified barriers to mentoring teachers, such as mentee load, travel/logistics problems, and teacher mobility. The supervisors reported having more resource barriers than the RAMP coaches, who are supported directly by the IP for mentoring work.

STAKEHOLDER BUY-IN

Although not officially part of the research questions for this study, data suggests there is clear and strong buy-in of RAMP by most of the stakeholders. This achievement cannot be undervalued. Stakeholder buy-in is one of the primary reasons many programs sustain over time. The IP also has accessed strong partners able to locally support RAMP, including the MOE, Queen Rania Teacher Academy (QRTA), CADER, and others who worked to launch RAMP nationally.

RECOMMENDATIONS

- **Create materials for teachers easily accessible in multiple formats,** including electronic tablets, paper copies, and supplementary manuals to support RAMP implementation.
- **Conduct internal assessments of the data collected** to identify patterns and issues of RAMP implementation. This can help guide allocation of future resources.
- **Provide math training to teachers prior to RAMP training**, so when receiving the RAMP dosage, trainers can focus on routines only. More broadly, assess teacher quality prior to trainings to provide tailored trainings, where feasible.
- **Test larger dosages of training on selected teachers** to see if this affects student performance.
- **Add informal learning channels for teachers** to pursue additional capacity building, such as learning working groups through WhatsApp, online YouTube training videos like Khan Academy, and webinars from QRTA in math (a subject area for which additional support was requested).
- **RAMP should restructure the supervising/coaching component to give teachers more one-to-one time with a mentor** regardless of the number of visits thereby increasing the quality of the visit. The current model of mentorship, according to conversations with the IP, is based on using the mentor visit as more of a "check-in;" the

presence of the mentor is to signal the teacher that they will be monitored for the implementation of RAMP to incentivize the application of RAMP. However, other effective models of mentoring are based on the quality of the visit and not the number of visits by the mentor. If the IP focuses resources on improving the quality of the visit, the mentoring dosage may be more effective than by continuing the current model – focused on number of visits regardless of the time or quality of the visit. This could in turn improve the impact of RAMP on students.

- **Test other capacity building models of RAMP.** The cascade model used in RAMP may be less effective when the underlying assumptions (i.e., equal dosage to all teachers, equal levels of capacity at the start, equal levels of resources/incentives by teachers to apply RAMP, similar views of RAMP, similar patterns of implementation by teachers) do not hold.
- **Improve training of coaches and mentors to ensure a more standardized approach to coaching and mentoring**. This should include a more uniform understanding of the teacher observation process, including how the observation is done, how the observation data is entered, as well as the approach to providing feedback to teachers. The observation data needs to be analyzed more effectively and in a timely manner to identify specific areas/approaches that need to be improved.
- **Create a plan and process to document the changes to the intervention during the lifetime of the activity, allowing for assessment of the fidelity of implementation over the course of the RAMP intervention.** Many of the changes to the RAMP design are part of the learning and adaptation most projects encounter during large, national scale-up processes. However, when changes are substantial, the actual original intervention may no longer be attributable to the observed impact. The IP can add the fidelity component part of the M&E plan so as to minimize costs and burned on staff.
- **To better estimate the implementation barriers and possible changes** required to meet the program design and intervention, this study suggests USAID consider, for large scale-up programs derived from pilot studies, conducting a needs assesment prior to the launch of the national activity. This process could help identify unforseen issues affecting the dosage or exposure times of the intervention due to scalee-up challenges.

INTRODUCTION

To further understand the findings of the MESP RAMP Impact Evaluation (IE), the evaluation team added a qualitative study.[15] This report presents findings on the fidelity of implementation of RAMP and recommendations for improvements.

Fidelity of implementation assessments often highlight whether a project outcome can be attributed to the intervention itself or its implementation (Fidelity Ontario report, 2013). Studying the implementation process of a program is critical to understanding *how* and *why* and under what conditions an educational intervention is effective/impactful (Century and Cassata, 2014).

> *"Fidelity studies can help prevent potentially false conclusions about an*

For this report, *implementation fidelity* is defined as the level to which an intervention has been delivered as planned (Carroll et al., 2007) or the extent to which an intervention has adhered to the program/model/protocol (Bell, 2009). Researchers focus on five dimensions when reviewing fidelity of implementation: adherence, exposure, quality of delivery, participant responsiveness, and program differentiation (Bell, 2009). This study examines **adherence** and dosage/**exposure**, and also looks at **participant responsiveness/perspectives**.

During implementation, programs often undergo changes due to unanticipated real-time events. This can be problematic for programmatic fidelity measurements. At the start, projects/programs have continuous quality improvements and rapid- cycle problem solving (Kershner, et al., 2014; Metz and Barley, 2012). To the extent possible, the evaluation team attempted to capture the changes made to RAMP over time, albeit recognizing not all changes may be documented. There is value in mapping the implementation process and understanding the areas of divergence from the intended plan. Implementation fidelity assessments are often conducted to understand program implementation, examine theoretical assumptions, interpret outcome findings, provide feedback for continuous quality improvements, and provide feedback to program developers about the program (Bell, 2009).

PURPOSE

The purpose of this assessment is to provide insight into where changes/deviations to the implementation of the RAMP intervention occurred and unpack the potential effects (positive or negative) these changes could have had on the target outcome – student performance. This study also facilitates understanding of the IE results and gives USAID and the IP actionable recommendations for changes to teacher training/mentoring – the key RAMP intervention component causing change in student scores.

[15] In accord with the request of the USAID Jordan Mission, the team has expanded the qualitative research to assess RAMP's fidelity of implementation and mentorship component. This was added late 2017 and expanded in January 2018 to include mentoring in the scope of the activity.

RESEARCH QUESTIONS

This qualitative study addresses the following questions:

1. Was the planned intervention for teacher training and mentoring implemented to the RAMP design specifications? (**Adherence**)
2. What, if any, were barriers in the full implementation of these two training elements (teacher training and mentoring) that could potentially affect/dilute/diminish the effectiveness[16] of RAMP on students? (**Exposure/Dosage**)
3. What were the perceptions by stakeholders of these two training elements? (**Participant Responsiveness**)
4. What are suggested recommendations to improve the RAMP training components?

This report is divided into the following sections: 1) introduction presenting the purpose and research questions; 2) background describing the implementation of the RAMP intervention as designed by RTI; 3) methods and data section of the assessment; 4-6) findings for research questions 1 to 3; 7) conclusions and recommendations; and a reference and annex section.

BACKGROUND

The RAMP intervention hypothesizes that in-service training of teachers will improve student performance in reading and math, specifically for grades 2 and 3.[17] Although other elements support factors in the intervention, such as parental and community support, the primary treatment is teacher training. In-service training is comprised of: 1) a workshop-based training, and 2) in-class mentoring by MoE supervisors and/or RAMP coaches.

The workshop training covers three modules, two of which are taught over the course of two weeks before the start of the school year. Module 3 is presented during the holiday period between semester 1 and 2. According to RTI, RAMP Module 1 measures students'

IP's description of the RAMP model in their midline survey report:

"By investing in building MoE staff capacity, especially that of public school teachers and supervisors, to use appropriate materials; research pedagogies; and differentiated support to students according to their needs, RAMP will contribute to a substantially higher proportion of grade 2 and grade 3 public school students being able to read with comprehension and do mathematics with understanding by the end of the initiative. By also involving parents and communities in general in support of RAMP, the impact of the initiative will be significantly enhanced. Also, these gains will be sustained and built upon beyond the life of the initiative through institutionalization of RAMP's research

[16] Effectiveness in this context is defined as the ability of RAMP to implement fully the designed intervention and reach the outlined student target goals.

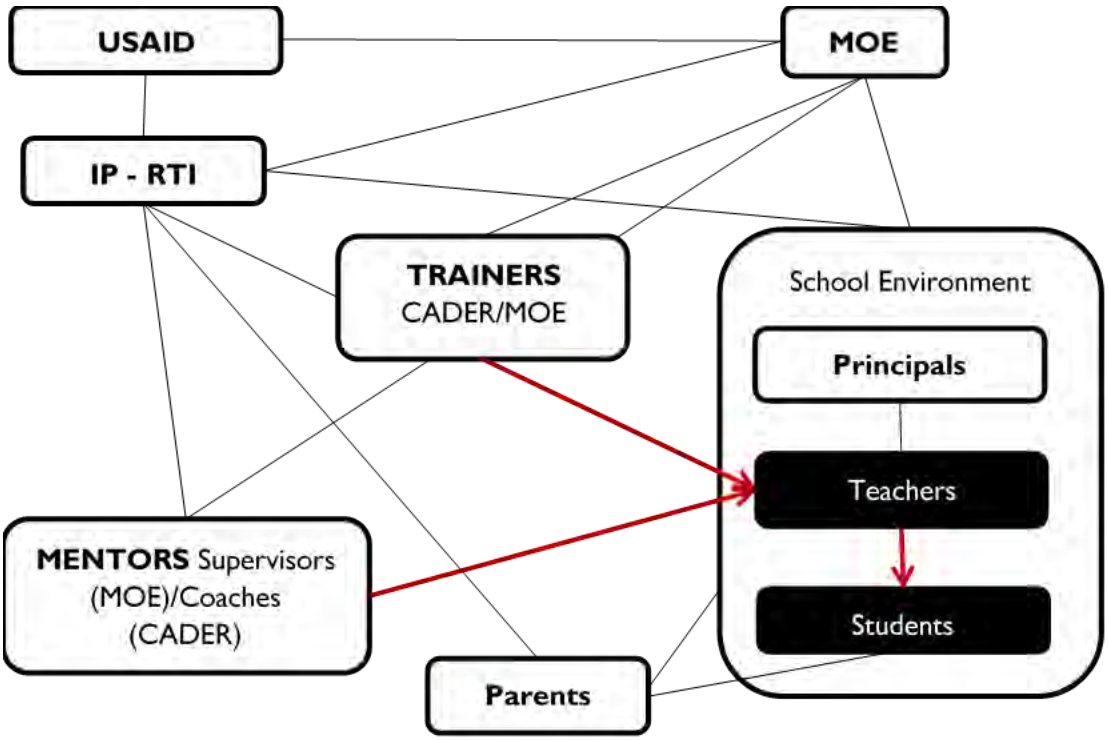
[17] Detailed description of the theory of change for RAMP by RTI can be found in their report, "Early Grade Reading and Mathematics Initiative (RAMP): 2017 Midline Survey Report."

foundational skills in reading and how to effectively implement these for students to improve reading comprehension. Module 2 focuses on mathematics and has the same goal as Module 1.

Per the proposed design, the full implementation of these two training components would result in changes in student performance in math and reading.

Figure 1 depicts the causal model implied by RTI's results framework. The red arrows from mentoring and training to teachers depicts the RAMP intervention with teachers to improve teaching techniques, which would cause improved student performance. The black lines denote the connections between the various stakeholder groups. These include the IP working with the MOE to integrate RAMP into the official curriculum and meeting with parents/communities to share the RAMP intervention effort. During implementation, the RAMP model underwent several key changes through internal learning and to overcome unanticipated obstacles. The following chapters describe these changes and their possible effects of RAMP to date.

FIGURE V.1. MAPPING THE RAMP INTERVENTION



Note: Red arrows denote the ramp intervention components (including mentoring and training) on teachers. Black boxes represent intervention target groups.

METHODS AND DATA

The team collected data through interviews and observations. Primary data was collected through interviews and from observations. The team also reviewed secondary data, mainly mentoring data shared by RTI. Comments on this dataset are presented in the findings section.[18]

INTERVIEW DATA

The evaluation team's key informant interviews (KIIs) were conducted with teachers, principals, mentors (coaches/supervisors), trainers, implementing partner RTI, and RAMP partners CADER, Dajani, Queen Rania Teaching Academy (QRTA), and Kaizen.

Starting in the summer of 2017, the team began interviews with key stakeholders. Table 2 lists the types of interviews conducted. The evaluation team worked with the IP to request interview times with their staff and partners. Although the team planned to interview USAID and the MOE, per USAID these interviews were cancelled due to limited staff availability. In addition, the team conducted interviews with teachers, principals, and mentors (both from CADER and MOE). Each person was contacted via phone and or email to schedule an interview. Interviews were conducted in person and in some cases two team members were present to help with note taking.

The sample was a convenience sample, although the team did attempt to capture views from all regions. The team worked with the IP to identify individuals who were available and who were directly involved with RAMP activities. For the mentoring sample, that included pairing of mentors with their respective teachers and respective principal. This pairing approach was selected in hopes of conducting analysis across the mentor-mentee pair. However, after data was collected, the team decoupled the data to protect the personal identifications of the mentors (as these are relatively few) and the teachers. For the pilot interviews of teachers and principals, conducted in late December 2017/early January 2018, the team selected individuals who were easily accessible (to reduce costs) and schools willing to participate.

For the teacher/mentor/principal interviews (i.e., 16 pairs of mentors-teachers), this sample included individuals who were scheduled for a mentoring session during the time of data collection (spring of 2018). The teacher and mentor both had to be willing to participate in the interviews to ensure matching pairs. This resulted in 16 pairs of mentor-teacher-principal interviews and matching observation data points. The pairs included schools from all regions (North, Center, and South).

Analysis of the interview data was done through content analysis and a summary of the comments by type of respondent per research question. Many of the participants can easily be identified

[18] Due to data limitations, the evaluation team could not conduct a full analysis of the RTI mentoring data. However, this report provides some recommendations to help improve the future utility of the data.

from the content of the interview; therefore, to ensure their anonymity, the evaluation team collapsed responses.

TABLE V.2. KII INTERVIEWEE CATEGORIES AND NUMBER OF INTERVIEWS COMPLETED

Interviewee Category	Description	Number of Interviews Completed
Implementing Partner	COP, Technical Director, Monitoring and Evaluation Technical Staff, Community Leadership Staff	4
Partners	QRTA Math Specialist, QRTA Reading Specialist, QRTA curriculum specialist, Dajani Technical Lead, CADER staff	5
Teachers	Pilot interviews to assess qualitative questions for future teacher/mentor/principal interviews	22
Principals		8
Teachers[19]*	Teachers paired with principal mentor data from all regions of Jordan (North, Center, South)	16
Principals*		16
Mentors	Supervisors (MOE) and RAMP Coaches (CADER)	16
Total KIIs		87[20]

OBSERVATION DATA

Mentoring: The team observed 16 mentoring pairs (teacher and mentor) including pairs with MOE and CADER. Of these, 50 percent had RAMP coaches from CADER and the other half were MOE supervisors. Observation data was collected in the spring of 2018. Principals of the respective mentor-mentee pair observed were included in the KII sample.

Training: In the summer of 2017, the team observed trainings in schools from the three regions. Over a period of a week, the team sat in on teacher training sessions. During these meetings, the team members were able to observe the trainer provide instruction on RAMP for math and reading. The team was able to observe both MOE and CADER staff trainers. In addition, for some of the schools the team members were able to talk informally with teachers, principals and trainers to gather additional views on the RAMP training component.

For training observation data, team members (1-2 people maximum) sat in the back of the classroom and observed the teacher and mentor. The team only wrote notes and waited to ask any clarification questions at the end of the class to reduce disruptions. There was no formal instrument used to observe training as the team had short notice on the location and time of trainings. The team was able to talk with teachers, trainers and principals after some of the training sessions to discuss RAMP components.

For mentoring observation data, team members were sent to schools to observe teachers and their RAMP mentors in real-time (see Annex B for a copy of the instrument). The observer would

[19] The teachers and principals with "*" are reflective of the unique sample where the evaluation team observed the pair (teacher-mentor) during the mentoring process. They were also interviewed.

[20] Several informational interviews were conducted with teachers and principals during the collection of training observation data by the evaluation team. However, their conversations were not formally recorded, but were used to inform design of the formal KII guides.

sit in the class and observe the teacher and mentor interaction. In some cases, the observer was able to talk with the mentor after the mentoring session and gather additional feedback about the mentoring session.

DATA COLLECTION INSTRUMENTS

Instruments were created and shared with USAID and RTI for comments. See Annex A, B, and C for the instruments used to gather qualitative data for this report.[21] Training observations were compiled both informally as part of a scoping field activity and formally by the team for mentor/teacher pairs.

HUMAN SUBJECT PROTECTION

All participants were asked for consent before each interview and for permission for observers to gather data on training of teachers and mentoring visits. Each participant was verbally read their rights to decline participation, to stop at any time of the activity, to decline specific questions, and to ask questions at any time of the activity. Personal identifying information (PII) has been removed from the data to keep responses anonymous in the resulting data set.

LIMITATIONS

Sample selection was not random, although efforts were made to ensure the list of potential interview respondents included stakeholders from the three regions. As with most qualitative studies, samples are not always representative of the popluations and therefore results are not statistically generalizable. This was a convenience sample which included schools that were available during the time frame for data collection, and whose principals were part of ongoing data collection efforts for the impact study. The sample is further limited by not having data from schools in cohort 1 – reflective of the impact study which only included cohorts 2 and 3. Therefore, this study's responses are reflective of views by participants in cohort 2 and 3 only. Response bias is likely present in the data, but attempts were made to mitigate this issue through piloting of the instruments/interview questions.

RESEARCH QUESTION 1 (RQ1) FINDINGS

Question: Was the planned intervention for training implemented to the RAMP design specifications? (Adherence)

[21] Note: for Annex A, an example of the KII guide for one of the interview groups has been provided. The evaluation team used several other very similar protocols, one for each KII target group.

Results for this research question focus on the following areas of intervention:

1. Training of teachers
2. Teachers' application of RAMP
3. Mentoring[22]
4. Community Participation
5. Monitoring of RAMP

Adherence: The extent to which specified components were delivered as prescribed in the program design (Century

Findings:

- **Teacher training was implemented according to the design of 15 days**. Observation data found some trainings to spend more time on math instruction, but less time on RAMP math routines.
- Interviews with the IP revealed the reason behind the **length of the trainings**, as a reflection of the constraints by the MOE due to availability of free time by teachers during their summer break. Although the IP would have desired a longer training period for RAMP, the negotiated time was the only amount permitted.
- **Partners to RAMP (QRTA, Dajani and CADER) suggested more coordination and transparency** is needed, including more in person meetings. Since RAMP has complex activities depending on each other, the partners desired additional information be shared with them more frequently on the status of each activity. This would permit them to better manage their tasks and ensure they were aligned with the larger programs of RAMP as a whole.
- **According to teacher interviews, when they first encountered RAMP, teachers found the implementation of the routines difficult** due to limitations in the availability of paper and photocopying access.
- **Mentoring was not implemented in accord with the plan.** Teachers received much lower dosages (see the RQ2 results) than planned. There were far larger numbers of teachers trained and therefore requiring a mentor. The IP increased mentor numbers through the addition of CADER RAMP coaches to existing MOE supervisors who were also mentoring teachers.
- **The planned feedback loops within the RAMP design were not as informative as**

RQ1: ADHERENCE: CONCLUSIONS:

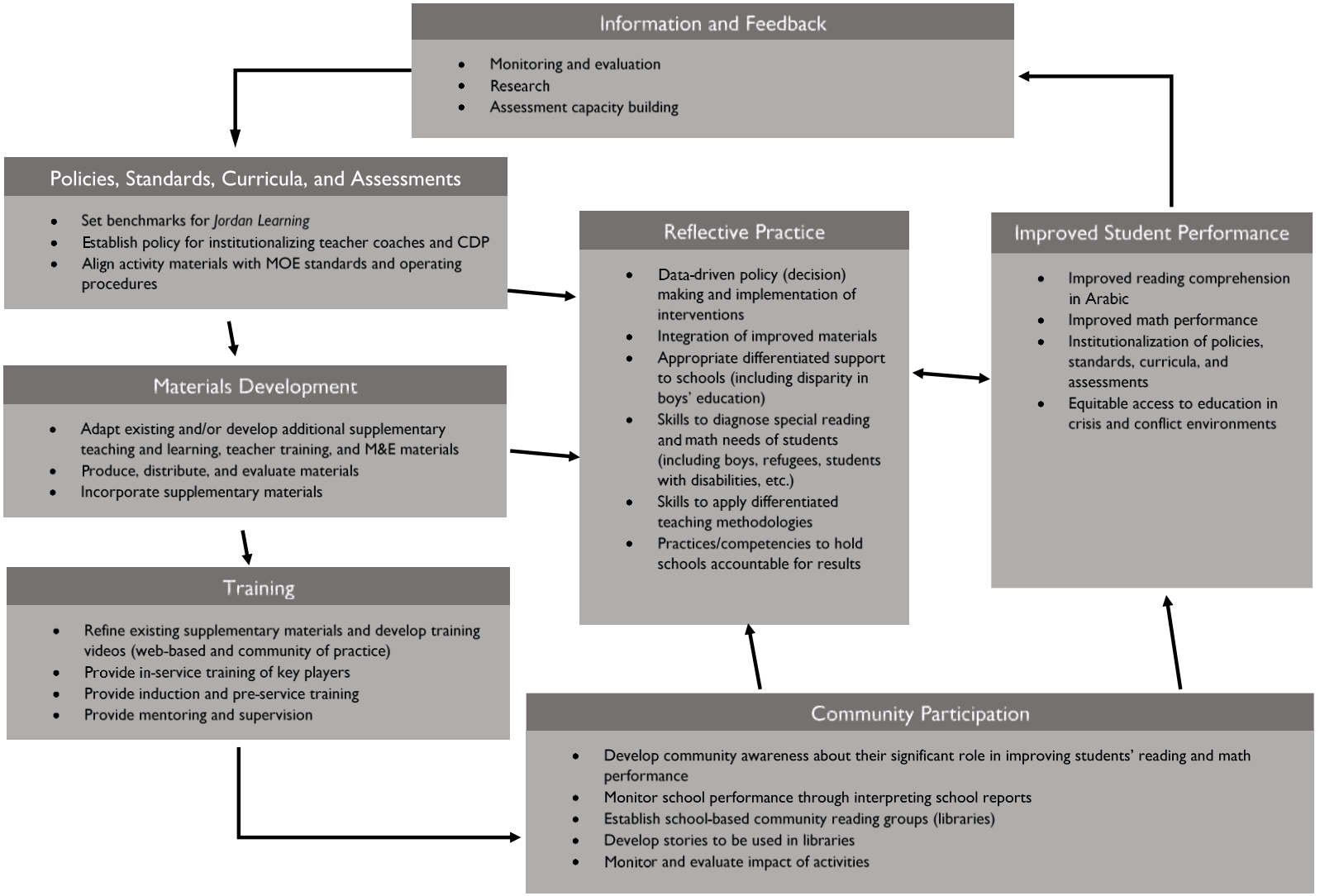
- Even though teachers attended training per the planned design, the content covered may have been more for some teachers and less for others. These differences may be accounting for observed variance in RAMP impacts. The dilution of the RAMP modules could also be affecting student outcomes.

[22] This study focused on assessing the views of teachers, mentors, and principals regarding the RAMP mentoring component. However, the timing of the study, budget, and staffing constraints (due to field work for the IE), limited data collection around the trainings of mentors. The team was not able to observe that process in the field. Questions for the mentors, CADER, and the teachers (receiving the mentoring) focused on the process of mentoring teachers in class.

- Limitations in resources for teachers could have also contributed to less-than-expected effects at the start of RAMP back in 2016. Although this issue has been corrected, and teachers are being, or have been, provided additional materials, there may be a lag effect not yet accounted for in existing studies (that is, the impact evaluation study and this qualitative study).
- There is substantial data gathered by the IP, both known and unknown to this team, which may provide additional information about the effects of RAMP.

Data suggested that all intervention areas were implemented. However, adherence levels varied greatly in areas 2 (teachers' application of RAMP) and 3 (mentoring). Area 4, the community participation component of RAMP, was found to be the least documented component, whereas the monitoring of RAMP was the area with the most need for improvement. Details are provided below. Figure 2 presents the original model for RAMP as shared with the team by the IP.

FIGURE V.2. ORIGINAL RAMP RESULTS FRAMEWORK[23]



[23] This results framework was provided by RAMP.

1. TEACHER TRAINING

The training component of RAMP, per the design, started with training of trainers by the IP. These included supervisors from MOE as well as education staff from CADER. The two groups trained together as one group. The trainers were then sent to provide the following to all teachers:

TEACHER TRAINING

- In August of each year, teachers received a 2-week workshop covering reading and math RAMP routines.
- During December of each year (mid-point in the academic year), teachers received 1 week of additional training.
- In some cases, a third session was offered, to those who might have missed the Summer/August workshop. Per conversations with IP staff, this session, if offered, was done in November, and was not the length of the regular 2-week session in the summer.

Interview data from the IP suggested there was a significant amount of effort put into negotiating the time of the training with the MOE. The length of the training, which affects dosage, was not fully set by the IP but by limitations in teacher's time, costs to the MOE, and other possible conflicts. The IP would have preferred a longer training period for teachers.

The IP, as of 2017 (per conversations with IP technical staff), monitors teacher attendance to trainings. Each teacher signs in daily to training; these data entry points are then logged by partner staff on paper forms and then entered into the monitoring system at RTI.

PARTNERS IN TRAINING

- RTI interviewees discussed how each partner focused on a component of RAMP for training. QRTA worked with the IP on the training curriculum and materials for math and reading instruction. Kaizen partnered with RTI to work on community outreach. Another of the IP's partners worked on the library activity, CADER provided support for training, and Dajani was responsible for data capture including administration of midline and endline survey.
- When the team spoke with the various partners about communication with the IP, some of the partners felt there was high quality interaction in terms of frequency and topics – both CADER and QRTA expressed positive views on their interaction with RTI. Other partners provided mixed views, suggesting the number of meetings was too infrequent and they did not always feel they were informed of RAMP's progress.
- According to RTI staff, teams meet regularly, including staff in the regional offices.[24] All partners did express a desire for increased transparency about the progress being made,

[24] RTI had staff in three offices covering the three regions: North, Center, and South.

and to have additional meetings where all partners were present together (based on interview data with all partners including the IP).

- QRTA shared a desire to know more regarding how RAMP was doing in the field, and potentially to attend and visit classes to see real-time math and reading routines. The interviews from QRTA highlighted an interest in having more connection to the training materials in the field, compared to just their design. This would help them revise and improve them with in-field perspectives.

TRAINING CURRICULUM

Interviews with the IP, CADER, and QRTA staff suggested the curriculum for RAMP training was created rigorously and underwent various iterations. Respondents involved in the creation of the training did not seem to have issues with the final training modules. By permitting input from various technical groups (i.e. MOE, QRTA, and other), RAMP's training curriculum was vetted significantly, ensuring potential issues were corrected. Also, the IP benefited from the pilot which had already tested the version of the training with a large sample.

RAMP was structured around the design's intermediate results (IRs): material development, training/coaching, community and parents, and institutionalization which was a crosscutting IR. However, RAMP was not officially integrated into the teaching curriculum. The IP shared steps they had been taking to address this issue:

- Discussions with the MOE to integrate RAMP officially into the Jordanian curriculum – a process that takes time due to all the approvals required;
- The creation of a teacher's guide from the MOE as currently there is no teacher's guide. This specifically is to help teachers better utilize the coarse and fine grain tools; and
- Using other monitoring tools like lot quality assurance sampling (LQAS) to allow review of RAMP results more frequently.

What seems somewhat unclear is how training changed for each cohort, and/or whether it remained the same. The interviews with the IP suggest the training and instruments used for RAMP were kept the same for all cohorts. However, RAMP added the additional training session in November for those who missed RAMP during the regular workshop period before the start of the school year (for new and transfer teachers).

Barriers to Adherence to RAMP Design in Teacher Training:

The following is a summary of barriers around teacher training identified during the study.

"Given the results [at midline], refocus activities to what makes a difference, we are covering too much…we have pre-service training, struggling with universities, in charge or pre-service for early grades, so I recommend we focus more on

- Length of the training was only 10 days (with a few additional days throughout the academic year for additional reinforcement of the intervention). This report discusses (in question 2) the variations around dosage.
- Another limitation to RAMP training may be due to the distribution of staff covering too many tasks, which were not anticipated in the scope of the original project. For example, the size of work on KG students (not reported in the RTI midline survey) increased workload for IP staff, decreasing time available for RAMP activities (interview data with IP).
- RAMP is not an official MOE routine. This issue affects the implementation of RAMP (teachers' adherence levels to implementing RAMP as designed and thus RAMP dosage levels to students). If schools do not feel it is required, teachers will focus on what is officially in the curriculum. Teachers are already burdened heavily, and interviews with them suggested limited time to add additional routines.
- Data suggests teachers were not always able to grasp the RAMP routine concept during training. Trainers also sometimes had some difficulty giving RAMP instructions. During classroom visits of RAMP training, observation data showed instructors in a few instances giving math instruction, versus giving teachers skills to teach math. In other words, the trainers were teaching the subject of math. Conversations with trainers suggested they often needed to first help teachers understand the math portion, before RAMP training routines could be given. This was more often the case for grade 3 teachers, as the math is more advanced. Teachers shared with the evaluation team that when attending school,

2. TEACHERS' APPLICATION OF RAMP

Interview data found variance on the application of RAMP in the classroom. Conversations with the IP suggested teachers would be more likely to implement RAMP if this was an official part of the curricula and if doing in-service training would count towards promotion. Teacher interview data suggested teachers do RAMP routines to varying degrees. Some teachers spend the full allocated time to the routine as per the design, while others felt class size and MOE curricula requirements limited time for RAMP. Other teachers did not have the materials to carry out RAMP in class. The IP has resolved this issue for cohort 2 and 3 teachers through the printing of RAMP materials. Adherence to the planned RAMP activity may have been less for cohort 1 when compared to cohorts 2 and 3; this could lead to lower dosage levels of RAMP on children.

During visits to training sessions of RAMP in 2017, the team observed a few cases where math instruction was focused on the subject matter and not the RAMP math routine. It may be math routines are applied less by teachers who struggle with the subject. Identifying these teachers who require additional support, would help the intervention be more effective.

Principals were asked during interviews to share their views on what part of RAMP might be more difficult for teachers to implement, reading or math sections. Of the 16 principals interviewed, 50 percent said math was the harder RAMP routine to implement for teachers. About 13 percent said reading was harder than math, and the rest felt both routines were equally hard for teachers to apply.

Teachers were asked during interviews if they had implemented RAMP, and most said they were implementing RAMP more now than before. Several teachers responded that RAMP was slightly difficult to implement at first but now they felt more comfortable doing the routines. Teachers were also asked if there are more difficulties implementing by subject matter or whether the student affected their instruction of RAMP. Only 1 or 2 teachers of the 16 interviewed in the mentor-teacher pairs suggested student type affected how they applied RAMP. These teachers discussed level of the student could affected what parts of RAMP could be applied to the student. Most teachers found both subject matters equal in terms of implementation difficulty. When talking to principals and teachers and trainers, the team was informed that teachers sometimes struggled with applying RAMP due to time constraints. Teachers are required to complete the MOE set curricula and not RAMP. They have, in some cases, larger classrooms requiring more of their time for instruction, in which a RAMP routine might not be feasible.

Barriers to Adherence to RAMP Design in Teacher's Application of RAMP:

1. Trainers, teachers, and partners discussed math instruction barriers limiting teachers' ability to provide RAMP routines. QRTA discussed possible issues for teachers in applying math routines of RAMP, based on feedback they receive from calls; some teachers struggle with the subject and would likely have some difficulty applying RAMP routines.

2. Teachers discussed work load and their mandate to cover MOE math and reading curriculum within the class period. Currently, there is very little time available for additional instruction.

3. MENTORING

According to the plan by the IP, RAMP teachers should receive six visits per semester for a total of twelve mentor visits per year. In the next chapter of this report, the team discusses dosage and number of mentoring sessions completed. For this chapter on adherence, data from interviews with teachers and mentors suggested RAMP could not adhere to the planned number of visits.

RAMP has trained far more teachers than anticipated in the design, which has increased the demand of mentoring visits. RTI increased mentors by adding CADER education specialists to also provide coaching visits (i.e. these individuals were called RAMP Coaches). The MOE continued to send supervisors to conduct mentoring visits as planned, however they carried out fewer visits than set by the design. This is due in part to issues around schedules, as supervisors have very demanding MOE responsibilities, leaving them with less time to conduct mentoring visits. Logistics always played a role in blocking mentor-mentee check-ins as some supervisors would have to travel far to see a teacher and were not offered additional funds to conduct that activity.

The purpose of mentor visits was described during IP technical staff interviews. The goal of mentoring sessions was for these visits to serve as incentives for teachers to apply RAMP. The mentoring process and expectations of the impact of mentor-mentee visits was not explicit in documentation shared with the evaluation team. Observations of mentoring sessions suggested mentors spend very little time with the teacher. Often, mentors watch the class routine, then spend a few minutes after class with the teacher providing some additional guidance (observation data on mentoring). Findings for questions 2 and 3 provide more details on the time spent by mentors with teachers and teachers' views of mentoring sessions.

According to the IP, if teachers knew they would be visited by MOE supervisors or RAMP coaches, then they would be more likely to conduct the RAMP routines. Mentor sessions were to work more as site visits than face-to-face sessions with teachers; the frequency of the visit is more important, as stated by the IP, than the content of the visit. Although not discouraged, mentors could provide instruction to teachers during class, and some did so (observation data of mentoring).

During additional interviews with the IP, the team learned about the new mentoring plan, which had been created after RAMP's implementation for cohort 1. The mentoring element of the RAMP intervention has evolved into a three-phase approach. Given the national scale of RAMP activity, a shift from one phase to the next is not based on satisfaction on any specific criteria, but rather based on maturity of the Cohort (i.e. the passage of time).

- **Phase 1:** According to the RAMP team, coach/supervisor visits and frequency of visits are the most important aspect of this phase. As there is typically a gap of four weeks between the two-week RAMP training and the actual start of the school, phase 1 visits are meant to prompt the teachers to implement RAMP methods.
- **Phase 2:** During this phase, the coaches provide feedback to the teachers. This phase has primarily started in cohort 1 schools.
- **Phase 3:** During this phase, schools will be receiving differentiated support based on student performance. RAMP has been able to classify schools based on student performance using their LQAS study. The classification includes A, B, and C categories.
    - Category A schools will be receiving teacher level support.
    - Category B schools will be receiving support at the school community of practice level.
    - Category C schools will be receiving support at the cluster level.

Based on the interview data and observations, the teachers are all receiving visits and some verbal feedback by mentors. At the time of this study, the team did not acquire data to support the implementation of the third phase of mentoring.

Barriers to Adherence to RAMP Design in Mentoring:

1. Adherence to the six visits per semester is a challenge due to the considerable number of teachers. RAMP coaches can visit more teachers in a week when compared to supervisors (MOE mentors) who have other responsibilities beyond mentor visits.
2. There is limited evidence to know whether the three-phased approach is being implemented fully at this time, whether the participants are clear about the new approach, and whether this mentoring plan will be more effective on RAMP

4. COMMUNITY PARTICIPATION

Of the various activities carried out by RAMP to achieve the intended outcomes, community participation is the least reported effort. Conversations with the IP suggest more has been achieved in this area, but the reports (shared by the IP) do not seem to reflect this in much detail. In the most recent report available to the team from RTI (Midline Survey, August 2017), there is no mention of this effort throughout the entire report. Per the IP, the role of the community outreach staff member at RTI is to work closely with the MOE – specifically the heads of the divisions. According to the IP, three units within the MOE work with parents and the IP works with all three units. These are Parents-Teachers Council, Program Parent Involvement Program, and School and Directorate Development.

These three MOE units make up a working group with the IP that includes 42 directorates. Overall, the working group includes 12 people from the three units and 250 individuals from the 42 directorates across Jordan.

Overall, the IP, based on what they see in the field, feels participation at community engagement meetings is increasing. To increase this further, the IP suggests schools give students praise during these meetings to incentivize parental presence.

Conversations with the IP offer insight on the status of RAMP community participation activity:

- The IP has been interacting with this working group to raise awareness within schools (teachers and principals) and parents of the importance of supporting RAMP - children's learning of math and reading.

- The IP mentioned during the interview three objectives, that until the time of this assessment had not been known to the evaluation team, nor seen in any document to date[25]:
    - Objective 1: To raise awareness in school staff of parental benefits in their participation;
    - Objective 2: To increase the number of parents visiting schools and parental participation; and
    - Objective 3: To increase the number of initiatives suggested by parents (IP interview, 2017)
- The role of RAMP is to build capacity, and therefore the monitoring of community engagement is left to the various MOE units and not RAMP.
- For sustainability, the IP has helped identify coordinators for the working groups (at the time of the interview in Aug. 2017). The working groups were to determine, within a year's time, how to track and monitor progress.

> "*When I visit the schools and parents, the feedback [on RAMP] is very positive, I have started with each cohort 1 and 2, will meet with cohort 3 in September [2017]…we will meet [working groups] and exchange stories, lessons learned, challenges,*

The team was not able to confirm with MOE or directorate participants the extent to which RAMP has adhered to their community engagement activity. This is because specific targets and objectives do not seem explicitly stated in documents and therefore the details on RAMP's design around this effort and its contribution to student performance is unclear. Secondly, there was limited access to the MOE staff due to their schedules,[26] limiting the evaluation team's ability to get feedback from working group members.

Interviews with teachers and principals provided some additional information about RAMP efforts in community engagement which cover the classroom library program.[27] Teachers discussed informally with the evaluation team during class visits (through data collection activities for the Impact Evaluation) student participation in the library program. Students do appear to be signing out books for readings.

Principals were asked about the score cards RAMP shared with the schools. Score cards could be considered part of RAMP under "monitor school performance through interpreting school reports" (Figure 2: Original Ramp Results Framework), within the community engagement component of RAMP's design. Principals had varied understanding of score cards (quotes from principal interview data – pilot see Annex D). Some principals did not have any cards but seemed

[25] Documents available to date to the evaluation team.

[26] To not disrupt RAMP implementation and increase further workload on MOE staff, per communications with USAID and the IP, the evaluation team was requested to pause interviews with the MOE.

[27] To encourage reading, RAMP has provided students with books they can check out for reading at home. This is to encourage additional reading practice at home and parental participation. Students can sign out a book with their teacher.

aware of them, others did not have any knowledge of the cards, and one respondent said they knew the card helped them identify remedial students so to help them improve.

Principals provided other feedback regarding parents' involvement in RAMP (quotes from principal interview data – pilot) with mixed responses. Overall, the responses by the principals did not find parental feedback common; a few principals stated a negative response by the parents regarding RAMP while1 or 2 principals noted a positive view by the parents on the initiative (Annex D).

The team obtained limited information about the library activity in the classroom. This was in part to the scope of the qualitative study (where the focus of effort and time was on training, teacher participation/perspectives, and mentoring). During conversations with teachers the team did hear from instructors that students were checking out books. There is anecdotal evidence that some students are using the library book RAMP supported activity, suggesting this component is being implemented. However, the team is unable to determine the level to which classes are offering the program.

This study cannot provide a full review of the adherence by RAMP to the community engagement portion of their design. The team was not able to obtain data on:

- How are students learning about the library program?
- What is the participation rate by students?
- How do the parents perceive this activity and to what extent are they involved in supporting their children in reading? Or
- How is the library activity supporting student performance?

Although the IP is likely to have data on the number of students checking out books, this information has not been shown in any of the available reports. Several teachers shared with the team the sign-out sheets for books in their classrooms, but the results for this effort are unknown.

The teachers also were asked about their views on parental participation. More than half of the teacher comments (Annex D) did not see changes in parental involvement. The remaining responses thought maybe some parents might have visited the school slightly more than before. Overall, the teachers did not feel there was a substantial change in parental involvement with the school due to RAMP.

> "The b*iggest barrier [for parent participation] are school staff and principals…if they feel parents are important they respond." (IP

Comments from the teachers suggest mixed views. Due to the small number of responses, the study cannot generalize these views across all RAMP teachers. However, the results suggest further research is needed to improve understanding about RAMP's expectation of the role parents play in the intervention.

Barriers to Adherence to RAMP Design in Community Engagement:

1. Data suggested in some cases school staff and principals were the barrier to parent participation.
2. Funding is a limitation to engagement of the commun according to interviews with the IP.
3. Principal involvement is key to community participatio
4. Increased coordination within the working group is needed as they don't often know what each group is doing; this also is reflected in the MOE who does not always have all the information about how schools are engaging the community and parents.
5. Lack of incentives and rewards for increased community engagement – to date it does not appear that either teachers, principals, parents, or MOE staff receive additional resources or rewards for community engagement activities. Conversations with teachers and the IP suggested "recognition" through an award to those doing these efforts would be very welcome.
6. Increased accountability by supervisors, who are conducting the visits at the schools, is needed per conversations with stakeholders (schools and IP).

> *"Principals are key…they set procedures…they can make the parents feel welcome; principals are the most important in the field."* (IP

5. MONITORING OF RAMP

For purposes of this report, the evaluation team has focused on adherence around monitoring data of RAMP mentoring. Due to the interest at midline by USAID to improve understanding of this component, the evaluation team looked mostly at RAMP's tracking in mentoring. At the midline, the IP shared with the evaluation team and USAID their data sets on mentors; RTI had not yet been able to fully analyze the data due to staff bandwidth/resource constraints, as the teams were still rolling out RAMP. The evaluation team was asked to review the monitoring data and provide some recommendations. This is discussed in the next chapter under Research Question 2 (RQ2).

At midline, the IP had not utilized their mentoring monitoring data to inform, as suggested in their RAMP model (Figure 2 above), management adjustments to the intervention to maximize outcomes. This is a divergence from the intended plan. It is possible the data could have provided early guidance to helpful changes to the mentorship program for improved efficacy, as stated by their results framework. The IP team receives feedback during team meetings from their field offices, CADER, and MOE supervisors. However, detailed analytical data from the field on of the mentorship RAMP component was very limited.

At the time of the interview with RTI (November 2017), a new M&E system had just been created for RAMP. Technical staff at RTI shared some of the issues encountered prior to the new system, including data collection format, storage, and platform/software. Some have been addressed by the IP:

- Data from the field had been collected in paper format; these documents were not always in order or clearly coded by respondent or location.
- Through interviews with the data collection partner for RAMP, the evaluation team learned about issues regarding (a) lack of software or the appropriate platform for data collection, (b) utilization of subcontractors in other countries, and (c) limited interaction with RTI leadership about scale-up issues. The partner suggested that more in-person meetings among all partners might have helped all teams ensure clarity across all the components of RAMP as it grew.
- The M&E team gathered all documents from the three field offices and organized these creating specific cover sheets by teacher for each data entry.
- This data was then entered to a database and a dashboard was designed to show multiple views of the information.
- This work, although extremely time consuming according to the respondents, was key in getting the monitoring data organized for analysis. This may be one of the main reasons the analysis for mentors had not been conducted by the IP by midline.
- Other changes made to improve the monitoring process included a switch from paper to tablet data format for data entry. Mentors were given tablets to use during their mentoring sessions. The evaluation team was not able to view an example of the tablet format, but this was due to the timing of the interview.

- The RAMP M&E team created a "primary key" coding system, allowing the IP to identify duplicates and ensure high standards of data quality. They also implemented version control features, limiting access to only a few staff to ensure data is kept organized for analysis.

During interviews, the IP discussed other types of monitoring required for the project, including quarterly reports, reporting against indicators (annually 34 indicators, 8 related to the national survey, and some for USAID), and other reporting to RTI headquarters.

Other possible factors which could have affected how RAMP adherred to the original plan relate to team composition and personnel. At the early-stages of RAMP, the IP's team comprised a centralized technical team where all tools, instruments, data, and implementation protocols were primarily managed by one person (per conversations with the IP staff). However, after RAMP's first year, the IP expanded their team and underwent staff changes, including change in the Chief of Party, addition of an M&E team, and decentralization of activities. Interviews with RTI suggested the changes were made in recognition of the needs of the intervention as these emerged from the roll-out of the intervention. Streams of data for RAMP were more complex, requiring systems to manage and organize the information (interview data with M&E staff at RTI).

Changes in the structure of RAMP's management diverged from the original vision presented in the RAMP model. These changes likely improved the implementation of RAMP, and therefore would have increased the probability of the program meeting the targeted impacts. Yet, even with this adaptive management approach, RAMP was not found to impact student outcomes.[28] Some of the barriers presented below are linked to scaling issues and are likely the primary cause of a divergence by RAMP from intended levels of outcome.

[28] The RAMP Impact Evaluation did not find large impacts for students or teachers attributable to RAMP. Within the allotted evaluation periods, few aspects of RAMP were significantly related to math and reading student scores.

Barriers to Adherence to RAMP Design in Monitoring:

1. Tracking and monitoring activities was a challenge in the first year. However, the IP was able to identify obstacles, such as mentoring visit monitoring. The barrier was partially due to issues around scope of roles of partners. Conversations with the IP and Dajani highlight the differing perspectives on the scope of work for them. During the interview with the COP, this issue had been clarified through changes in contractual language and by the addition of the M&E team to RTI.
2. Partners viewed RAMP to have limited clarity of processes and protocols and recommended the teams (a) create explicit plans shared with everyone, and (b) set more templates for data entry.
3. Per the RAMP design, the tracking of mentorship was expected to be as previously done for the other activities – through notes, paper, and using excel macros. However, the size and scale of the activity, to include much larger numbers of teachers and mentors than planned, pushed RAMP to revise their approach.
4. Also, partners suggested that another likely issue contributing to why RAMP might have needed to diverge from the original plan was that the plan was not fully transparent. During interviews, some partners expressed a desire to know more about the status of RAMP efforts, to help them align their work. Partners felt increased clarity of templates, protocols, and expectation could have also helped them ensure RAMP's model was followed more closely.
5. Another issue identified by the IP regarded payments to teachers for transportation costs for training. When training started, teachers experienced significant delays (some with over 7 months) in their payments, with over 600 complaints (phone calls by teachers to regional managers). Through a new system established in the fall of 2017, teachers were now paid through tablets. In August 2017, teachers attended training and were paid 2 weeks after the training, with no complaints filed.

RESEARCH QUESTION 2 (RQ2) FINDINGS

Question: What, if any, were barriers in the full implementation of these two training elements that could potentially affect dilute/diminish the effectiveness[29] of RAMP on students? (Exposure)

[29] Effectiveness in this context is defined as the ability of RAMP to implement fully the designed intervention and reach the outlined student target goals.

Training and mentoring components encountered several obstacles during the implementation of RAMP. Many of these obstacles reflected scale-up issues which are common to projects when going from a small pilot to large national programs.

Exposure: Level or degree of exposure to the treatment. This can include number of sessions implemented, length of time of sessions, or frequency of a technique being

Findings for RQ2 focus on dosage levels estimated for training and mentoring, which ultimately affect student dosage of RAMP – per the program design. This chapter presents results on training dosage and the findings on teachers' exposure to RAMP mentoring. Mentoring results shown are derived from mentoring observation data, supervisor/coaches/teacher interviews, and mentoring data shared by RTI.

Findings:

Teacher Training Exposure and Dosage

1. Teachers did receive the training dosage of three weeks per the RAMP design.
2. The amount of RAMP training matched the design, but the content of the training may have varied across teachers. Some instructors were observed teaching math, as a subject, and not showing teachers the RAMP math routine. This reduced the amount of dosage in RAMP math instruction.

Mentoring Exposure and Dosage

1. *Dosage Amount*: Data collected in this study did not find evidence to support full dosage of RAMP mentoring for teachers as per the design. On average, the data suggested teachers received 3-4 mentor visits per year. Some cohorts may have received more mentoring exposure than others.
2. *Dosage Quality*: The data suggests there is variance of mentoring dosage across the teachers. Some teachers received multiple visits with one-on-one meetings with mentors. Other teachers stated they received supervisor/coaches with no follow-on discussion after the observation period in class.
3. *Mentoring Data Inconsistencies*: There are inconsistencies between what people remember as the number of visits (based on interview data with teachers), with the logged data by the IP on the number of mentoring visits, and with the actual number of visits conducted. Some teachers in their interviews reported 6 visits per semester,

RQ2: EXPOSURE/DOSAGE CONCLUSIONS:

- The actual dosage of math instruction, as a RAMP routine, was found to vary in trainings. Not all teachers received a full dosage of math.
- Mentoring dosage (number of visits, quality of visit, perception of mentoring process) was less than planned by the RAMP design (See Table 3).

- The impact study did not find RAMP to significantly affect student outcomes. One possible explanation could be the variance in training content and mentoring dosage.

TABLE V.3. SUMMARY OF RAMP TRAINING AND MENTORING DESIGN VS. ACTUAL

Type of Dosage	Dosage Specified in RAMP Design	Actual
Training Dosage	• 3 weeks	• 3 weeks • Math training content dosage varied
Mentoring Dosage	• 12 visits per year equal for all cohorts	• Between 2-6 per year (large variance) • Variance in exposure to mentoring by cohort

1. TRAINING DOSAGE

Although teachers attended the training (based on interviews with teachers, principals, IP, from sign-in sheets logged by the IP, and from observations by the evaluation team of training sessions), it is uncertain whether the total amount of training provided is sufficient to cause the intended impact. Therefore, while teachers adhered to the required training attendance levels (3 weeks), the dosage may be too low to create the required pedagogical behavioral change.

The IP had limitations around the length of time permitted to build capacity for teachers in RAMP routines; it was limited to only the 15 approved days of training. The results from the impact study did not find notable change in student performance scores in math and reading; one possible explanation could be that the RAMP dosage on teachers was not high enough. When teachers were interviewed and asked about their views on RAMP, some suggested the routines were difficult to implement at first. This variance in the application of RAMP may be contributing to an uneven/inconsistent administration of the RAMP intervention on students.

As discussed in the recent global meta-analysis of educational programs supported by USAID.[30], cascade training models can have pitfalls. Trainers often have different training approaches, and therefore can present RAMP differently to teachers. Observation data on training by the evaluation team showed different approaches to RAMP training across the schools visited. In some cases, the trainer worked on the actual subject matter; this was mostly commonly seen for math, where instead of a trainer working on the math routine (ways to teach math) as per RAMP, the trainer worked on teaching math. Observation data also noted that some training classes were very large, with a trainer having more than 30 teachers to train. This limits the time and ability of teachers to get all their questions answered.

Training materials were created for each of the three modules. Teachers and trainers had extensive guides to use during the training. Observation data from trainings found some training sessions only had 1-2 copies of the materials for the teachers. The evaluation team did, however, observe trainers and teachers conducting exercises and following guidance from the training materials.

[30] *Synthesis of Findings and Lessons Learned from USAID-Funded Evaluations, Education Sector 2013-2016*, March 2, 2018.

2. MENTORING DOSAGE

The mentoring dosage achieved was much lower than planned. RAMP envisioned twelve visits per teacher per year. On average most teachers (according to observation and interview data, and RTI mentoring data) had between 2 and 4 mentoring visits in total for the school year.

Several issues provide some explanation for the reduced mentoring dosage:

1. **The number of teachers receiving RAMP training was much higher than planned.** This placed an unexpected burden on mentors. In particular, supervisors who were already limited in their availability to support mentoring were now given additional teachers to mentor. The IP added more mentors through RAMP coaches from CADER. These individuals, unlike the supervisors, had the sole purpose of conducting mentoring visits and did not have to split their time with other responsibilities. This permitted RAMP coaches to conduct higher numbers of mentoring visits when compared to MOE mentors (interview data).
2. **The mentoring pairing between teacher and mentor is not transparent**. Per RAMP's design, each teacher who received RAMP training is designated a mentor from CADER or MOE. However, the designation process is somewhat obscure. The evaluation team was not able to get further clarity on the assignment process of mentors to teachers, or what criteria was used to assign larger mentees to certain mentors versus others. The only known information acquired from the interviews was around geographical region; some mentors had more teachers to visit because their region had more schools.
3. **Tracking mentor visits and their impact has not been captured fully or well by the IP.** Furthermore, the reports provided thus far by the IP had contained very limited information on mentoring and its effect on teachers' ability to do RAMP better or the effects on student scores. Each mentor, whether a RAMP Coach (coaches) or a MOE Supervisor (mentors), is asked by the IP to complete a form reflective of the visit. The data is then collected by a partner and transferred to RTI for analysis. RAMP monitors mentoring visits through these RAMP forms/data entry sheets. According to the IP, this data has not yet fully been analyzed due to team/staffing constraints.

These three issues influence the degree to which a teacher receives the full dosage of RAMP mentorship. The following sections in this chapter highlight: 1) summary findings about mentoring from the RTI data shared with the evaluation team, 2) description of the mentoring process from the observation data, 3) supervisor/coach teacher load results for mentoring from the interview data, 4) mentoring dosage results from the interview data, and 5) obstacles to mentoring based on observation and interview data.

SUMMARY FINDINGS ON MENTORING FROM THE RTI MENTORING DATA

The RTI data shows a total of **102,498 reported teacher mentoring sessions**. This number is reflective of all entries submitted to RTI, including planned but not completed sessions between

a mentor and teacher (i.e., teacher was not present during mentor visit). For more information on the data shared by the IP, see Annex E and Table 12.

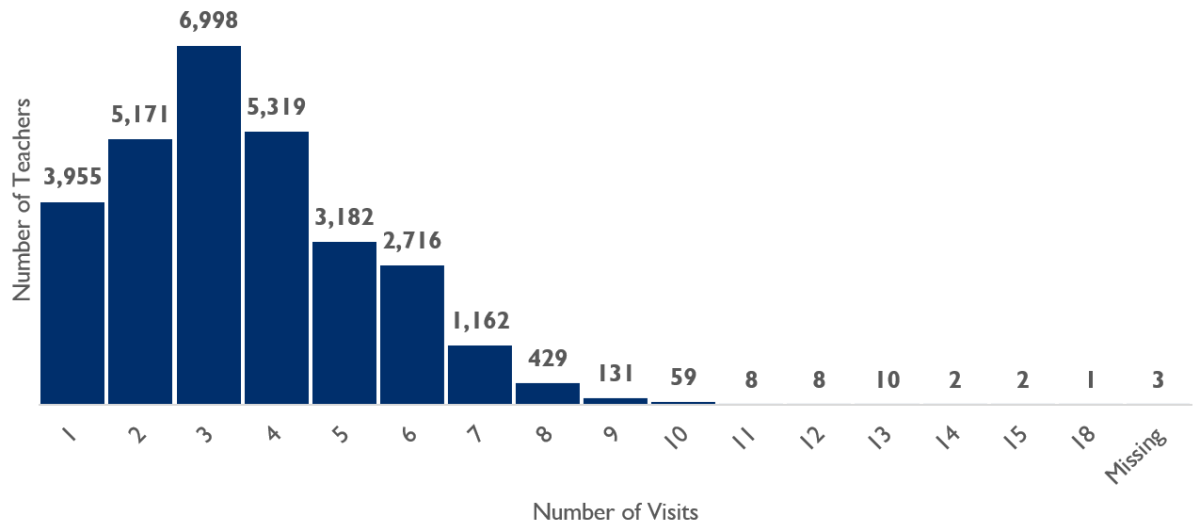
When the data is disaggregated by the number of visits per unique teacher-school code:

- Most teachers were visited 3 times by a coach/supervisor; 24 percent of the data (Table 13, Annex E; Figure 3) shows teachers with 3 mentoring visits across the entire data timeframe. This could be reflecting newer teachers from Cohort 3 for S1 2017-2018 for which the data is incomplete.
- Approximately 5,319 (18.2 percent) teachers had 4 mentoring visits; this includes teachers from all cohorts and is the second largest category of total reported mentor visits.
- Teachers who had 2 visits comprised 17.5 percent of all teachers.
- For teachers with only 1 reported mentoring session, this group totaled 13.6 percent of the reported sessions.

Overall, teachers appear to have a range of 2-4 mentor sessions. The actual total completed number of mentoring visits may be different, as the data is incomplete. Due to the gaps in the RTI data on mentoring, the team cannot provide further information about the mentorship across cohorts. Additional analysis may be possible with further clarification regarding the data shared.

Even with the limitations, the RTI data set (Annex E and Figure 3) suggests the mentoring dosage was lower than the target of six visits per the RAMP design. As the next few sections will show, interview data from coaches and mentors align with the much lower numbers of mentoring visits than the goal of six per semester.

FIGURE V.3. RTI DATA ON FREQUENCIES OF TEACHERS BY MENTOR VISITS[31]



Source: RTI Data.

DESCRIPTION OF THE MENTORING PROCESS FROM OBSERVATION DATA

The mentoring process was observed between mentor and mentee by the team for 16 pairs of teachers with their respective coaches/mentors. The team observed the following:

- Coaches/mentors started their mentoring session by checking in with the school principal's office. Mentors and coaches have pre-arranged the visit with the teacher. Some call their teachers in advance of the visit.
- Data shows the mentors/supervisors were all seen focused on the teacher during the mentoring observation process.
- During the class observation period, 62.5 percent of mentors/supervisors did not give direct feedback, but 37.5 percent did give the teacher advice during the class time, with a larger percentage being RAMP coaches (observation data).
    - Feedback was in the form of answering questions, verbal class instruction, and/or helping with a class routine.
- During class, except for one person, 15 of the 16 coaches/mentors wrote down notes and of these, 8 wrote down notes in a table. The evaluation team could not always observe if the RTI form[32] was being filled or not.
- The mentoring sessions observed were primarily in the morning (15/16 mentoring sessions observed were between 7:20-9:30am). One session was at noon. However, this is

[31] Source of the data for this figure is data shared by RTI on teacher visits by mentors. Additional information can be found in Annex E.

[32] RTI has a form which supervisors/coaches are required to complete during each visit. These started out as paper forms but changes in 2017 to the IP's process had the supervisors/coaches switch to electronic forms on tablets.

reflective of the access provided by the school to the data team. Mentoring visits can occur at any time during the day, at the convenience of the teacher and school schedules.

- Supervisor/coach in-class time varied by mentor. Maximum time the team observed a coach/supervisor in a class was 45 minutes, and the minimum was 15 minutes. The average time across 15 mentoring sessions (missing data for one session) was 29 minutes.
- Feedback observed from supervisor/coach was primarily verbal (87 percent of sessions observed by the team showed verbal feedback only).
- After the class ended, 5 of the 16 supervisors/coaches left the class with no further interaction with the teacher. More than half of the supervisors/coaches (9/16 sessions – 56 percent) held a short meeting in the hall or in the employees' room with the teacher.

- **Maximum** time a coach/supervisor was in a class was **45 minutes**
- **Minimum** time of mentoring was **15 minutes**
- **Average time** for a mentoring

One of the tasks for the team gathering observation data was to observe the supervisor/supervisee interaction. This was one of the questions and observations recorded:

"What did the supervisor/coach do in the class?" (Observation data):

☐	Greeted teacher and introduced himself to the students
☐	Coach interacted with students by moving among student groups and participating with them in the solution
☐	The coach focused on observing the teacher
☐	Coach is quiet and writes notes down
☐	He [coach/supervisor] asked the children about the work paper individually at their seats
☐	Coach/Supervisor evaluates the students' level in analysis and structure. Participated with the teacher in class and teaching
☐	He [coach/supervisor] offered feedback to the teacher

Of the mentoring sessions observed, more than five supervisor/coaches were seen praising the teachers for their efforts. The following are some examples of types of comments by the supervisors/coaches to the teachers as observed by the data collection team:

- Teachers were observed providing multiple pathways to solving a problem.
- Coaches were observed providing encouragement to the teachers during class when using the routines. One example observed included the mentor talking with the teacher about the division method specifically.
- One coach was observed providing instruction to the teacher on how to communicate with students of diverse levels.
- Another coach was seen giving suggestions to the teacher in class on the need to leave something written on the board at the end of the day and the importance of visuals for

the students. The solution also lets students be inspired to be creative as to how they get to the answer.

Overall, the in-class mentoring observation data was positive, suggesting in-class mentoring does not appear to disrupt the teacher's class performance. Close to 90 percent of the observations of mentor with teachers found, "Yes, the teacher was comfortable with the presence of the supervisor/coach – gave class smoothly." Only one person in the sample of the observation data felt there was a disruption due to the presence of the mentor.

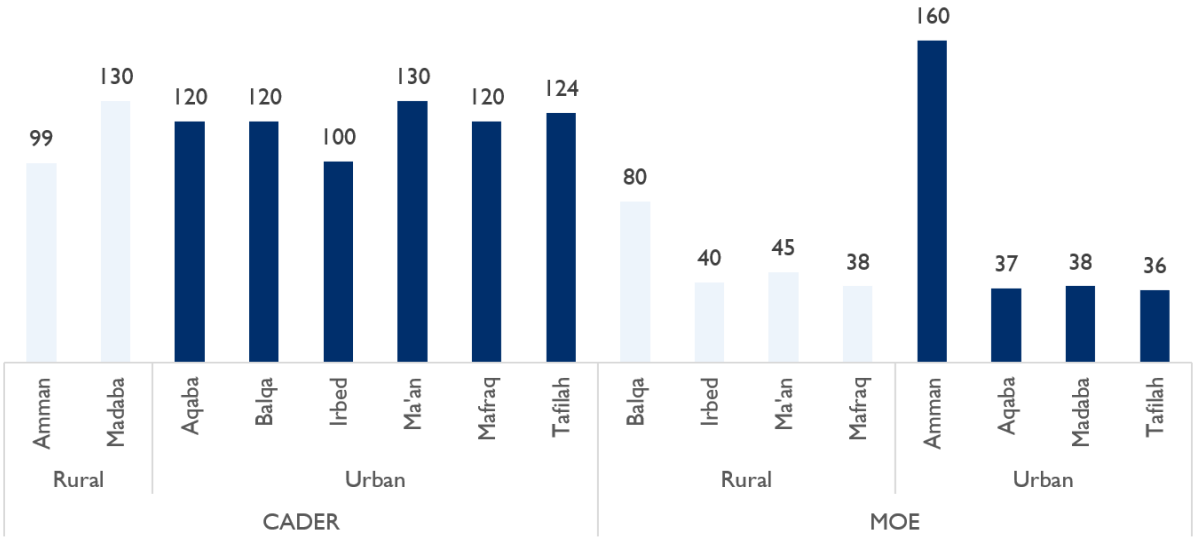
RESULTS OF SUPERVISOR/COACH TEACHER LOAD FOR MENTORING FROM THE INTERVIEW DATA

Interview data showed RAMP coaches (CADER) conducting a larger proportion of the teacher visits, when compared to MOE supervisors. Overall, RAMP coaches are more in number and tasked specifically with visiting teachers, allowing them to have larger teacher loads for mentoring and conducting more visits per teacher.

Of the 16 mentors and coaches interviewed for this study, half were RAMP coaches and half were MOE supervisors/RAMP mentors. This sample of 16 mentors and coaches reported in their interviews visiting 1,417 teachers in total, as part of the RAMP mentoring component. MOE supervisors reported visiting a total of 474 teachers, whereas RAMP Coaches stated they covered 943 teachers. These figures are for a total of 16 supervisor/coaches interviewed. On average, one CADER mentor would have conducted 117 mentoring visits (this is based on the interview data gathered for this study from mentors and coaches).

RAMP coaches – per the interview data – reported more urban visits and therefore likely larger schools (Figure 4). This may be one reason why they appear to have more teachers under their purview when compared to MOE supervisors.

FIGURE V.4. NUMBER OF TEACHERS COVERED BY MENTOR TYPE AND LOCATION



Source: Interview data

There are differences in teacher load between mentors (MOE supervisors) and RAMP coaches (CADER staff), as shown in Figure 5. The latter have the full-time role of mentoring compared to the MOE supervisors who must balance RAMP mentoring with their existing work load. The data gathered on the 16 supervisors/coaches show higher number of teachers and visits by RAMP coaches compared to MOE supervisors, almost double in difference. Supervisors are more limited in terms of number of staff available to conduct mentoring visits, as well as in the teacher load they can accept. MOE staff conducting mentoring visits are not compensated for the additional work. However, supervisors come with the full weight and authority of the MOE versus RAMP coaches who do not. This is both an advantage and disadvantage to RAMP coaches. On the positive side, some teachers stated their presence was seen more as educative, compared to a visit from a supervisor – who was sometimes seen as more evaluative with higher consequences. Thinking of the possible negative aspects of being a RAMP coach, they may not be viewed with as much authority as members of the MOE. Some coaches stated in their interviews that teachers declined to do the RAMP routine during the visit. The coach felt a supervisor could have helped change the attitude of the teacher.

The assignment process of teacher load for supervisors is unknown. However, the team has information about the process for teacher assignment of RAMP coaches. RTI has partnered with CADER to provide additional coaching to teachers, as the number of supervisors by the MOE is limited and the number of teachers targeted by RAMP was 15,000.

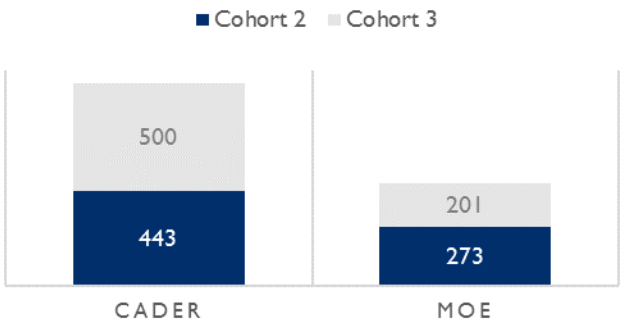
RESULTS ON MENTORING DOSAGE PER THE INTERVIEW DATA

Dosage per Cohort (interview data): Earlier cohorts likely received longer periods of mentoring visits and, thus, dosage amounts.

Overall, both types of mentors have relatively equal loads of teachers from each cohort (2 and 3 only). RAMP coaches had 500 and 443 teachers for Cohort 3 and 2, respectively. MOE Supervisors had 201 teachers for Cohort 3 and 273 teachers for Cohort 2. This study does not have mentors in its sample covering Cohort 1. Mentoring activities for Cohort 1 ended in 2017 and data collected for this study was in 2018.

Cohort 1 started their mentorship in semester 2 (spring of 2016) and continued visits by coaches/mentors in the fall semester of 2016 and spring of 2017 (three semester's worth of mentoring). Cohorts 2 and 3 received only 2 semesters of mentorship (fall 2016 and spring 2017 for Cohort 2 and fall 2017 and spring 2018 for Cohort 3) Cohort 3, the last group of the intervention, is likely to benefit from the project's maturity (i.e. learning over time and making corrections to improve RAMP). However, Cohort 3 also comes in when supervisors/coaches are at their maximum teacher load levels (i.e. number of teachers to coach/supervisor). This could potentially affect the quality of visits for Cohort 3 teachers vs. those from Cohort 1 where their supervisors/coaches had a much smaller list of teachers to mentor.

FIGURE V.5. NUMBER OF TEACHERS MENTORED BY MENTOR TYPE AND COHORT

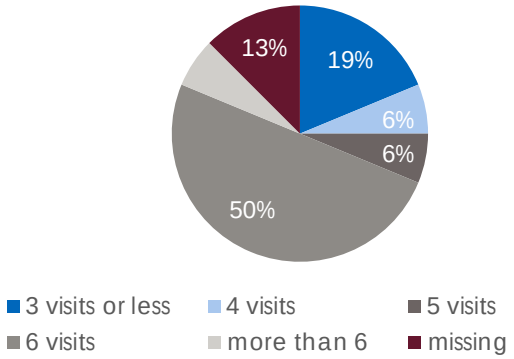


Dosage by Teacher - Number of Supervisor/Coach Visits per Teacher (interview data): Teachers perceived much higher numbers of mentoring visits when compared to interview data from supervisors/coaches and RTI data.

Teachers were also asked about the number of visits received by their supervisor/coach per semester. Half responded with 6 visits per semester by their supervisor/coach (Figure 6). There were more teachers from grade 3 who responded with higher visits by mentors (4/16 teachers) when compared to G1, G2 and KG.

When asked, "*What would you change to benefit more from the mentoring process?*" 75 percent of teachers stated they would not change anything, and 12 percent (2/16 teachers) said they would change the feedback process to incorporate teachers' opinions and invite outstanding teachers to share their success stories.

FIGURE V.6. SUPERVISOR/COACH VISITS PER SEMESTER AS REPORTED BY TEACHERS

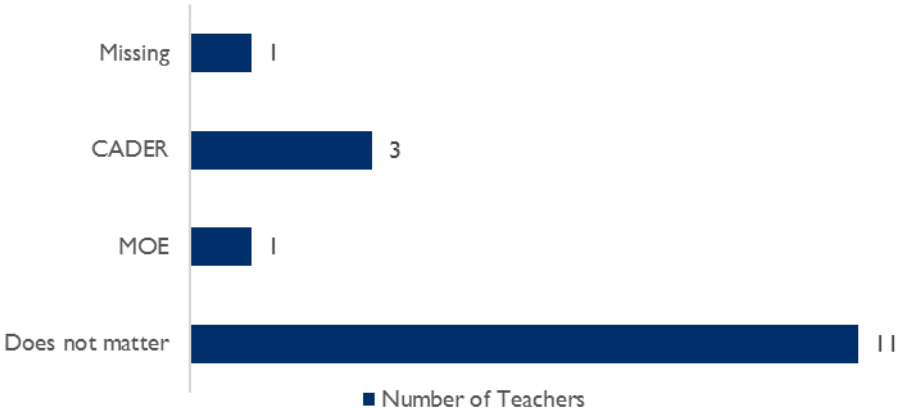


Regarding whether a teacher preferred the type of mentor - whether an MOE Supervisor or RAMP Coach from CADER - 68 percent of the teachers interviewed did not have a preference for the type of mentor (Figure 7).

Those who did prefer RAMP coaches found them more familiar with the material and less formal/ supervisory. Whereas teachers who preferred MOE mentors felt a more collegial connection.

- Teachers with RAMP Coach preference:
  - "…Because he is more aware of the initiative and his presence is not supervisory but evaluative."
  - "…Prefer staff members [RAMP Coach] because when the Ministry of Education visits I do not know what to do."
- Teachers with MOE Supervisor preference:
  - "…He is treated as a colleague more than a supervisor."

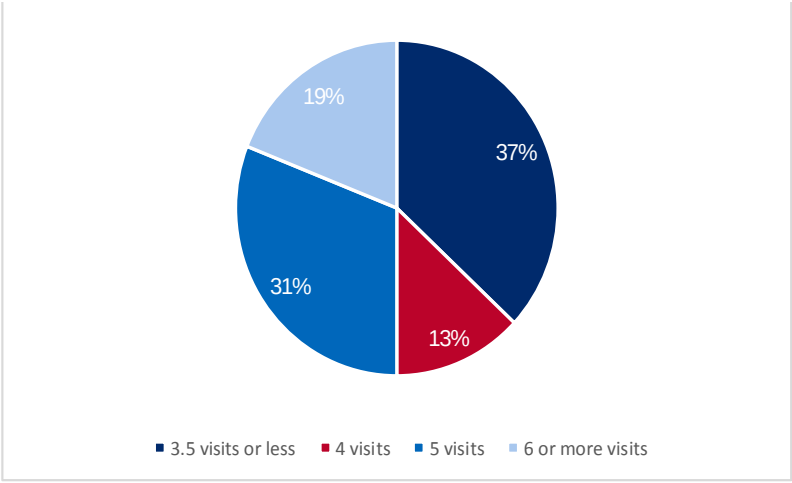
FIGURE V.7. TEACHER PREFERENCES ON MENTOR TYPE



Multiple components can affect the dosage of mentoring including the **number of visits a teacher receives from a supervisor/coach**. The design lays out 6 visits per semester for each teacher by a coach/supervisor. The design does not, however, specify how a teacher may be affected if the amount of mentoring is less than the target (6 visits). The assumption is likely to suggest more supervisor/coach visits for a teacher leads to an increased probability of RAMP being implemented more accurately and thus students receiving the full RAMP dosage.

The evaluation team asked both mentors and teachers to share the number of visits conducted/received during a semester. In contradiction to the reported number of supervisor/coach visits by teachers (which were closer to the 6 visits per term), supervisors/coaches stated fewer visits per semester (Figure 8).

FIGURE V.8. NUMBER OF MENTORING VISITS PER TEACHER PER SEMESTER AS REPORTED BY MENTOR INTERVIEW DATA



"What is the number of visits per teacher per semester?" (Interview data Supervisors/Coaches)

- 37 percent of the interviewed coaches/supervisors stated they had carried out 3.5 teacher mentoring visits or fewer per semester.
- 31 percent stated they have visited a teacher 5 times a semester.
- 19 percent of coaches and supervisors stated they had completed 6 visits per teacher per semester

Another factor that could affect the exposure level of the RAMP mentoring component is the **number of teachers or mentoring load for each coach/supervisor**. The RAMP design is not clear as to what the required minimum or maximum number of teachers per supervisor/coach should be to ensure full implementation of RAMP – and thus ensure full dosage of mentoring.

As supervisors/coaches increase in number of teachers they must cover, so do the number of required mentoring visits. This number does not reflect the distance to each teacher and time required for a supervisor/coach to travel to a school, which further augments the burden on the supervisors/coaches. Careful consideration must be given to ensuring the data on teacher visits is not misinterpreted. A coach/supervisor with a larger number of teachers and lower number of visits per teacher is not an indication of poor mentorship. It could be distance to each teacher is much higher for this supervisor/ coach when compared to another with similar teacher load (whose teachers may be located much closer to him/her or each other, reducing travel time, and increasing the chance to conduct more visits).

"What is the number of teachers you supervise/coach?" (Interview data of supervisors/coaches.) The minimum number of teachers supervised/coached was 36 and the maximum reported by the sample was 160. The total number of teachers covered across the 16 coaches/mentors was 1,417. The average number of teachers per supervisor/coach would be 88 teachers per supervisor/coach. If, then, each supervisor/coach (based on the data collected in the interviews of supervisors/coaches) was to fulfill the RAMP model of 6 visits per teacher per semester, that would equal 1,056 mentoring sessions per year per supervisor/coach covering 88 teachers each.

The **length and frequency of a mentoring visit may also impact the quality of the RAMP mentoring element a teacher receives**. Per the teachers' views, as discussed in the earlier section, mentoring frequency or amount of time per session was not presented as a problem or issue. Regarding frequency and length of supervising/coaching visits, coaches and mentors felt the repetitiveness of the visits was a positive factor for the teacher with an appropriate length of time. However, some respondents suggested:

- **Repetition may be feasible for a teacher but place a high burden on a supervisor** who have other tasks (2/16 respondents).
- **Repetition is more important than the duration of the visit** (1/16 respondents).
- **Increasing the time of the visits and reducing the number of visits** because the initiative is merging with the class curriculum/MOE policy and increased number of visits means increased paper work burden (1/16 respondents).

OBSTACLES TO MENTORING BASED ON OBSERVATION AND INTERVIEW DATA

Mentors and coaches were asked about the obstacles they face mentoring teachers through RAMP (Table 4). The following categories were identified through the interview data with mentors and coaches:

TABLE V.4. OBSTACLES TO MENTORING IDENTIFIED BY MENTORS DURING THEIR INTERVIEWS

Barrier/Obstacle	Findings and Respondent Comments
Distance to School	The location of the teacher can affect both the potential for a visit as well as subsequent teacher mentoring sessions by a supervisor/coach.
Resources	12.5% of mentors stated limitations to implementing their job due to costs – some schools can be costly to visit due to the distance from the mentor's home base. Some visits do not result in mentoring sessions (teacher not present), and there is no recompensation to MOE supervisor for additional tasks.
Transportation	19% of mentors revealed transportation and distance to be an obstacle in implementing RAMP mentoring.
Classroom Size	One of the mentors mentioned classroom size to affect a session because the routine takes much longer to implement in-class.
Teacher Participation/Absence	25% stated either teachers were not present when they visited, or they moved schools.
Workload	Paperwork was mentioned by the supervisors/coaches as a barrier to implementing RAMP. Of the mentor responses, 12.5% were on the fingerprinting process for attendance. There is a high burden on reporting attendance which can be at the directorate level (this office can be often far from where the mentor is located).
Other	Other obstacles mentioned included overlapping interventions, limited time for routines for the teachers and thus limited time for the mentor to see the routine, and teacher refusal to implement RAMP.

Figure 9 presents some of the mentor quotes on obstacles to mentoring, reflective of the categories identified above in Table 4:

FIGURE V.9. QUOTES FROM MENTOR INTERVIEWS ON OBSTACLES TO MENTORING

RESEARCH QUESTION 3 (RQ3) FINDINGS

Question: What were the perceptions by stakeholders of these two training elements? (Participant Responsiveness)

TRAINING AND MENTORING

Participant Responsiveness: Participant responsiveness…may include levels of participation and enthusiasm (Century and

This section summarizes the views gathered during interviews with teachers, supervisors/coaches, IP staff, and other RAMP partners. It presents stakeholders' views about training and mentoring. These views are important as they can potentially impact the degree to which RAMP is implemented by each group. Views on RAMP may also provide further explanation around the impact study results.

Findings:

Training Perceptions by Stakeholders

- Incentives for teachers to implement RAMP are not aligned with MOE curricula; therefore, some teachers may be implementing RAMP less than others.32F[1]
- Access to resources and materials limited teachers' ability to implement RAMP in the classroom; this likely led to variations in RAMP dosage to students.
- Interview data suggested use of coarse- and fine-grain tools.

Possible variance in the amount of math RAMP training received by each teacher, as some trainers spent more time on the subject of math. At the onset of RAMP there was likely an underlying assumption about teacher math skills being similar; data from this study suggest otherwise.

Mentoring Perceptions by Stakeholders

Perceptions by teachers on the mentors varied by mentor type. Some of the teachers viewed the MOE supervisors to have more authority, and thus their comments were taken more seriously. CADER mentors, on the other hand, were seen more as peers and their comments were suggestions – less pressure was felt by the teacher.

The implementation level of the mentoring stages is unclear; therefore, there is limited

RQ3 – Conclusions on Stakeholder Perceptions of RAMP:

- **Overall views on RAMP were positive** by those receiving RAMP intervention components (training and mentoring), such as the teachers and principals.
- **Data does suggest some variance in teachers' understanding and views of what RAMP** is, in terms of it being a routine. This issue may be less so today at year 3 of RAMP. The maturity of RAMP has permitted teachers to receive reinforcement directly or

indirectly on what the RAMP intervention means. This has been through the mentoring visits, teachers sharing among themselves questions about RAMP through WhatsApp groups, and principals with more knowledge of the program.

- **Partners did request more transparency on the achievements of RAMP** goals as these progressed. They suggested more face-to-face meetings with the IP to keep everyone on the team informed of the various pieces of RAMP being implemented simultaneously.

1. TRAINING

According to USAID and RTI, based on their interaction with the MOE, the ministry has been supportive, helpful, and enthusiastic about the RAMP intervention. Interviews with RAMP partners (RTI, CADER, QRTA) suggested the MOE has felt a direct linkage to RAMP (further information about the partners' perspectives can be found above in the findings section for RQ1). The MOE has permitted the IP as well as USAID evaluation teams to access schools, teachers, principals, and students. The IP and MOE have negotiated the RAMP training length and reviewed the modules being offered to teachers.

Through interviews, the evaluation team asked teachers, principals, and partners to share their perspectives on RAMP training.

TEACHERS

Half of the teacher interviewees described the training as "good," "useful," and "appropriate." Overall, the teacher views of the training were very positive. Teachers were also asked during interviews their perspectives on the trainer who provided RAMP instruction. Respondents as a group felt the trainers delivered RAMP in an easy format, used direct implementation methods/ exercises, and seemed well trained. When teachers were asked "Do you feel that the benefits of reading and math are equal among students?" over half (56 percent) stated yes, 31 percent said no, and the remaining 12 percent made comments about the differences.

Other questions in the teacher interviews invited description of expectations around whether RAMP would continue being implemented and what needs teachers might have going forward. Table 5 presents some of their responses regarding expectations to continue using RAMP in the classroom. Most of the responses were positive, but some responses were also tentative.

TABLE V.5. TEACHERS' RESPONSES REGARDING THEIR CONTINUED APPLICATION OF RAMP

Do You Expect to Continue Using the Reading and Math Initiative? Why? (Teacher Interview Data)		
Yes	**No**	**Maybe**
"Yes, because it is useful to students, and I got used to the teaching method"	"No"	"It makes class interesting and varied in styles, and it addressed individual differences"
"Yes, because it became part of teaching"		"I expect it will continue because the teacher got used to the style and will not change it"
"Yes, because it organized teaching methods and facilitated the delivery of information to students"		"I expect to continue using it. The teacher cannot implement the initiative for more than a year and then stop applying it."
"Yes, because it facilitated learning for students"		"I expect it will continue because of the results of exams and the better academic levels attained"
"Yes, because it serves the curriculum"		"I hope it will continue because there will be communication between teachers and students over new methods"
"Yes, because it is convenient and benefits students"		
"Yes, because it works with the initiative and applies it"		
"Yes, I expect it will continue"		
"Yes, because its goals are noble, coming from the Minister of Education, and there is media focus on it"		

Perspectives by teachers on what types of resources they may require to sustain RAMP in the classrooms are shown in Table 6.

TABLE V.6. TEACHERS' RESPONSES ON THEIR NEEDS TO SUSTAIN RAMP

Response Categories	**What Will You Need to Continue Implementing the Reading and Math Initiative?** (Teacher Interview Data)
Teaching Aids/Tools	"Teaching aids should be made available to the teacher"
	"Providing teaching aids and tools"
	"Teaching aids that help students"
Materials/Stationary	"Material support like stationery…"
	"Including booklets and activities and increasing support materials"
	"Providing material support for the classroom environment"
Incentives/Recognition	"Doing repeat visits, giving financial incentives and thank-you letters"
	"The availability of supervision and follow-up helps the initiative to continue."
Relief from Paperwork	"Relieving the pressure of written work off the teacher"
	"…Reducing the load of teachers"
	"Continuing till the tenth grade because the method changes between the third grade and higher grades"

Additional Funding	"Providing means"
	"Sources of financial and moral support"
Other	"Would love to raise the level of students and their performance in reading and math"
	"Cooperation between teachers and supervisors, and holding a course for new and old teachers"

To summarize, teachers felt they would continue implementing RAMP. They did highlight some of the resources needed for them to continue RAMP. These included: additional financial assistance, incentives, recognition/appreciation letters, additional supervision, provision of aids/tools, and teacher aids to support class learning.

PRINCIPALS

Of those principals interviewed from the sixteen mentor-teacher pairs schools, over half the responses were in praise of RAMP as a positive activity for their teachers. A few principals discussed some teachers having difficulty early in the RAMP process. Another principal felt teachers needed to implement RAMP over a longer period for it to be more helpful to students. The following is a summary of principal responses (Table 7):

TABLE V.7. PRINCIPALS' VIEWS OF RAMP

Response Category	**"Can You Share Your View of the RAMP Initiative?"** (Principal Interview Data)
Benefits to Students	"I wish it could be applied to all primary upper grades up to grade 10, as it is excellent, based on the results obtained"
	"It is a wonderful initiative. It raised the level of the student from the stage level to the basic level. Students learned concepts that were only heard in the higher grades, and that earned them more appropriate life skills."
	"Students benefit from it if there is feedback from the teacher, and the initiative would be successful only if most students are outstanding. It takes time to implement."
	"It is purely educational, and we hope to implement it with the higher grades"
Benefits to Teachers	"The initiative was successful and helped the teacher in using modern teaching methods and improved the level of students"
	"It is a useful initiative in terms of teaching and supporting teaching methods and the educational process. It helped in rehabilitating teachers"
	"First, it made a change in the teacher and added new experiences, methods and strategies and gave him the ability to change the behavior of his students"
	"Teachers are happy, as the initiative has brought out their technical and educational creativity as well as creativity in teaching this program"
Financial Needs	"It is a good and useful initiative, but more resources, means and financial allocations should be provided to support the initiative"
Increased Burden	"I am supportive of the initiative, but I object to the endless paperwork"
	"It is brilliant, but teachers are concerned about the heavy burden of writing, and the activities are clear but need some organization"
Other	"At first there was difficulty in implementing it, but after a while the idea became clearer and easier"
	"It is excellent and achieves its goals if it is implemented correctly and if verified properly as well"

	"Despite all my prior reservations and the teachers' objections for the same reasons, I found justification for this situation, as I give binding instructions from the Ministry of Education to apply them as they are despite the objection"

PARTNERS AND IP

Partners (CADER, KAIZEN, Dajani, QRTA) and the IP had the following views on the training (Table 8):

TABLE V.8. PERSPECTIVES ON TRAINING BY THE IP (RTI) AND PARTNERS (QRTA, CADER, KAIZEN)

Partners	RTI
• Significant effort was given to the development of the training curriculum with investments by QRTA and the MOE to ensure the training provided in math and reading was appropriate. • Partners felt they had been given opportunities to contribute in the process and found the IP open and willing to their suggestions. • QRTA interviewees felt teachers would likely, even with all the training given to them through RAMP, still struggle with the math component. They discussed how teachers in the past had contacted them regarding math as a subject for early grades and the difficulties teachers sometimes faced in teaching. • The QRTA respondents felt math was an area that likely would need additional support. • The training provided to the trainers (based on trainer comments) was viewed as sufficient and appropriate for the intervention. The few trainers who spoke with the evaluation team members stated they felt prepared to conduct the training, but they felt some teachers might not be ready after the training to do RAMP.	• Interviews with the IP revealed positive views on the training process and curriculum. The staff at RTI felt the training components were appropriate but also based on what was permitted by the MOE. The IP would have liked to have had a longer training period for the teachers; however, the MOE could only permit 10 days of RAMP training during the summer period with additional days between the semesters.

2. MENTORING

TEACHERS

The data from the interviews found that teachers viewed the visits by the supervisors/coaches positively. According to their response, these mentoring visits provide helpful feedback and guidance for in-class instruction, give teachers information to improve the implementation of RAMP, and help identify strengths/weaknesses around instruction/application of math/reading routines.

What the data does not provide is information on whether this component of RAMP is significantly affecting student outcomes. Would the absence of this RAMP element affect the impact on student scores? How much of the change in performance of math and reading scores is attributed to the mentoring component? This is an important question as it helps guide future adaptations

to RAMP to exponentially increase the effectiveness of the program. If mentoring is helping teachers implement better or more accurately RAMP routines, this could be reflected in dosage and thus RAMP impact on students. However, the impact study of RAMP did not find significant changes on students' scores. Would the impact of RAMP change if the mentoring dosage changed? If RAMP was designed for teachers to have 12 visits per year, reduced mentoring dosage could provide some explanation for the impact study results.

During interviews, teachers were asked to provide their views on how mentoring helps them as teachers, helps them implement RAMP, and how a supervisor/coach visit helps. Teachers found the feedback helpful, felt supported, and generally viewed positively the supervisor/coach visit. They also rated their coach/supervisor as highly supportive, cooperative and helpful. Table 9 summarizes the views on mentoring held by teachers:

TABLE 9: TEACHER VIEWS ON MENTORING FROM INTERVIEW DATA

Interview Questions	Responses
"How do mentors/supervisors help or not in your roles as a teacher?"	• Helps provide feedback and encouragement • Explains things and gives suggestions for improvements • Supervisor/coach is supportive by giving creative examples in the classroom • Helps to implement the initiative • Provides real-time corrections during class
"Did the supervisor/coach help you with RAMP?"	• Yes, "…he gave me help and support and advice when needed" • Yes, "…he helped me implement the initiative" • Yes, "…helped by showing the methods that suit the students and how to communicate with them" • Overall, responses were all positive and teachers found the supervisors/coaches helpful
"Is their [coach/supervisor] visit helpful? If yes, how? If no, why not?"	• Everyone in the sample responded "yes" (16/16 teachers) • Teachers felt supervisors/coaches provided feedback and guidance regarding the implementation of RAMP • The visits helped teacher identify strengths and weaknesses • The visits provided direction towards improvements • Supervisors/coaches helped to identify gifted students • According to one teacher, the visits were helpful because the supervisor/coach helped her improve her performance
"What do you think of your supervisor/coach?"	• "He helped me a lot by providing feedback after each visit and opened the door to communicate with him in matters that are not clear in the initiative" • "the supervisor is good and directs us to the best way to deal with students" • "…patient and supportive, offers guidance" • "He helped me a lot in improving performance" • "He helped me with the implementation process and treated me really well" • "He gave me help, advice and feedback" • "…very cooperative and gives moral support"

This study did not ask the teachers how many times they changed supervisors/coaches during a semester or over the RAMP mentoring period. However, this event is very plausible, potentially affecting the quality of the mentoring experience. Based on conversations with teachers/mentors/principals/CADER, the team heard that some teachers had and could change supervisors/coaches during a semester and possibly during a school year. In the section on

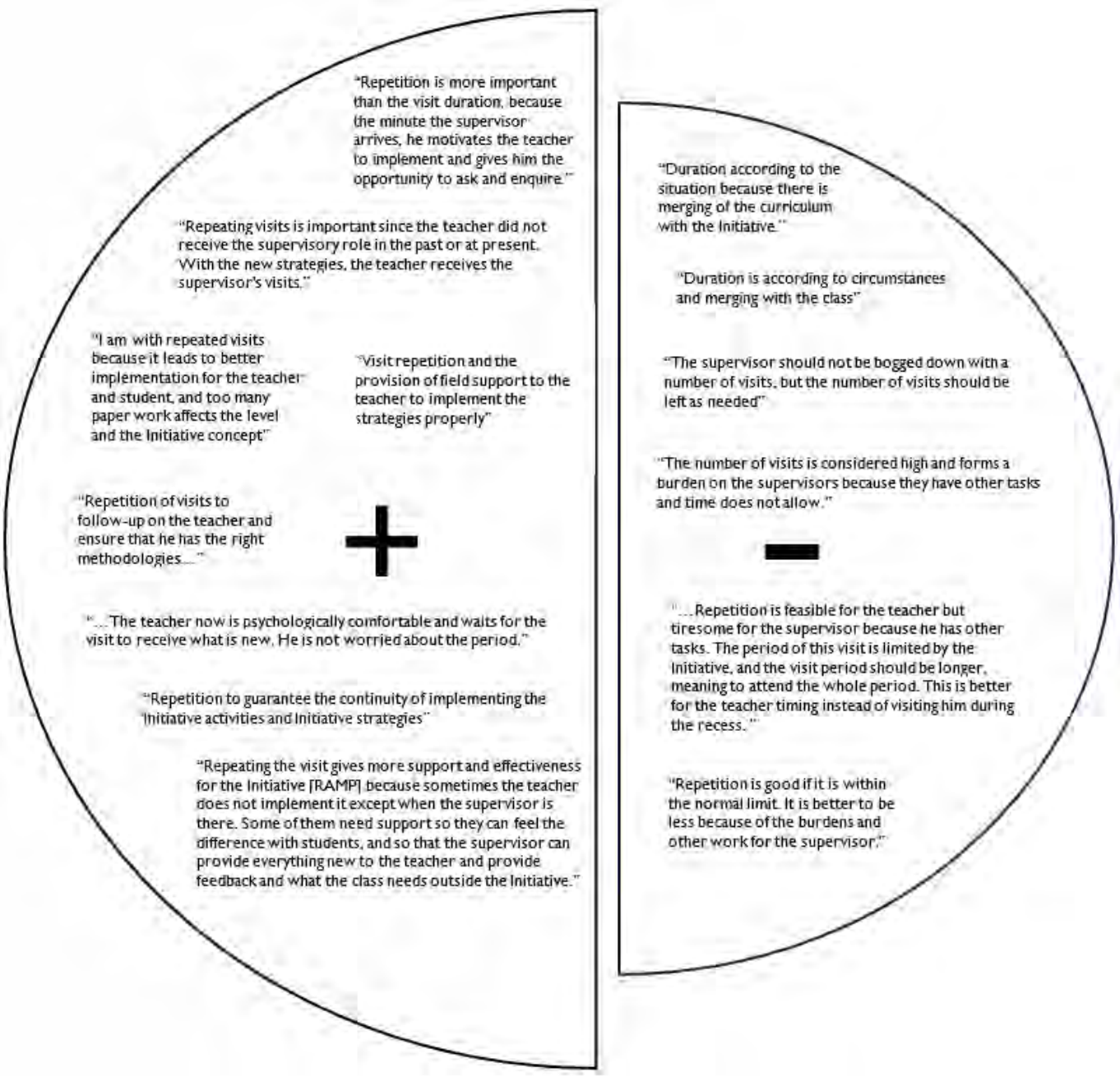
supervisor/coach views, they discuss how an obstacle to their job is teacher mobility. Frequent changes in supervisor/coach could affect the dosage of mentorship.

SUPERVISORS/COACHES

The team asked coaches and mentors to share their views regarding duration of mentoring. Figure 10 highlights some of their responses, showing a larger number of positive responses around the frequency and length of mentoring visits. Supervisors/coaches who had positive views stated the importance of repeat visits to support implementation, ensure the right methodology is present with teachers, and incentivize the teacher to apply RAMP. Some of the less positive responses noted the burden on the mentors to meet the targeted number of visits, and that duration/frequency should be teacher specific.

FIGURE 10: SUPERVISOR/COACH VIEWS ON THE FREQUENCY AND DURATION OF MENTORING VISITS UNDER THE CURRENT RAMP DESIGN

"What is your opinion on the frequency and duration of the coaching/supervising visit?"
(Supervisors/Coaches interview data)



Mentors and coaches were asked about their views on what mentoring components were most effective. The responses fell into three categories: in-class observation, post-class time with teacher, and overall field visitation.

- The most common response by a supervisor/coach was on the importance of having the in-class observation component.

- The second most popular response on what makes the mentoring visits effective was just having a visit; field support for a teacher was viewed as important to provide direct and real-time feedback to the teacher (See Table 10).

- Finally, mentors and coaches also found the time spent with the teacher after class observation was an effective component of the mentoring visit.

TABLE V.10. MENTORS' PERSPECTIVES ON MENTORING

Interview Questions	Responses
"In your opinion, which part of the mentoring is effective?"	• Post-classroom time with the teacher (2/16): 12%
	• Overall visit because it provides field support to the teacher (4/16): 25%
	• In-class observation because supervisors/coaches can see the student reactions and teacher capacity (9/16): 63%
"How do you think the teachers and headmasters (male and female) look at the supervision and mentoring process"	**Open to RAMP Mentoring** • Teachers seem open and view mentoring as collaborative • More than 50% of supervisors/coaches stated the words "open communication," "cooperative" and "supervision" • Overall supervisors/coaches felt the teachers were interested in RAMP • One respondent stated "They are open to supervision and all that is new…" • Principal shows open communication with supervisor/coach and teacher • One response stated teachers "accept more" the visit by the supervisor (MOE mentor), than from the RAMP Coach (CADER). This person believes it is because the teachers received more diversity in teaching options from the MOE mentor than the RAMP coach • One respondent stated teachers do not seem to see the mentoring as disruptive to their schedules
	Resistant to RAMP Mentoring • One mentor stated some teachers refuse to implement RAMP. • Another supervisor/coach stated they found male teachers more resistant to implementing RAMP when compared to their female teachers
"What are your thoughts on the continuity of this process (supervision and mentoring) for the RAMP reading and math?"	**Increased Monitoring is Needed** • "Grade three (Cohort 1) was outside the equation and were not trained, so we hope they will be reconsidered. In general, cohorts 1, 2, and 3 should be monitored (first, second, and third). They need to be monitored to ensure the Initiative is implemented in terms of the presence of the Initiative strategies and their being implemented from grade 1 to 6, to ensure that the ratio is 100%. We have now reached 55% and we can reach 100% at grade 6."
	Targeted Supervision and Monitoring is Needed • "Focus on people who refuse the initiative by training them more. Increase motivation for teaching (attempt to convince them)" • "Emphasize the need for supervision and follow-up so that the effect is there in a larger way." • "Implement the initiative to the curriculum, the teacher should be convinced of the implementation. Monitoring and evaluation are necessities. Currently, applied tasks are being implemented (not everybody implemented them.)" • "Intensify field visits for teachers." • "Continue visits and follow-up for teachers and prepare a monitoring tool to be implemented by the teacher in the reading and math initiative, and the way in which the headmaster can be involved and placed in the picture."

Interview Questions	Responses
	Merge Initiatives with the MOE Curriculum and Increase MOE Involvement • "Merge with the Ministry curriculum or other initiatives. Coordinate and connect with other Initiatives to provide more effectiveness for the material" • "Merge activities with the book." • "Important to form a team that supports the directorate." • "Offer the role to the supervisor to create a friendly atmosphere between the supervisor and the teacher."
	Improvement of RAMP Incentives and Teacher Buy-in Level • "Full conviction by the teacher of the initiative's importance. This does not mean that the teacher needs monitoring but through rehabilitating school headmaster and selecting the best teacher participating in the Initiative for coordination and implementing the initiative activities." • "Organize a competition among schools or offer incentives and encouragement Initiatives."
	Reduce Burden on Teachers • "Reduce the burden for the teacher because the Initiative requires a lot of work, and the teacher needs to follow up as needed." • "In case there is not sufficient support, visits must be reduced because they form a burden for the supervisor." • "Reduce training hours when starting to train new teachers."
	Increase Funding and Resources for RAMP implementation • "Provide all supplies related to implementing the Initiative to the school." • "Financial support should be provided so that the Initiative continues."
	Increase/Certify Training and Capacity of MOE staff • "Additional training for teachers who attend. The Initiative should be a course in universities or part of teaching in universities." • "A suggestion for the Ministry that the skills the teacher learned should have a guide or manual inside the Ministry's teachers' guide." • "Training for everyone (teachers, headmasters concerned, supervision) so that everyone knows what is inside the Initiative." • "The person responsible for the training should be confident that Ministry cadres are capable of following up."
	Exchange RAMP Knowledge between Stakeholders • "Exchange visits with other directorates to be exposed to experiences and exchange them. Visit exhibitions related to the Initiative." • "Select an outstanding teacher who knows the initiative well so that each school has a coordinator to communicate with the teachers and provide assistance and support for the continuity of activating the Initiative through documentation in a register that is shown at the end of a specific period, presenting the experience and its results, and honoring those with best results."

Regarding teacher/principal views on the mentoring process as perceived by the mentors, more than half the respondents held positive views about the RAMP mentoring component.

- According to the mentors, teachers seemed to be overall open to the mentoring process.
- Some of the mentors stated teachers had refused to implement RAMP and found sometimes the male teachers more resistant when compared to the female instructors.

Coaches and mentors were also asked about sustainability of mentoring visits. As presented earlier, the RAMP mentorship is supported by the IP through RAMP coaches via CADER as well as through MOE supervisors who also mentor teachers. Some of the barriers to sustainability shared by mentors/ coaches related to logistical support for MOE supervisors as well as their workload limiting time for mentoring.

PRINCIPALS

Principals were also interviewed and asked their views on the mentoring component of RAMP. Their views have been presented earlier in this report and are illustrated in Figure 11. Overall, the views were in support of mentoring, and identified this component as a helpful activity for teachers and for the implementation of RAMP in the classroom.

FIGURE 11: PRINCIPAL INTERVIEW RESPONSES ON RAMP MENTORING

"What do you think of the [RAMP] coach/supervisor?" (Principal interview data)

CONCLUSIONS AND RECOMMENDATIONS

The purpose of this qualitative study is to provide additional information about the effects of RAMP and the degree to which the implementation of two components (training and mentoring) held or deviated from the original program design. To help improve understanding around the results of the RAMP IE being conducted by MESP, USAID requested this study be carried out to provide further explanation around the fidelity of the implementation of RAMP. When a program is not implemented according to the intended design, the deviations may be reasons for limited impact.

The study focused on the following questions:

1. Was the planned intervention for teacher training and mentoring implemented to the RAMP design specifications? (**Adherence**)
2. What, if any, were barriers in the full implementation of these two training elements (teacher training and mentoring) that could potentially affect dilute/diminish the effectiveness[33] of RAMP on students? (**Exposure/Dosage**)
3. What were the perceptions by stakeholders of these two training elements? (**Participant Responsiveness**)
4. What are suggested recommendations to improve the RAMP training components?

This section presents conclusions for research questions 1-3, each with recommendations (RQ4), respectively.

RESEARCH QUESTION 1: WAS THE PLANNED INTERVENTION FOR TRAINING IMPLEMENTED TO THE RAMP DESIGN SPECIFICATIONS? (ADHERENCE)

1. Scale-up complications: Yes, the training was implemented to the RAMP design, but the program underestimated the number of teachers needing training, actual resources available to teachers, and degree to which teachers move.

Scale-up issues are common for all programs going from a pilot-scale effort to a nationally implemented activity. The estimates required to plan national-scale programs can have sizeable error – in part because schools can make final decisions about the number of classrooms and teachers as late as the beginning of the semester. If the estimates for teacher and student numbers are based on the prior year, these could be significantly flawed (either significantly underestimating or overestimating RAMP needs). Additionally, the time required to provide mentoring by a mentor (mentor prep time, travel time, mentoring time, and reporting time), might have been somewhat underestimated. The estimated number of mentors required to conduct the

[33] Effectiveness in this context is defined as the ability of RAMP to implement fully the designed intervention and reach the outlined student target goals.

number of sessions per the design might have been too low. Scale-up challenges could be a reason as to why the RAMP design deviated from its original plan. Based on results from the RAMP IE, some of the interviewee teachers stated they had taken the RAMP training more than once. This could be also why the estimates for teacher numbers are higher. The IP was able to mitigate some of the issues that arose, such as provision of additional materials for teachers unable to photocopy these at the schools as expected, provision of additional mentors due to increased teacher numbers, and additional training to cover teacher mobility between schools/grades. However, even with these changes to help RAMP adhere to the original design, the interview data suggest the design did not take into account variation in how the training was administered.

More teachers than anticipated were given the RAMP training, and this increased the number of trainers, mentoring visits, and monitoring data. The IP did provide more training sessions with an additional module in November to help teachers who had missed training or had joined schools after RAMP module 1 and 2 in August. Materials were created for teachers to implement RAMP in classrooms which had not been part of the original RAMP design. Teachers were found to be struggling to implement RAMP (during the first days of RAMP and based on interview data) because the instructors did not have enough copies of the materials they needed; schools have limited paper and photocopying capacity. Some teachers stated they would hand copy the materials to ensure students had a copy.

2. Limited feedback loops during the first year of RAMP: The RAMP design was not fully implemented with respect to the envisioned feedback loops from the monitoring data. A more comprehensive and robust system was needed to track RAMP activities. The evaluation team understands that the IP now has this in place.

RAMP did not follow its original design around monitoring and evaluation (M&E). The actual size and level of effort required to monitor the RAMP components, such as training and mentoring, were not sufficient. During interviews with the IP staff, they discussed how stretched their team was across multiple tasks. Some of the IP tasks added were indirectly part of RAMP, added after RAMP had started being implemented. Other tasks added to the IP were not part of RAMP but set as a priority. The IP did revise their M&E system to better track field data, store information, and organize for analysis. Although there has been limited analysis by the IP of their own monitoring data (for example, of mentoring), they are now positioned to conduct the needed assessments of their activities. However, the capacity of RAMP during the early years (1 and 2) to use implementation feedback loops was limited by the original M&E system. This likely led to missed opportunities for timely program adjustments.

Potential issues in the design assumptions: Although RAMP training was implemented as designed, data in this study highlights possible issues around the design assumptions regarding the time to achieve the scale of the impact envisioned.

Although the scaling and dosage discussed may present a plausible reason for limited findings of the impact of RAMP on student performance, the underlying assumptions about the causal model

of RAMP may not hold true. This might be a more significant problem to the RAMP intervention. The qualitive data highlights challenges implementing RAMP but finds overall individuals willing and open to carry out this activity. Buy-in for an innovative approach is not usually easy in education, but RAMP does appear, within the span of 2 years, to have become widely known.[34] Interview data from partners and stakeholders suggested RAMP may have more impact in a few years and that the effects might not be visible at this time (where impact measurements were made). To change the approaches and views of an entire educational system on how math and reading are taught usually takes several years to take hold.

Recommendations

1. Create materials for teachers that are easily accessible in multiple formats, including electronic tablets, paper copies, and supplementary manuals to support RAMP implementation.

2. Conduct internal assessments of the data collected to identify patterns and issues of RAMP implementation. This can help guide allocation of future resources.

3. Testing larger dosages of training on selected teachers to see if this affects student performance.

4. Adding informal learning channels for teachers to pursue additional capacity building, such as learning working groups through WhatsApp, online YouTube training videos like Kahn Academy, and webinars from QRTA in math (a subject area for which additional support was requested).

5. Restructure the supervising/coaching component to give teachers more one-to-one time with a mentor. The mentoring dosage may then have a higher potential of helping increase

RESEARCH QUESTION 2: WHAT, IF ANY, WERE BARRIERS IN THE FULL IMPLEMENTATION OF THESE TWO TRAINING ELEMENTS THAT COULD POTENTIALLY AFFECT/ DILUTE/DIMINISH THE EFFECTIVENESS OF RAMP ON STUDENTS? (EXPOSURE/DOSAGE)

TRAINING DOSAGE

There were several barriers to the full implementation of RAMP. The data suggest the most prevalent was differential approaches to training teachers on RAMP math routines affecting dosage levels, variance in mentoring visits (dosage), and differences in participant views of RAMP leading to diverse levels of application of RAMP components.

[34] Ibid.

Exposure to RAMP may have differed or been diluted due to early challenges in implementation: teachers' differential access to resources, variations in RAMP math teaching, and teacher mobility. This would lead to some teachers implementing RAMP more than others, causing some students to have higher dosages of the intervention when compared to other students. Interview data and results from the IE found teacher mobility was present during RAMP. Some teachers might have received double dosages of RAMP (as interview data suggested they stated to have attended more than one RAMP training), making them stronger RAMP implementers.

Variations in training were also present, with some trainers teaching the math subject versus the math RAMP routine. This may have led to limited time for teachers to learn about the RAMP intervention and how-to execute in class.

Another dosage issue relates to the amount of RAMP training received by teachers (15 days). The RAMP design was planned for 10 days pre-fall semester, with a few additional days for follow-up between semesters. The length of the training dosage was set by the MOE. However, this amount of time for training in an intervention cascade model, such as RAMP, may not be sufficient time according to the IP teachers and other interviewed stakeholder.

The amount of training given to teachers was the maximum allowable by the MOE to the IP. Based on the interview data, most teachers and principals felt the amount of time was sufficient. Other studies for in-service training suggest this may not be sufficient or the cascade model used may be problematic.[35] Limited incentives, lack of additional class time to implement RAMP, and missing directive from MOE curricula for RAMP added further challenges to an effective RAMP intervention.

MENTORING DOSAGE

The mentoring dosage envisioned in the design of 12 visits per year per teacher was not achieved. The actual number of visits per data collected in this study suggests teachers had between 2 and 4 mentoring visits per year.

Obstacles to the implementation of full mentoring dosages include: increased number of teachers leading to increased load on mentors, limited understanding of the monitoring data on mentoring to assess the full dosage of this component, and coach/supervisor resource restrictions.

[35] Synthesis of Findings and Lessons Learned from USAID-Funded Evaluations, Education Sector, 2013-2016. March 2, 2018.

Recommendations:

1. Provide math training to teachers prior to RAMP training, so when receiving the RAMP dosage, trainers can focus on routines only.

2. Create a specialized unit in the MOE solely focused on training modules for math to improve math prior to RAMP routine training.

3. Work with the MOE to increase the number of days training is provided or to permit additional training during the semester as part of teacher professional development plans. This would increase the dosage of training on teachers, especially if the training is provided through multi-model formats (i.e. in-person, online/webinar, workbooks). An example is Kahn Academy, which has formed partnerships with countries to provide online courses and training modules across multiple fields, including mathematics for early grades.

4. To reduce variance in training, create a "trainer team" who is the sole provider for training. RAMP could be implemented in a way that provides official educational credits or certification for teachers. This in turn could be part of an advancement option for teachers. By providing standardized training modules, the intervention could ensure more even distribution of the RAMP dosage to teachers.

RESEARCH QUESTION 3: WHAT WERE THE PERCEPTIONS BY STAKEHOLDERS OF THESE TWO TRAINING ELEMENTS? (PARTICIPANT RESPONSIVENESS)

The team gathered participant recommendations for RAMP via interviews. Teachers and principals found RAMP helpful and the trainers prepared and effective in giving RAMP training.

Overall, respondents have a positive view of RAMP and an interest in continued participation. Teachers and principals expressed a desire for continued education, such as the RAMP training; teachers and mentors stated mentoring visits were of benefits to teachers and thus their classroom. A few respondents did mention some difficultly implementing RAMP at first, but this view diminished with time. Table 11 presents stakeholders' suggested changes for improving RAMP:

TABLE 11: RECOMMENDATIONS FOR RAMP BY PARTICIPANT TYPE FOR TRAINING AND MENTORING

Recommendation Topic	Respondent Type	Suggested Changes/Recommendations
Training Type	**Teachers**	• Provide practical, not theoretical, trainings.
Resources/Logistics	**IP and Partners**	• Provide resources to teachers to implement RAMP as this had been identified as a barrier early in the RAMP process. Some of these have already been implemented by the IP.
		• The creation of manuals and integration of RAMP into approved MOE curricula, which was in process according to those interviewed by the evaluation team.
	Teachers	• Provide teaching aids to support the teacher.
		• Create a detailed and integrated manual for RAMP and the curriculum to make it easier for teachers.
		• Create more Arabic activity books.
	Principals	• This group had similar responses to the teachers. They also suggested additional support in materials and manuals.
	RAMP Coaches (CADER)	• Type of training and logistical conditions - Trainers must feel comfortable in terms of location and date. Distance from the training venue causes an obstacle to the trainer, especially if the trainer is a female and has children.
Recognition and Incentives	**Teachers**	• Hold competitions among governorates for math and reading.
	Principals	• Like teachers, principals also recommended RAMP to find ways to recognize the teachers through simple certificates/awards to encourage them to continue the intervention.
Paperwork/Burden Reduction	**MOE Supervisors (Mentors)**	• Fingerprinting attendance of the educational supervisor is recommended to be changed.
		• Reduce the amount of paper work per visit. "Too much paper work for the supervisor…The supervision visits log and tablet which take a long time".
Knowledge Sharing	**MOE Supervisors (Mentors)**	• Ensure supervisors are involved in the activity/models early on, before the mentoring starts, as well as ensure the MOE changes the supervisor work model.
	Teachers	• There should be an exchange of ideas and experiences with other teachers.
Training/Mentoring Load	**IP and Partners**	• Increased support to mentoring by the addition of mentors from CADER (RAMP coaches) as MOE supervisors are limited in number and in availability to mentor.

Recommendation Topic	Respondent Type	Suggested Changes/Recommendations
	Principals	• Increased visits from supervisors, parental participation and community involvement.
	MOE Supervisors (Mentors)	• Number of visits is very high and recommend 3 visits per semester only. Supervisors work among the school teachers (to be a coordinator and monitor regarding the Initiative implementation).
		• Reduce the number of teachers per supervisor and not add any additional teachers until they implement RAMP correctly.
	RAMP Coaches (CADER)	• Transfer the teacher to a MOE mentor, when the CADER coach is rejected. This will ensure the teacher received the mentoring visits required.
		• Reduce the number of teachers per mentor.
		• Also, identify which teachers require more or less visits to have more effective mentoring.
Targeted Intervention	**MOE Supervisors (Mentors)**	• Provide training and capacity opportunities based on teacher levels. "Give the Initiative a certain specificity, evaluate outstanding classes even in teacher and supervisor training."
	Teachers	• Consider school conditions, as these differ and affect RAMP implementation.
		• The initiative should include other subjects.
		• Help and involve parents in the math learning activity.
Training/Mentoring Time	**MOE Supervisors (Mentors)**	• The number of visits must not be set, and the supervisor should be the one who identifies that because he/she knows the situation of teachers.
	Teachers	• Ensure the class is enjoyable for the student so it is more effective.
		• Use official hours of work for RAMP activities.
	RAMP Coaches (CADER)	• Consider holding trainings with the mentor and teachers inside the school (hold training with teachers to increase benefit during visits or hold meetings and group training with teachers).
		• There should not be a specific length of time for the mentoring session. Duration of the mentoring session should be individualized to each teacher based on need.
Project Data and Monitoring	**IP and Partners**	• Improving the collection and storage of monitoring data, which the team has done by creating a new tablet-based system with an online dashboard. New protocols have been put in place to track and ensure quality in the monitoring data of RAMP.

Recommendation Topic	Respondent Type	Suggested Changes/Recommendations
	Evaluation Team	• Create a plan and process to document the changes to the intervention. Several key changes were made to RAMP over the course of implementing the program nationally. Tracking these changes is important for understanding the impact of RAMP, as these changes my diverge significantly from the original design. Thus, the final impacts may be due to a different program design than what had originally been envisioned.

Although the sustainability of RAMP remains nebulous, some of the components of RAMP are clearly welcomed by teachers, principals, and students. Teachers and principals frequently stated positive views about the RAMP training goal and mentoring component. Further discussions with the MOE should ensure focus on which of the RAMP elements require future support to continue, such as mentoring visits and materials for teachers. Although the impact of mentoring is still unclear, as it pertains to increased effectiveness of RAMP routines and thus improved students' outcomes, further research should be conducted with the IP mentoring data to assess the outcome of mentoring on effective adoption of RAMP routines, and the appropriate dosage of mentoring to achieve the desired results on teacher practices and subsequently on student learning outcomes. The IP should consider tracking a sample of teachers over time regarding the benefits of mentoring to ensure the characteristics with the most effect is retained in future variations of RAMP.

Measuring implementation fidelity can help identify areas where the project diverged from the planned design. To better estimate the implementation barriers and possible changes required to meet the program design-intervention, this study suggests USAID consider, for large scale-up programs dervied from pilot studies, conduct a needs assesment prior to the launch of the national activity. This process could help identify unforseen issues affecting the dosage or exposure times of the intervention due to scalee-up challenges. This study has suggested various areas where RAMP may be strengthened, and potentially improve the likelihood of improved student math and reading scores.

Although not discussed in detail in this study, there is clearly buy-in of RAMP by most of the stakeholders. This achievement cannot be undervalued, as one of the reasons many programs fail to implement or sustain over time is due to the lack of engagement by stakeholders. The IP also has access to strong partners able to locally support RAMP, including the MOE, QRTA, CADER, and others who worked to launch RAMP nationally. Both findings from this and the RAMP IE study provide suggestions to help RAMP reach the identified targets. The team for this study recommends the IP consider conducting further assessments at the end of RAMP and one-year post-RAMP, as there could be lag effects not captured in prior assessments.

SUB-ANNEX A. INTERVIEW GUIDE: IMPLEMENTING PARTNERS

Interview Guide – Implementing Partner RTI

The MSI MESP RAMP IE team would like to set up interview times with various RTI staff involved in the implementation of RAMP. The objective of this activity is to:

- Map out the intervention process through conversations with the implementing organization, observation data of the RAMP training, and KIIs of participants (trainers/teachers/supervisors).
- This information will be used partially to assess the fidelity of implementation of the RAMP intervention, as well as expand qualitative data around the "why" of the IE results.

The evaluation team would like to talk with the following team:

- **RTI Management and Technical** leads/staff on the activity are useful in understanding how the activity has progressed over time, whether there have been any key changes/adaptations that may have a bearing on the impact evaluation design.

Date:			
Interviewer Name:	☐ Carolyn ☐ Afnan ☐ Other _____		
Participant(s) Name:	***Name***	***Organization***	***Role/Position***
Group Interview:	☐ Yes ☐ No		

Consent:

Your participation is completely voluntary, and you can withdraw at any time. The information you provide will be not keep any personal identifier information. Participating in this interview is of minimal risk to you and should not cause you any harm.

Do you consent to participate in this interview?	☐	**Yes**	☐	**No**
Are you comfortable doing this interview in English?	☐	**Yes**	☐	**No**
We will be taking some notes (laptop/notebook)	☐	**Yes**	☐	**No**

[Interviewer describes interview process]

- **"We would like to ask you a few questions about your knowledge and experience in RAMP"**
- **"There is no need to prepare in advance"**
- **"Please feel free to stop us at any time to clarify or add further comments"**
- **"Should you wish further clarity on any question please feel free to ask us"**
- **"You can stop at any time or decline to answer any question"**

[Interviewer – we do not have to cover every question, some might be answered in a prior question, there may not be enough time, the respondent may not wish to answer, the respondent may not know, and or the respondent may wish to send us information to answer that question.]

[Interviewer - ASK FOR EXAMPLES when possible]

Interview Questions

ROLES AND RTI STAFF

1. Can you please tell us who is on the Management/Technical Team?
    a. What is the organizational structure of the different units (and who is in these units)?
    b. Reporting plan within the unit and management
    c. Who are the team leads for each unit and how do they work with each other?

HISTORY OF THE PROJECT/INTERVENTION

2. Please describe the overall history of the intervention, walk us through the project?
    a. What was the original design/plan?
    b. Key stages and rationale
    c. What is primary assumption underlying the theory of change?

PARTNERS

3. Please tell us about your key partners/subcontractors on this intervention? What is nature of your relationship and their contributions to RAMP?
    a. What role did RTI play in writing and creating the materials, the training design?
    b. What role did the Academy, Kader, and MOE play?
    c. Where there any other partners who helped shape the training?
4. What have been some key challenges and opportunities when it comes to these partners?
    a. Are there any partners you wish you had after having implemented RAMP, if so who would they be? What partner do you think you might have been missing?

DESIGN

5. How did you arrive at your training design (modules, schedule, length)?
    a. Who participated in the development of the materials?
    b. What role did RTI play and others (like MOE) to what level, HQ?
    c. Did you get input form teachers and supervisors?
6. When designing this project in Jordan, what customization did you make, as EGRA/EGMA is in other countries?
    a. What was unique about Jordan?
7. What is the rationale behind the staggered implementation approach, and is there an advantage to this implementation process?
    a. What would you change?
    b. What have you changed from cohort 1, 2 to now 3
8. What would you say are the Strengths and Limitations of RAMP/of the design?
    c. What would you do differently?
    d. Any changes in your overall approach since the start of your activity? Why?
9. What are some key advantages and challenges associated with a staggered cohort-based implementation process?
    e. What would you change?

f. What have you changed from cohort 1, 2 to now 3

CHANGES

10. Have there been any key shifts/changes/adaptation in your activity? If so, why?
11. What are your mechanisms, in case any, for tailoring your intervention, whether by geography or based on other technical considerations?
    a. What type of field visits or observations does the management team conduct?
    b. Is it direct, quality check, or is the based primarily on feedback forms?
    c. How do you decide who and where to visit?
12. If your work identifies any low performing schools, is any action taken or any changes made to improve their scores?

FEEDBACK

13. From whom, and how, do you elicit feedback concerning RAMP? Is this feedback used and or does it lead to changes/adjustments?
14. How do you incorporate MOE feedback, when and in what areas, give examples?
15. How often do you communicate and how with your trainers, supervisors, MOE, academy, other?

M&E and EFFECTIVENESS

16. How do you assess the overall effectiveness of your approach?

SUSTAINBILITY (SCALE-UP, REPLICATION)

17. Could you tell us about your sustainability approach, what would sustainability look like for you if you achieved it?
    a. What is a measure of sustainability for you?
    b. Could you share the challenges in scaling up your activity across Jordan?
    c. What are some key assumptions
18. What are some key threats to the likelihood of sustaining your intervention, in case any?
19. What have been some key challenges, in case any, in scaling up your pilot intervention?

END of Interview THANK YOU:

[Thank your participant]

- **"We would like thank you for your time and for sharing with us your views about RAMP"**
- **"The information collected will be stripped of all personal identifiers"**
- **"Once all stakeholders have provided their views we will write a final report about the various perspectives of RAMP by each group. This work will help inform the ongoing RAMP Impact Evaluation"**

SUB-ANNEX B. OBSERVATION TOOL: MENTORING

Observation Tool – Supervisors/Coaches

The MSI MESP RAMP IE team would like to observe supervisors/coaches during the supervising/coaching process from both CADER and MOE. The objective of this activity is to:

- Map out the intervention process through conversations with the implementing organization, observation data of the RAMP training, and KIIs of participants (trainers/teachers/supervisors/mentors/coaches).
- This information will be used partially to assess the fidelity of implementation of the RAMP intervention, as well as expand qualitative data around the "why" of the impact evaluation (IE) results.
- Data collection of supervisors/coaches is also an area of inquiry specifically requested by the USAID education team post RAMP midline findings.

The evaluation team would like to set up observation sessions with individuals mentored/coached by CADER and the MOE at the following schools based on timing of session and willingness of participation by supervisor/coach and respective teacher:

- **CADER**
  - [TBD NAMES/numbers/location from which the sample will be selected]
- **MOE**
  - [TBD NAMES/numbers/location from which the sample will be selected]

Class ID: |___|___|___|___|___|___|___|___|
Teacher ID: |___|___|___|___|___|
Gender (Teacher): ☐ **MALE** ☐ **FEMALE**
Teacher name: _____
Teacher phone number or email address (if available): _____

Observer ID: |___|___|___|___|___|
Date of Observation: |___|___| / |___|___| / |2|0|1|8|
Month **Day** **Year**

Consent:

Your participation is completely voluntary, and you can withdraw at any time. The information you provide will be not keep any personal identifier information. Participating in this activity is of minimal risk to you and should not cause you any harm.

Do you consent to being observed during your supervising/coaching process?

Supervisor/Coach	☐	**Yes**	☐	**No**
Teacher/Mentee	☐	**Yes**	☐	**No**

[Observation process]

- **"We would like to gather data about the mentoring coaching process by observing your supervisor/coach and their interaction with you"**
- **"There is no need to prepare in advance"**
- **"We will not be asking questions, and be silent observers"**
- **"You can stop us at any time or decline to participate"**
- **"The observation time should last the duration of your supervising/coaching session"**

[Interviewer – you may wish to walk them through the process of what it would look like, where you might sit, that you would be taking notes, and other information.]

Observation Questions / Categories			
Theme of Inquiry	**Question**	**INTERNAL**	
		Observation Recorded	**Comments**
Classroom Characteristics	1. Classroom Size	☐ _____ number of students (numeric)	
	2. What was the teacher doing at the time when the supervisor/coach arrived?	☐ Teaching a math lesson ☐ Teaching a reading lesson ☐ Teaching other _____ ☐ Could not determine	
Supervisor/Coach in Class Observing Teacher	3. When did the coach/supervisor arrive, and how long did he/she observe the teacher?	☐ ___arrival time (hours minutes) ☐ ___ minutes into the class ☐ ___ number of minutes supervisor/coach stayed in the class observing the teacher	
	4. How did the coach/supervisor introduce themselves to the teacher? How did they introduce the purpose of their visit?	☐ Please describe ☐ _____	
	5. Did the supervisor/coach focus on the teacher and class?	☐ Yes ☐ No, the supervisor/coach seemed distracted ☐ Could not determine	
	6. Did the supervisor/coach provide any verbal 'real-time' feedback to the teacher during class?	☐ Yes ☐ No ☐ Could not observe/determine	
	7. What did the supervisor/coach do in the classroom?	☐ Please describe ☐ _____	
Supervisor/Coach Post Class Interaction	8. How did the supervisor/coach interact with the teacher the class?	☐ _____ ☐ _____	
	9. Please select all that apply regarding the type of feedback given to the teacher from the supervisor/coach:	☐ Supervisor/coach gave verbal feedback to teacher ONLY ☐ Supervisor/coach gave written feedback to teacher ONLY ☐ Supervisor/coach gave BOTH written feedback and verbal to teacher ☐ Supervisor/coach gave no feedback (that I could observe)	

		□ Other _____	
	10. If the supervisor/coach interacted with teacher after observing them in class, what did this interaction encompass [mark all that apply]:	□ Teach and Supervisor/Coach had a meeting □ Teacher and supervisor/Coach did NOT have a meeting, coach/supervisor left after class observation □ Could not observe/Do not know	
	11. What feedback did the supervisor/coach provide to the teacher	□ Please describe □ _____	
Other Observations	12. Were you able to see how the supervisor/coach was collecting data?	□ Supervisor/coach wrote notes □ Supervisor/coach filled out form □ Could not determine	
Your opinions	13. In your view, did the teacher appear comfortable with the presence of the supervisor/coach in the classroom?	□ If "Yes", how □ No □ Could not determine	
	14. In your view, was the classroom disrupted (i.e. students were distracted, lesson was delayed, other) during the time of the mentor?	□ If "Yes", how □ No □ Could not determine	
	15. Overall, if you could rate the interaction between supervisor/coach and teacher, where 1 is "the interaction seems very unfriendly/most uncomfortable", and 5 is "very friendly/comfortable interaction", what would level would you rate the interaction at:	□ 1 – very unfriendly/uncomfortable □ 2 – somewhat □ 3 – neutral □ 4 – somewhat □ 5 – very friendly/comfortable □ 99 – did not observe any interaction between them □ 8 – could not determine	Please provide some comments as to what you observed to determine your response
STOP			

END of Interview THANK YOU:

[Thank your participant]

- **"We would like thank you for your time and for sharing with us your views about RAMP"**
- **"The information collected will be stripped of all personal identifiers"**
- **"Once all stakeholders have provided their views we will write a final report about the various perspectives of RAMP by each group. This work will help inform the ongoing RAMP Impact Evaluation"**

SUB-ANNEX C. INTERVIEW GUIDE: SUPERVISORS/COACHES

Interview Guide – Supervisors/Coaches

The MSI MESP RAMP IE team would like to set up interview times with individuals working as supervisors/coaches of RAMP schools from both CADER and the MOE to learn about their experience. The objective of this activity is to:

- Map out the intervention process through conversations with the implementing organization, observation data of the RAMP training, and KIIs of participants (trainers/teachers/supervisors/mentors/coaches).
- This information will be used partially to assess the fidelity of implementation of the RAMP intervention, as well as expand qualitative data around the "why" of the impact evaluation (IE) results.
- Data collection of supervisors/coaches is also an area of inquiry specifically requested by the USAID education team post RAMP midline findings.

The evaluation team would like to set up interviews with the following individuals at **CADER and the MOE**:

- **CADER**
  - [TBD NAMES/numbers/location from which the sample will be selected]
- **MOE**
  - [TBD NAMES/numbers/location from which the sample will be selected]

Date:			
Interviewer Name:	☐ Bandar ☐ Amer ☐ Afnan ☐ Other _____		
Participant(s) Name:	***Name***	***Organization/School***	***Role/Position***
Gender:	☐ Male ☐ Female		
Type of Mentor	☐ CADER ☐ MOE		

Consent:
Your participation is completely voluntary, and you can withdraw at any time. The information you provide will be not keep any personal identifier information. Participating in this interview is of minimal risk to you and should not cause you any harm.

Do you consent to participate in this interview? ☐ **Yes** ☐ **No**

We will be taking some notes (laptop/notebook) ☐ **Yes** ☐ **No**

[Interviewer describes interview process]

- **"We would like to ask you a few questions about your knowledge and experience in RAMP"**
- **"There is no need to prepare in advance"**
- **"Should you wish further clarity on any question please feel free to ask us"**
- **"You can stop at any time or decline to answer any question"**
- **"The interview should take between 40-60 minutes"**

[Interviewer – we do not have to cover every question, some might be answered in a prior question, there may not be enough time, the respondent may not wish to answer, the respondent may not know, and or the respondent may wish to send us information to answer that question.]

[Interviewer - ASK FOR EXAMPLES when possible]

Interview Questions			
Theme of Inquiry	**Question**	**INTERNAL**	
		Prompt	**Responses**
Warm-up/Background	16. Please describe your role and responsibilities as a supervisor/coach	☐ What are they tasked to do as a supervisor/coach? ☐ Confirm they are currently supervisor/coaches	
	17. Can you please tell how long you have been a coach/supervisor for RAMP?	☐ Number of months or years	
	18. What schools/geographic areas are you responsible for?	☐ Have you always covered these areas?	
	19. Where you a trainer during the RAMP training?	☐ [we want to confirm the number of roles they might have played in RAMP as some coaches/supervisors were also trainers]	
Selection/Preparation to be a Supervisor/Coach	20. How were you selected to be a supervisor/coach?	☐ What were the requirements to be a supervisor/coach? ☐ How did they find out?	
	21. Can you please share with us the training you received by RAMP to be a supervisor/coach	☐ Who provided the training and for how long? ☐ What did the training cover?	
	22. [if yes to training] In your view, did the training prepare you for your position as a supervisor/coach?	☐ What was missing/What would change/add?	
Supervisor/Coach Activity	23. Please describe for us the selection process for your mentees/people you coach	☐ Do you select the people and the schools where you supervisor/coach? ☐ How many mentees/people to coach do you get a week?	
	24. How many mentees do you have? [estimated number is fine]	☐ Total number they are responsible for ☐ Does this number change by semester? ☐ Do they keep the same teachers for the whole year	
	25. How often do you visit them per semester? [estimated number is fine]	☐	
	26. Can you walk us through a typical session of supervising/coaching	☐ Please include the entire process, if more than one meeting with the teacher, etc. ☐ How do you collect the information about the person?	
	27. What data do you collect during our session?	☐ How do you collect the information? ☐ Where does it go after you collect it? ☐ Who do you report to – who do you send the data to?	

Interview Questions			
Theme of Inquiry	**Question**	**INTERNAL**	
		Prompt	**Responses**
	28. How do you provide feedback to the teacher?	☐ Verbal in person while there during my observation ☐ Written in person while there during my observation ☐ Verbal over the phone at a later time ☐ Written via email/mail at a later time ☐ A combination of the above ☐ Do not provide feedback	
	29. What is your view on the frequency of visits and mentoring time?	☐ Does it matter how often they [the mentor] visits a teacher? ☐ Is frequency more important than length of time?	
	30. How do you think the teachers and principals view the mentoring and coaching?	☐ Are they interested? ☐ Do they seem open and willing? ☐ Does it fit with their schedule? ☐ Is the time allotted enough time? ☐ Do you get contacted by the teacher/mentee for advice?	
	31. What do you think are the benefits of this activity, to you, to your mentee, to the school, to the students?	☐ Who benefits? ☐ How do they benefit? If at all ☐ Share examples if possible	
	32. What is the biggest obstacle to doing your work as a supervisor/coach?	☐ _____	
Sustainability	33. Are you compensated/paid to be a mentor?	☐ Is it sufficient compensation?	
	34. What incentives do you think would be helpful for supervisors/coaches to continue mentoring?	☐	
	35. What are your thoughts on the sustainability of this activity, supervising/coaching?	☐ What does sustainable look like?	
	36. Are you aware of the various phases of mentoring for RAMP?	☐ Try to determine how familiar the respondent is with the phase 1, 2 and 3 of mentoring outlined by RTI ☐ Do they have any comments about these?	
Coordination	37. Can you please share with me your perspective on the relationship between CADER and MOE supervisors/coaches	☐ Are there opportunities to co-supervisor/coach? ☐ What are the strengths of each group? ☐ What are the weaknesses of each group?	

Interview Questions			
Theme of Inquiry	**Question**	**INTERNAL**	
		Prompt	**Responses**
	38. Do you interact with any of the following groups, and if yes, in what capacity? CADER (if MOE) MOE (if CADER) RTI USAID Other _____	☐ If they can provide examples of these interactions	
Recommendations	39. In your view, what part of the supervising/coaching process is working well?	☐ Example of something working well	
	40. In your view, what part of the supervising/coaching process would you change?	☐ Example of something not working so well	
	41. If you could design a supervising/coaching activity, to benefit teachers, what would this look like?	☐ Please provide details if possible	
Closing Questions/Time Permitting/Background Information	42. What is your background and level of education?	☐ Was this person a teacher? ☐ Did they take additional classes to be a trainer/mentor?	
	43. Prior to this job, had you done supervising/coaching before and in what capacity?	☐ In their view what is their skill level	
STOP			

END of Interview THANK YOU:

[Thank your participant]

- **"We would like thank you for your time and for sharing with us your views about RAMP"**
- **"The information collected will be stripped of all personal identifiers"**
- **"Once all stakeholders have provided their views we will write a final report about the various perspectives of RAMP by each group. This work will help inform the ongoing RAMP Impact Evaluation"**

SUB-ANNEX D. ADDITONAL INTERVIEW DATA

Principals had varied understanding of score cards (quotes from principal interview data – pilot):

With regards to RAMP, did the school use the score card?
"The mentor conducted the assessment two days ago and I haven't received any card."
"There is no card."
"There is no card / it is only for schools that perform school evaluation and my school has no card."
"We have a score card / we have remedial plans for weak students and we have also met with their parents and most of them were illiterate."
"Yes."
"I have never heard of it."
"We do not have."

Principals provided other feedback regarding parents' involvement in RAMP (quotes from principal interview data – pilot) with mixed responses:

What was the Parents Feedback on the RAMP Initiative?	How did the RAMP Initiative Impact the Relation of Students' Families (Parents) with the School?
"There is a lack of feedback from the parents, despite of the meetings and clarifying RAMP to them, they did not show any response."	"Normally, parents would visit to thank the principal and teachers."
"A meeting, we introduced them to RAMP and they were excited and thanked us."	"I would like to gather parents and teach them how to teach their children with the new methods and how to help them."
"As an encouragement we honored mothers whose kids read a lot."	"A positive impact, now parents come to take the new thing."
"They were excited about the volunteering of attending the morning routine or story telling or hand crafts in the school events and Ramadan lantern after the community participation."	"A positive relation even before the initiative."
"There was no response, my interactions with parents is more in regard to the issues that their children might face at school."	"It had no impact due to lack of response from their part."
"Positive feedback, but there is a percentage of family disintegration hindering the student's benefit from the initiative."	"We daily have 10 parents with inquiries."
"They become more interactive and participate in making teaching aids and meetings."	"The visit is conducted once or twice a month. The supervisor comes and attends a part of the class, gives advice and direction to the teacher. The supervisor sometimes conducts model classes, follows-up with the weaker teachers and shares her notes and the Ministry's results with me."
"Negative; in mothers meeting, the mothers went out because the trainer is a man. Positive; it improves the relationship between children and their parents."	"…during parents meeting, we distribute cards to parents about the desires and needs to see how we can address it"

Teacher interview data regarding parental involvement:

"How did the RAMP Initiative Impact the Relation of Students' Families (Parents) with the School?" (quotes from teacher interview data - pilot)
"It did not. There was a student's mother who volunteered to participate with the second grade, but unfortunately she was busy because she is a member of the Development Board and has no time. Parents have a great role, but unfortunately there is a lack of interest."
"We try with parents, but the area that we are in parents do not care. When we invited them, they did a group breakfast."
"When we invited them, they did a group breakfast."
"There is no interaction."
We have noticed a difference after we met with them, we explained the idea and they started to prepare for the activities.
We felt that the parents liked the initiative through community participation.
There is no impact.
It had an impact through the open meeting, there is also communication and community participation. The visits file has become a burden, we need to fill in visits a month and it is a voluntary work and we cannot force parents to participate.
It did not have an impact.
I have been asking parents to visit from the first day.
I did not feel any impact, when I invited parents only 9 out of 28 came. When I gave the mothers the participation paper none of them responded.
In parents' meeting the attendance is usually great and you feel that parents are interested, even their relationship with the school became better and their visits increased.
I did not notice any change because parents are not at all interested.
I did not notice any impact.
To me, I found that parents' participation very slightly increased.
Their communication increased with the teacher and the administration.
Mothers attended the mourning routine and helped students in crossing the street. Some mothers participated in the group breakfast and one of the teachers gave a lecture on food etiquettes and organization.
No change.
Although the principals have conducting a meeting for parents, there was no sufficient response.
They became more interactive and know better about their children's level and personality. We made an activity in the school in the form of a game where we would ask the mother about her sons' hobbies, things he likes and things he hates. After that, we ask the student the same questions away from the mother, and then we show the mother his answers, in many times their answers did not match. We also gave a lecture about the right teaching methods to avoid conflict between parents' and teacher's methods.
Community participation, now there is a Parents' Board and a Development Board … All schools participate and give the school teaching aids mostly made by parents.

Of the mentoring sessions observed, more than 5 supervisor/coaches were seen praising the teachers for their efforts. The following are some examples of types of comments by the supervisors/coaches to the teachers as observed by the data collection team:

- Teachers were observed providing multiple pathways to solving a problem
- Coaches were observed providing encouragement to the teachers during class when using the routines. One example observed included the mentor talking with the teacher about the division method specifically.

- One coach was observed providing instruction to the teacher on how to communicate with students of diverse levels.
- Another coach was seen giving suggestions to the teacher in class on the need to leave something written on the board at the end of the day and the important of visuals for the students.

	Focus on improving student level in math
	Offered more than one way to arrive at a solution
	The coach praised the teacher for the preparation and good modeling and how she offered the activity.
	Discussed with her the analysis and division method.
	The coach praised the routine implementation (modeling, mentor's directives, and direct practice).
	The coach praised the way the reinforcement was done and how the student participated.
	Pointed out the need not to focus on the skill of asking questions.
	The coach stressed merging the activities with the curriculum and the way she employs in the merging process
	He described to her the implementation method for all class levels to support the weak and enrich the outstanding.
	The coach described how the teacher communicates with the weak students and the right and proper ways for that.
	Asked to strengthen students and to apply RAMP steps in relation to reading skills and asked to remind of prior knowledge about class subject, to do a network of vocabulary and predictions, form visuals…the need to merge RAMP in curriculum
	The coach noted that it is preferable at the end that there is something visible on the board so that the solution gives students a chance to be creative.
	He asked about students' reading record. He asked about the refugees and their situation and considering the differences among the students
	He knew she did not take the reading and math initiative, and he explained the routine she should follow.
	He commented on the way of writing on the board so that the students would synchronize the head movement and the way to draw lines on the board regarding nice handwriting because it is model for all students.
	Coach/supervisor commented texts must be read loudly for the sake of comprehension and work at increasing student interaction in class through tools, breaking the ice and promoting their motives in the class as well as selecting a role model for the students on the board, and the need to use evaluation tasks in class.

Obstacles to Mentoring (interview data supervisors and coaches)

- 37.5 percent (6/16) supervisors/coaches found "no obstacles" to implementing RAMP mentoring.
- 12.5 percent (2/16) stated limitations to implementing their job due to costs – some schools are far, and it can be costly to visit them. Some visits do not result in mentoring sessions (teacher no present), and no recompense to MOE supervisor for additional tasks.
- 25 percent (4/16) stated either teachers were not present when they visited, or they moved schools.
- 12.5 percent (2/16) mentioned fingerprinting process for attendance and reporting attendance at the directorate level (this office can be often far from where the coach/supervisor is located).

- 19 percent 3/16 supervisors/coaches revealed transportation and distance to be an obstacle in implementing RAMP mentoring.
- 6 percent (1/16) felt male student only schools were not implementing RAMP.
- Another response 6 percent (1/16) found classroom size to affect a session because the routine takes much longer to implement in-class.
- Finally, some of the other obstacles mentioned included workload, paper work, teacher refusal to implement RAMP, and the presence of overlapping interventions, leaving little time to do RAMP routines

Mentors and coaches were asked to share an ideal design of mentoring that would benefit the teachers. Some of the responses matched the existing RAMP design, as presented by one supervisor/coach:

- "Contact the teacher and prepare for the visit in advance and work with the coordinating teacher so that the visit coincides with the teacher's schedule, and then inform the headmaster of the visit.
- Sit with the teacher to plan the class and identify the teacher situation.
- Plan the class, attend it, prepare a class situation if needed. Feedback.
- Meeting with teachers in person and individually or with all the teachers in the school depending on the situation
- Arrange for next visit and ask teacher to apply a model class and collect the data in paper or on the tablet
- Conduct model session by the teacher in front of other teachers in the school and outside it"

Other supervisors/coaches suggested:

- During the pre-visit period discuss with the teacher what is needed from the visit and specific questions/issues to review
- Increase communication between supervisor/coach with the teacher including direct and indirect modes of communication
- Have teachers exchange visits between classes (peer-to-peer)
- Provide mentors and coaches additional support for the teacher visit
- Select teachers for visits based on teacher excellence level
- Research and study the mechanisms for selecting and matching trainers/mentors/coaches
- Create new content based on the experience from prior visits
- Increase website visibility of RAMP and pick version to collect data (paper or tablet)
- hold meetings with the community and supervisors
- teacher training needs should be assessed before a specific type of training activity should be implemented – make a plan specific for the teacher's needs and skills
- Ask the teacher to build a self-development plan so the supervisor/coach can identify how best to assist the teacher

- Provide more freedom to the supervisor/coach on selection of teacher and session to attend
- Involve more the headmaster in the process

SUB-ANNEX E. RTI MENTORING DATA

Table 12 is a summary of the RTI mentoring data as it pertains to the number of visits logged by the mentors and teachers visited between 2015-2018 (first semester only). Because of the data quality, the team was only able to present descriptive information about this data set and the number of mentoring visits recorded by the IP. A total of 29,156 unique teacher-schools were identified.

TABLE V.12. RTI DATA FOR MENTORING BY SEMESTER

Dataset Shared	Freq.	Percent	Cum. Percent
2015-16 S2	2,569	8.81	8.81
2016-17 S1	8,470	29.05	37.86
2016-17 S2	8,843	30.33	68.19
2017-18 S1	9,274	31.81	100
Total	29,156	100	

As the IP for RAMP, RTI collected entry forms completed by mentors and coaches with the help of their partners. Per the earlier meetings between USAID, MSI, and RTI, the Mission asked the evaluation team to review the RTI mentoring data as part of the mentoring qualitative study.

The evaluation team was provided with a copy of the data for the following semesters: semester 2 (S2) 2015-2016, S1 and S2 for the school year 2016-2017, and partial data for S1 for the 2017-2018 school year. Through a preliminary review, the team created unique codes matching teacher code with school code to provide descriptive information about mentoring visits. RTI has conducted some additional scoring and created a measurement of effectiveness. However, this study did not utilize these nor review them based on the issues identified during early conversation with the IP about this data set.

The data is reflective of the RAMP roll-out design with Cohort 1 starting spring of 2016, the second semester of the school year 2015-2016. As the semesters move forward, the increase in teacher codes reflect the addition of new teacher cohorts.

When looking at the variable provided by RTI indicating whether an actual visit occurred, the number of total mentoring sessions (102,498) is less by 11,229 visits for **a total of 91,268 recorded sessions**. This result does not include data for the S2 2015-2016 data, which does not have the variable distinguishing between actual versus planned mentoring sessions. The data is missing the S2 2017-2018 semester for which teachers also received mentorship/coaching. Another limitation in the data lies in likely missing/not logged mentoring sessions, for those supervisors/coaches who did not submit a form after visiting a teacher, forms that could have been lost/misplaced, or other sources of data entry error.

TABLE V.13.: FREQUENCY OF MENTORING VISITS (RTI DATA)

Visit Number	Freq.	Percent	Cum.	Total Visits for Visit Number Category[36]
1	3,955	13.56	13.56	3955
2	5,171	17.74	31.3	10342
3	6,998	24	55.3	20994
4	5,319	18.24	73.55	21276
5	3,182	10.91	84.46	15910
6	2,716	9.32	93.77	16296
7	1,162	3.99	97.76	8134
8	429	1.47	99.23	3432
9	131	0.45	99.68	1179
10	59	0.2	99.88	590
11	8	0.03	99.91	88
12	8	0.03	99.94	96
13	10	0.03	99.97	130
14	2	0.01	99.98	28
15	2	0.01	99.99	30
18	1	0	99.99	18
Missing	3	0.01	100	-
Total	29,156	100		102,498

[36] The table represents the RTI data showing the total number of teachers per visit amount. For example, 13.6% of teachers had at least 1 mentoring visit.

REFERENCES

Austin, P. C. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." Multivariate Behavioral Research, vol. 46, no. 3, 2011, pp. 399-424. doi:10.1080/00273171.2011.568786

Kautz, Tim, Peter Z. Schochet, and Charles Tilley. "Comparing Impact Findings from Design-Based and Model-Based Methods: An Empirical Investigation." National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education, 2017.

McCaffrey, Daniel F., Greg Ridgeway, and Andrew R. Morral. "Propensity score estimation with boosted regression for evaluating causal effects in observational studies." Psychological methods vol. 9, no. 4, 2004, p. 403.

MESP. "RAMP Impact Evaluation Baseline Report." Amman, Jordan. May, 2017.

MESP. "RAMP Impact Evaluation Midline Report." Amman, Jordan. November, 2017.

Ridgeway G, Daniel McCaffrey, Andrew Morral, Beth A. Griffin and Lane Burgette. TWANG: Toolkit for Weighting and Analysis of Nonequivalent Groups. R package version 1.4-9.5. 2016.

Schochet, P. Z. "The late pretest problem in randomized control trials of education interventions." Journal of Educational and Behavioral Statistics, vol. 35, no. 4, 2010, 379-406.

Miller, C. Catalina Torrente, Heather Gordon, and Arif Mamun. "Estimating Impacts of Early-Grade Reading and Math Project (RAMP) in Jordan: Evaluation Design Report." Cambridge, MA: Mathematica Policy Research, 2016.

What Works Clearinghouse (WWC) (2014). Procedures and standards handbook, Version 3.0. Washington DC: Institute of Education Sciences.

What Works Clearinghouse (WWC) (2017). Standards Handbook, Version 4.0. Washington DC: Institute of Education Sciences.

Early Grade Reading and Mathematics Initiative

Endline Survey Report

This publication was prepared for review by the United States Agency for International Development. It was prepared by RTI International.

Early Grade Reading and Mathematics Initiative

Endline Survey Report

Cooperative Agreement Number: AID-278-A-15-00003

Prepared for
USAID/Jordan
Noor Majdalani, AOR
Office of Education & Youth

Prepared by Jonathan Stern, Ahmed Abdelgawad, Patrick Fayaud and Rula Al-Jundi
RTI International
3040 Cornwallis Road
Post Office Box 12194
Research Triangle Park, NC 27709-2194

Table of Contents

Page

List of Figures

List of Tables

Abbreviations

AAM	Assessor Accuracy Measuring
AMELP	Activity Monitoring Evaluation and Learning Plan
C1	Cohort 1
C2	Cohort 2
C3	Cohort 3
CADER	ChangeAgent for Arab Development and Education Reform
CISLE	Cultivating Inclusive and Supportive Learning Environments
DFID	United Kingdom's Department for International Development
EdData II	Education Data for Decision Making II
EGMA	Early Grade Mathematics Assessment
EGRA	Early Grade Reading Assessment
EMIS	Education Management Information System
ERSP	Education Reform Support Project
G2	Grade 2
G3	Grade 3
GL	Goal-level (indicator)
IR	Intermediate Result
K1/KG1	Kindergarten 1
K2/KG2	Kindergarten 2
L1	Level 1
L2	Level 2
LQAS	Lot Quality Assurance Sampling
M1	Module 1
M2	Module 2
M3	Module 3
M&E	Monitoring and Evaluation
MoE	Ministry of Education
MSI	Management Systems International
ORF	Oral Reading Fluency
PhD	Doctor of Philosophy
QRTA	Queen Rania Teacher Academy
RAMP	Early Grade Reading and Mathematics Initiative, Jordan
RTI	RTI International (a registered trademark and a trade name of Research Triangle Institute)
SSME	Snapshot of School Management Effectiveness
UKAID	United Kingdom's Department for International Development
USAID	United States Agency for International Development
WLR	We Love Reading

Acknowledgements

The authors want to acknowledge the important contributions of many people who have made the Early Grade Reading and Mathematics Initiative (RAMP), as well as the endline survey and report possible, particularly the following:

- The Ministry of Education (MoE), especially Eng. Wafa Al-Abdallat (Education Training Center) who provide invaluable support to RAMP overall and Dr. Hafs Mallouh who played a coordinating role from the outset; without their involvement and commitment, RAMP and this study would not have been so successful.
- In the Education Training Center, Dr. Sami El-Mahasis, Dr. Juma'a al-Sound, Dr. Mohamed Al-Jdou'a, Dr. Thaera Abu Dayyeh, Dr. Khetam El-Sawarees, Dr. Ahmad Masafeh, and Dr. Mohamed El-Zouby, who all played important roles in the design and planning of the national survey.
- In the Examination and Testing Directorate, Dr. Nawaf Al-Ajarma, Dr. Mohammad Kenana, and Dr. Ahmad Al-Ajarmeh, who led the successive activities and guaranteed the scientific validity of the survey approach.
- In the Queen Rania Center (OpenEMIS), Dr. Ruba Al-Omari, Dr. Marwan El-Torman, Dr. Essam El-Kfaween, and Dr. Ali Mahasis, who contributed to the successful completion of the whole process.
- The MoE team of trainers who did excellent work, rigorously preparing the assessors to collect quality data. This team included Dr. Thaera Abu Dayyeh, Dr. Ghada Al-Smady, Dr. Linda El-Saed, Dr. Fadi Abou Jouda, Dr. Saeda Abou Ouda, and Dr. Ahmad Al-Ajarmeh.
- Dr Wael Salah (Minia University, Egypt) and Kellie Betts (RTI International), who provided support to the training and ensured that the survey met international standards.
- Dr. Khalid Dajani, Dr. Samer Ghannam, and the team at Dajani Consulting, who managed the logistics of the endline survey.
- Many MoE staff, particularly the MoE Supervisors (too many to mention by name), who contributed to implementation of RAMP. Without their care, commitment, and enthusiasm, this intervention would not have been possible.
- Christine Pagen, Noor Majdalani, and Angie Haddad (United States Agency for International Development [USAID]/Jordan), who showed interest in and commitment throughout RAMP and this study.
- USAID, Jordan and the United Kingdom's Department for International Development (DFID or UKAID), for funding RAMP.
- RAMP partners Queen Rania Teacher Academy (QRTA), ChangeAgent for Arab Development and Education Reform (CADER), We Love Reading (WLR), Dajani Consulting. Kaizen, MercyCorps, and Prodigy Systems for providing leadership, technical expertise, and professionalism that were critical to RAMP's success.

- Maitri Punjabi, who performed the data cleaning, data processing, and analysis needed to create and manage the vast data set associated with this study.
- Michelle Ward-Brent, David Harbin, and Julianne Norman from RTI, who provided RTI Home Office oversight and management of RAMP.
- The RAMP team, particularly Dr. Ahmed Abdelgawad for his leadership and dedication in planning and implementing the national endline survey with the RAMP monitoring and evaluation (M&E) team, including Dr. Farouq Banihamad and Jumana Youssef. Rula Al-Jundi, Taghreed Souleyman, and Mohamed Badar also played critical roles in reviewing the instruments and training the assessors.
- Amy Morrow and Gail Hayes from RTI, for the editing and layout of this report.
- Finally, this work would not have succeeded without the cooperation and contributions of the school principals, teachers, students, and communities included in the study, who, for obvious reasons, must remain anonymous.

Executive summary

This report presents the findings of the Jordan Early Grade Reading and Mathematics Initiative (RAMP) endline survey conducted at the end of the 2018–2019 school year (in May 2019).

Background

USAID/Jordan, in partnership with the Jordanian Ministry of Education (MoE), contracted with RTI International in 2011 under the Education Data for Decision Making II (EdData II) project to conduct the Snapshot of School Management Effectiveness (SSME), Early Grade Reading Assessment (EGRA), and Early Grade Mathematics Assessment (EGMA). In response to the findings of the 2012 National Survey, it was decided to develop an intervention pilot program that would support teachers in providing deliberate, structured, and developmentally appropriate daily instruction to develop students' foundational skills for reading and mathematics. The intervention was implemented during the 2013/2014 academic year by more than 400 teachers in 347 classrooms across 43 schools, reaching approximately 12,000 students. An endline survey to measure the impact of the intervention pilot was conducted in May 2014. In the following year, the Remedial Pilot Research Activity was developed, in order to assist teachers in providing support to students performing below general performance levels of the class.

In response to findings from all of these activities, RAMP began on January 1, 2015 (scheduled for 5 years with an end date of December 31, 2019). Supported by USAID and (DFID, or UKAID), RTI is leading RAMP implementation and capacity building support for the Jordanian MoE.

RAMP is a nationwide effort designed to improve the reading and mathematics skills of students in Jordan in kindergarten 2 (K2) through grade 3 (G3). RAMP has been carried out for nearly 5 years and has delivered improved reading and mathematics instruction to all public school students in Jordan in grades K2–G3—approximately 400,000 students. RAMP's effectiveness and impact were evaluated by means of, among other activities, midline (2017) and endline (2019) surveys. A midline study was conducted in May 2017 to measure RAMP's impact in its first 2 years. This endline study was conducted in May 2019 to measure RAMP's impact to date and the progress of the initiative toward the RAMP indicator targets.

Methodology

Determining that it was not necessary to conduct a separate baseline for RAMP, the MoE and USAID agreed that it would be best to use data and results from the 110 control schools involved in the 2014 National Survey as the baseline for RAMP indicators. A total of 2,159 students from the 110 control schools were assessed in 2014. The 2017 midline study involved 240 schools (20 per governorate) and 4,769 students (20 students per school). For the 2019 endline survey, the study population (2,193

schools) contained an estimated 243,333 G2 and G3 students from which the sample was drawn based on three stages sampling (school, class/teacher, student). To account for disproportionate sampling to and ensure that results were representative of the national population, weights were calculated as the inverse of the selection probability for each student.

Given the MoE's interest in better understanding the impact of the Syrian refugee situation on education in Jordan, the endline survey also included a sample of students from "special" schools (i.e., War Child, Syrian only, and Refugee Camp schools).

Student performance data were collected at all three time points via EGRA and EGMA. Additional information on RAMP implementation was collected at both midline and endline through the administration of three survey instruments: 1) oral student questionnaire; 2) oral teacher questionnaire; 3) school or principal questionnaire.

It is important to note that although the 2014 data allow national-level point estimates to be calculated reliably, the levels of disaggregation from those data are limited. Most importantly, although the 2017 and 2019 data can be used for analyses at the governorate and cohort levels, the 2014 data do not enable such comparisons between 2014 and 2017/2019. Therefore, results are compared across baseline, midline and endline wherever possible but some comparisons can only be made from midline to endline.

Findings

Overall, RAMP successfully produced significant improvements in G2 and G3 student reading and mathematics performance from baseline to endline, across foundational and higher-order skills. The impacts were typically larger for G2 students. This finding is not surprising given the cohort-designed rollout of the program, which led to nearly one-third of G3 classrooms only implementing RAMP methodology during the final year of the program. However, G3 students from C1 (i.e., students and teachers with the greatest RAMP exposure) exhibited some of the largest gains overall.

EGRA Results

As displayed in ***Table ES 1***, reading gains from baseline to endline were statistically significant for subtasks measuring the three most foundational reading skills (i.e. letter sounds, syllable sounds, and invented words) but there was no impact found for oral reading (fluency or accuracy). This may be due, in part, to the lack of attention paid to ORF among teachers and supervisors. It seems, however, that the comprehension-focused work was quite successful, as evidenced by the gain in the EGRA reading comprehension subtask (from 33.8% to 38.6%) and the high silent reading comprehension scores at endline—which led to an estimated 60% of G2 and G3 students reading with comprehension on the task. These trends were nearly identical across G2 and G3.

Table ES 1. Overall EGRA Results by Year (G2 and G3 Students)

Subtask	Measure	Baseline G2 and G3	Midline G2 and G3	Endline G2 and G3
Letter sound	fluency (correct letters per min.)	37 [±3.9]	47.9 [±2.0]	50.1* [±1.9]
	% correct of items attempted	64.9% [±5.4]	76.1% [±2.2]	79.1%* [±2.1]
Syllable sound	fluency (correct syllables per min.)	25.2 [±2.0]	31.7 [±1.7]	33.0* [±1.7]
	% correct of items attempted	63.8% [±3.7]	72.5% [±2.4]	71%* [±2.3]
Invented words	fluency (correct words per min.)	8.9 [±.9]	12.2 [±.7]	15.1* [±.8]
	% correct of items attempted	39.4% [±3.5]	49.3% [±2.3]	57%* [±2.1]
Listening comprehension	% correct	67.2%* [±2.2]	60.6% [±2.0]	58% [±2.1]
Oral reading	ORF	26.6 [±2.2]	25.9 [±1.4]	27.1 [±1.7]
	% correct of items attempted	63.7% [±4.4]	60.4% [±2.6]	60.1% [±2.7]
Reading comprehension	% correct of items attempted	45.3% [±4.4]	49.8% [±2.8]	49.3% [±2.9]
	% correct	33.8% [±3.5]	36.7% [±2.4]	38.6%* [±2.9]
	% of students with 80% comp.	17.9% [±3.5]	21.3% [±2.4]	23.6%* [±3.0]
	% of students with 80% comp on silent reading			60.5% [±3.8]

* p<.05 (baseline vs. endline)

EGMA Results

Although gains in mathematics were small from baseline to midline, implementation changes and redoubled efforts for mathematics training appeared to have a positive impact, leading to strong gains in the final 2 years of the program. As shown in ***Table ES 2***, statistically significant gains were ultimately found on all mathematics subtasks from baseline to endline. While the gains were relatively small for the more foundational skills (i.e., because the baseline scores were quite high), the gains for conceptual skills, which is the focus of RAMP, were larger. Improvements were particularly large in addition and subtraction L2 (41.9% to 52.1%), as well as in word problems (57.6% to 63.6%). As with reading, overall mathematics gains were similar across G2 and G3.

Table ES 2. Overall EGMA Results by Year (G2 and G3 Students)

Subtask	Measure	Baseline G2 and G3	Midline G2 and G3	Endline G2 and G3
Number Identification	% correct of items attempted	88.1% [±1.9]	92.6% [±.8]	92.2%* [±1]
Quantity Comparison	% correct	78.9% [±2.3]	83.7% [±1.4]	85.1%* [±1.4]
Addition L1	fluency (correct items per min.)	11.9 [±.5]	12.5 [±.3]	13.1* [±.4]
Subtraction L1	fluency (correct items per min.)	9.5 [±.5]	10 [±.3]	10.4* [±.3]
Addition and Subtraction L2	% correct	41.9% [±2.8]	45.9% [±2.2]	52.1%* [±2.5]
Missing Number	% correct	60.3% [±2.9]	64.5% [±1.9]	64.1%* [±2.2]
Word Problems	% correct	57.6% [±2.7]	59.9% [±2.0]	63.6%* [±2.4]

* p<.05 (baseline vs. endline)

RAMP Indicator Results

One of the key purposes of this endline survey was to determine how RAMP student achievement improved compared to a range of preset indicators (as published in the RAMP Activity Monitoring Evaluation and Learning Plan (AMELP)). The results show that RAMP was able to meet two endline targets (based on 95% confidence intervals)—GL_02 and GL_07—both of which show impressive improvements in the mathematics proficiency of students in the early grades (***Table 11***). While the remaining five targets were not met, statistically significant improvement occurred from baseline to endline for three targets: GL_01 (G2 reading proficiency), GL_04 (G3 math proficiency), and GL_06 (G2/3 reading proficiency). Small gains in indicator GL_03 (G3 reading comprehension) were achieved, but they were not statistically significant. The confounding finding from ***Table 11*** is the increase in oral reading zero scores for students at the end of G2 (i.e., GL_05). It is interesting to note that this increase in zero scores occurs in the same population of G2 students among whom overall reading fluency and comprehension benchmark results (GL_01) increased significantly. This issue is explored further in Chapters 3 and 4 of the main report.

Table ES 3. Summary of Performance on RAMP Indicators by Year

Indicator Number	Indicator[1]	Baseline Value	Midline Value	Endline Target	Endline Value
GL_01	*Proportion of learners who, by the end of two grades of primary schooling, demonstrate reading fluency and comprehension of grade-level text.*	7.9%	11.4%	22%	13.5%*
GL_02	*Proportion of students who, by the end of two grades of primary schooling, demonstrate that they can do grade-level mathematics with understanding.*	9.8%	11.7%	17%	18.7%*
GL_03	*Proportion of students who, by the end of three grades of primary schooling, demonstrate that they can read and understand grade-level text.*	29.0%	31.5%	50%	33.3%
GL_04	*Proportion of students who, by the end of three grades of primary schooling, demonstrate that they can do grade-level mathematics with understanding.*	19.5%	21.2%	34%	29.2%*
GL_05	*Proportion of students obtaining zero scores on ORF at the end of G2.*	12.5%*	12.6%	6%	20.7%
GL_06	*Proportion of G2 and G3 students who demonstrate that they can read grade-level text with comprehension.*	17.9%	21.3%	35%	23.6%*
GL_07	*Proportion of G2 and G3 students who demonstrate that they can do grade-level mathematics with understanding.*	14.4%	16.0%	25%	24.1%*

* p<.05 (baseline vs. endline)

Performance by Key Indicators and Subgroups

With regard to gender, baseline (and midline) results showed that reading performance of female students was higher than that of male students, whereas male students performed slightly better in mathematics. By endline, however, the gap between female and male students closed considerably, leading to non-significant differences for nearly all indicators.

Governorate-level data show that changes in performance were not geographically consistent—with some governorates producing large gains, and others stagnating in their growth. For example, in terms of reading proficiency, only Aqaba and Karak showed statistically significant gains in both G2 and G3, with Karak achieving the largest gains overall. Mafraq produced significant gains in G3 only. Madaba showed marginally significant gains in G2, while Aljoun and Jarash had marginally significant gains in G3. These results may be tied in part to cohort membership, but additional investigation into the differential performance of governorates (i.e., to learn what does and does not lead to successful improvement) is warranted.

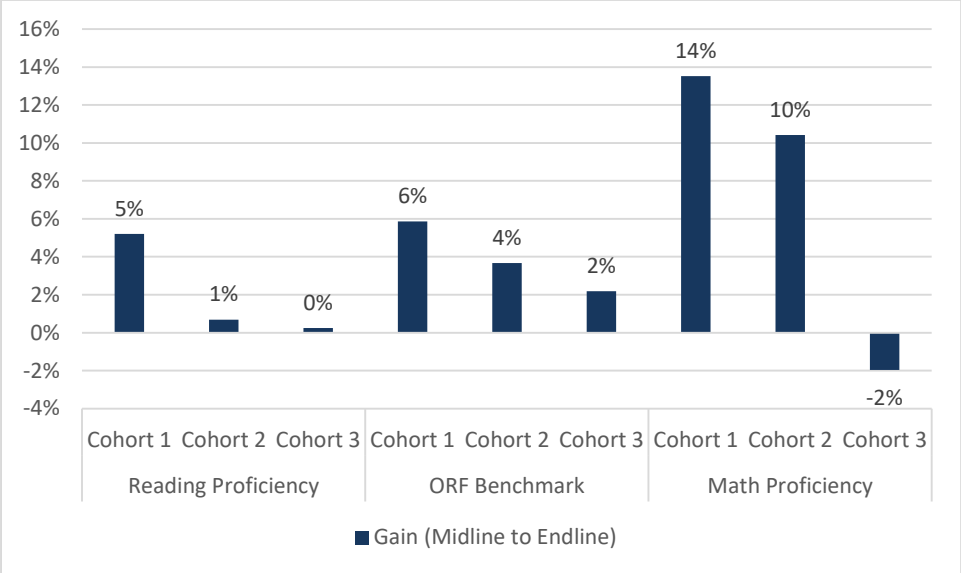
[1] For the purposes of these indicators and throughout the report, *reading with comprehension* and *doing mathematics with understanding* are defined as follows:

- To *read with comprehension*, the student must score at least 80% on the EGRA reading comprehension subtask.
- To *do mathematics with understanding*, the student must score at least 80% on the EMGA addition and subtraction (L2) subtask and at least 70% on the EGMA missing number subtask.

In terms of cohorts themselves, results show that Cohort 3 (C3) students performed significantly below Cohort 1 (C1) and Cohort 2 (C2) students at midline for nearly all indicators. At endline, C3 students in G2 were able to narrow the gap slightly in reading but C3 students struggled to produce gains in G3 (reading or math).

The overall gains for G3 performance by cohort (for three key indicators) are displayed in ***Figure ES 1. G3 Performance by Indicator, Cohort, and Year***. Clearly, G3 students in C1 showed the largest gains in reading proficiency, ORF benchmark, and math proficiency. Because G3 students and teachers in C1 had the longest exposure to RAMP methodologies, it is not surprising that they showed the largest gains. This finding indicates that longer exposure to the program may be essential for translating into larger learning gains, once both students and teachers are familiar with the teaching and learning process.

Figure ES 1. G3 Performance by Indicator, Cohort, and Year



A comparison of Jordanian and Syrian students in traditional, public schools showed that, counter to RAMP assumptions, Syrian refugee students were not outperformed by Jordanian students (and, in some cases, they actually showed larger gains than Jordanian students from midline to endline). However, the increasing number of Syrian refugee students in traditional schools is likely to have adversely impacted program progress, as it has led to an increasing proportions of two-shift schools, which dedicate less time to teaching reading and mathematics, leading to significant performance gaps between morning shift and all-day schools.

In terms of school type, results of students in traditional, refugee camp, Syrian day, and War Child Program schools were compared during the endline. Unsurprisingly, the lowest-performing students in reading and mathematics were in the refugee camp schools; the other three school types showed somewhat similar performance, except for the lower reading proficiency mark in War Child Program schools.

Key Student, Teacher, and School-Level Characteristics

In order to accommodate recent increases in the student population, the use of two-shift schools has expanded considerably in the past two years (with two-shift schools accounting for 28% of schools at endline, as compared to only 15% at midline). Furthermore, based on the teachers sampled in this study, it was estimated that the proportion of substitute teachers more than doubled from midline to endline (from 7% to 19%). Additionally, the proportion of classrooms that had a single teacher for the duration of the year decreased (from 92% to 86%), while the proportion of classrooms that had three or more teachers for the year increased by 5%. These system-level changes undoubtedly pose a challenge for RAMP or any intervention seeking to effect large-scale change in teacher and student performance.

With regard to RAMP-specific measures, the proportion of teachers who attended all days of all three RAMP teacher trainings increased significantly from baseline to endline (31% v. 67%) but 10% of teachers in G2 and G3 classrooms still did not attend any RAMP training at all. Additionally, only 1 in 10 teachers at endline reported that the RAMP materials (e.g., teacher's guides, student activity books) arrived on time. Despite these challenges, more than 90% of teachers reported that they understood RAMP goals and had confidence in implementing the methodology as intended, while more than 80% of teachers felt that RAMP improved math performance, reading performance, and student enthusiasm for learning in their classrooms.

At the student level, significant increases were found in the proportion of students who have time to read books in the classroom or library every day (from 58% at midline to 64% at endline), the proportion of students who bring books home from school (from 55% to 70%), and the proportion of students who meet with others in the community and listen to someone read (from 37% at midline to 45% at endline). All these results indicate increased access to print materials for students supported by RAMP but the percentages remain below ideal levels.

Conclusions

RAMP successfully produced reading and mathematics performance gains for students in both G2 and G3 over the life of the program (2015–2019). The largest reading gains were achieved in the more foundational letter sound, syllable sound, and invented word subtasks. There were no significant gains in oral reading scores throughout the program but but reading comprehension scores did improve. Perhaps most intriguing of all is the estimated 60% of students found to be reading with comprehension based on the silent reading task. Nevertheless, progress in reading has been slower than anticipated and below the target levels set by the program. MoE officials noted that one possible explanation for the slow progress was the fact that the number of periods (or lessons) allocated to Arabic language per week was recently reduced from nine to seven.

Regarding USAID performance indicators, RAMP was able to meet two endline targets, both of which relate to mathematics. After a slow start and lack of progress from baseline to midline, RAMP introduced a mathematics booster training to teachers,

which focused on their conceptual understanding of early grade mathematics and appears to have positively impacted performance.

The performance of male students improved from baseline to endline and narrowed the gender gap considerably—which the MOE attributes, in part, to the increased focus on male students by female teachers. Generally, G2 classes showed greater progress than G3 classes; similarly, C1 showed the strongest performance and gains of any cohort—providing evidence of the importance of longer exposure to RAMP programming. In contrast, C3 had the lowest results overall, which is attributed to the short duration of RAMP exposure (less than 1 year), the less desirable locations for teacher placements, and the large geographical areas of the four governorates (Mafraq, Ma'an, Madaba, and Balqa), which are obstacles to providing consistent, timely, high-quality support.

Increases in the student population has led to similar increases in the proportion of two-shift schools. With morning-shift schools producing significantly lower results than all-day schools, it is reasonable to conclude that the reduced class time for students may be adversely affecting student reading and mathematics progress.

Overall, teachers and head teachers view RAMP favorably, with the overwhelming majority of head teachers believing that the program provided sufficient training, materials, and support to their teachers, and the majority of teachers stating that the program was effective in increasing student performance.

Despite the support for RAMP within schools, a few system- and program-level factors likely had a negative impact on progress in performance. Most notably, the majority of teachers claimed that they did not receive materials on time for the start of the school year. This has clear implications for teachers' ability to implement the methodology as intended. The increases in the proportions of non-permanent teachers and those without pre-service training in early grade reading or mathematics, and the constant percentage of teachers who did not receive RAMP training, in addition to those issues mentioned above, are all likely to have been obstacles the consistent and faithful implementation of the new methodology and pedagogical approach in the classroom. Thus, these areas should be a continued focus for the system in the coming years.

Recommendations

One trend that arose from examinations of the endline data was that high-performing students are continuing to improve while low-performing students are remaining behind and not progressing as intended. Typically, teachers tend to try to cover the whole curriculum and the entire textbook during the school year, without verifying whether students are understanding and keeping up with the learning pace. This practice is likely to have the greatest impact on poorer-performing students who fall behind early and never catch up. Accordingly, the system must make particular effort to ensure that teachers apply differentiated instruction in all grades and subjects. ***It is recommended that the MoE provide additional guidance on how to employ differentiated instruction approaches in the classroom.***

Due to issues arising from the reduction in the number of Arabic language lessons per week and the shortened school day in two-shift schools, ***the MoE should consider***

increasing the time allocated to reading in the curriculum and should instruct teachers to give students time to read a wider variety of text and make sure they are involved in more engaged reading activities. Additionally, it is important for the MoE to identify strategies to provide equal treatment and learning opportunities to all students, particularly those in two-shift schools. Furthermore, ***it is recommended that the MoE conduct a time-on-task study to assess the amount of time in a school day that students actually learn and work actively and the amount of time students spend reading and engaged with print materials throughout the day*** (providing evidence on how time is used during the day and where students could be provided with additional opportunities to read and engage with text).

Furthermore, classroom observations have shown that teachers tend to focus on and interact with only a small group of students during their lessons. Accordingly, ***the next training program should focus on the following:***

- Differentiated instruction techniques and remedial activities/remedial group management
- Classroom management techniques designed to ensure the involvement of all students in the learning process
- Formative evaluation, which will help teachers to adjust instruction to the actual learning pace of students and differentiate instruction as needed.

The data reflect a positive association between reading outside the classroom and student performance. Therefore, ***it is recommended that the Reading Incentive Program be a focus of additional support***, including examinations of the fidelity of implementation and use by teachers and schools.

Lastly, it is clear from the endline results that the RAMP model works better in some governorates than in others. This variation is partly attributable to the level of fidelity with which the program is being implemented. High fidelity of implementation will ultimately only be achieved with similarly high levels of accountability and support. Therefore, ***principals should play an increasing role in monitoring teachers and supporting them to ensure that they apply new teaching methodologies***, including differentiated instruction. Additionally, ***supervisors should be held accountable for their coaching activities and the quality of the support they provide to teachers***, and ***field directorates should be held accountable for monitoring performance and implementation fidelity***.

The education system in Jordan is strong, and the RAMP methodology is sound. Working closely with the MoE at all stages, the program was able to improve instruction in the classroom and produce student performance gains in both reading and mathematics. The recommendations laid out in this section are intended to provide the MoE with guidance on how best to continue to build upon the successes of RAMP and how to overcome the obstacles encountered to ensure that all students in the country are provided with the opportunities to learn and succeed they deserve.

1 Background

This report presents the findings of the Jordan Early Grade Reading and Mathematics Initiative (RAMP) endline survey conducted at the end of the 2018–2019 school year (in May 2019). This chapter provides a background to both RAMP and the RAMP National endline study.

1.1 2012 National Survey

United States Agency for International Development (USAID)/Jordan, in partnership with the Jordanian Ministry of Education (MoE), contracted with RTI International in 2011 under the Education Data for Decision Making II (EdData II) project to conduct the Snapshot of School Management Effectiveness (SSME), Early Grade Reading Assessment (EGRA), and Early Grade Mathematics Assessment (EGMA). The purpose was to gain insight into student facility with foundational skills and to better understand the characteristics of Jordanian schools that are associated with student performance. These assessments were conducted in a sample of primary schools in Jordan at the end of the 2011/2012 academic year. The hope was that evidence resulting from the survey could inform future education policy decisions, as needed.

1.2 2013/2014 Intervention Pilot Research Activity and 2014 National Survey

In response to the findings of the 2012 National Survey, it was decided, after discussions with the MoE Curriculum Team and Senior Reading and Mathematics Supervisors, to develop an intervention pilot program that would support teachers in providing deliberate, structured, and developmentally appropriate daily instruction to develop students' foundational skills for reading and mathematics.

The intervention was implemented during the 2013/2014 academic year by more than 400 teachers in 347 classrooms across 43 schools, reaching approximately 12,000 students. An endline survey to measure the impact of the intervention pilot was conducted in May 2014. The Jordan Intervention Impact Analysis Report[2] noted that the percentages of non-readers or beginning readers and non-mathematicians or early mathematicians in the control group remained relatively consistent between baseline and endline. However, substantial reductions in the proportions of non-readers or beginning readers and non-mathematicians or early mathematicians were observed in treatment schools (from 32% to 19% in reading and 30% to 22% in mathematics). Additionally, while the proportions of readers and mathematicians remained constant in control schools, these values increased significantly in the treatment schools (from 13% to 24% in reading and 14% to 24% in mathematics).

[2] Brombacher, A., Stern, J., Nordstrum, L., Cummiskey, C., & Mulcahy-Dunn, A. (2015). *Education Data for Decision Making (EdData II): National Early Grade Literacy and Numeracy Survey–Jordan. Intervention impact analysis report.* Retrieved from https://jordankmportal.com/resources/jordan-intervention-report-english

1.3 2014/2015 Remedial Pilot Research Activity

In 2013, in response to the findings of the 2012 National Survey, it was decided to develop a Remedial Pilot Research Activity that would enable teachers to provide support to the students in their classes who are performing below the general performance level of the class.

The Remedial Pilot Research Activity was piloted in the first three grades at 41 treatment schools by 308 teachers during the 2014/2015 academic year, reaching approximately 10,000 students. The remedial diagnostic tools were also administered in the first three grades at 16 control schools, at both the start and the end of the 2014/2015 academic year, to determine the impact of the Remedial Pilot Research Activity. The assessment data revealed that remedial students benefited from the remedial support, in both reading and mathematics.

1.4 Early Grade Reading and Mathematics Initiative (RAMP)

In response to findings of the 2012 and 2014 surveys and the success of the Intervention and Remedial Pilot Research Activities, RAMP began on January 1, 2015, and was scheduled for 5 years, ending on December 31, 2019. Supported by USAID and the United Kingdom's Department for International Development (DFID, or UKAID), RTI is leading RAMP implementation for the Jordanian MoE.

RAMP is a nationwide effort designed to improve the reading and mathematics skills of students in Jordan in kindergarten 2 (K2) through grade 3 (G3). The initiative works with the MoE to (1) develop and distribute improved learning materials to every K2–G3 classroom in Jordan; (2) train teachers, principals, supervisors, and field directorate and MoE administrators to provide more effective instruction; (3) promote community participation in reading and mathematics education; and (4) support nationwide adoption of early grade reading and mathematics policies, standards, curricula, and assessments. RAMP has been carried out for nearly 5 years and has delivered improved reading and mathematics instruction to all public school students in Jordan in grades K2–G3—approximately 400,000 students. In addition to improving early grade reading and mathematics performance, RAMP is charged with ensuring the integration of gender, disability, and refugee issues as cross-cutting themes into its activities and deliverables. The initiative was formally launched by Her Majesty Queen Rania Al-Abdullah as part of the broader MoE initiative to improve education. RTI and its lead partners Queen Rania Teacher Academy (QRTA), ChangeAgent for Arab Development and Education Reform (CADER), and We Love Reading (WLR) implement RAMP.

RAMP is capitalizing on the contributions and successes of previous projects in Jordan. These include the materials and findings of the 2012–2014 Intervention Pilot Research Activity; the materials and results of the 2013–2015 Remedial Pilot Research Activity, the Cultivating Inclusive and Supportive Learning Environments (CISLE) in Jordan's Schools project, and the Education Reform Support Project (ERSP).

For the sake of efficiency, this report does not provide a detailed description of RAMP. A brief description is provided of the specific RAMP aspects that have a direct bearing

on the achievement of the key results described in the initiative's Activity Monitoring Evaluation and Learning Plan (AMELP).

1.4.1 RAMP Theory of Change and Results Framework

The 2012 and 2014 National Surveys demonstrated that the majority of students in the early grades in Jordan were not reading with comprehension or doing mathematics with understanding.[3] A range of conditions contributed to this situation. For reading, teachers were not specifically trained on early grade reading instructional techniques that develop phonemic awareness nor to provide deliberate instruction focused on phonics, fluency, vocabulary, and comprehension. For mathematics, teaching and learning were typically focused on memorizing facts, rules, and formulas instead of understanding and actively interacting with the subject matter.

RAMP is a response to these challenges for early grade education in Jordan. Its primary goal is to improve or change the learning outcomes for reading and mathematics from K2 to G3. The activity was constructed based on the following development hypothesis:

By investing in building MoE staff capacity, especially that of public school teachers and supervisors, to use appropriate materials; research-based pedagogies; and differentiated support to students according to their needs, RAMP will contribute to a substantially higher proportion of grade 2 and grade 3 public school students being able to read with comprehension and do mathematics with understanding by the end of the initiative. By also involving parents and communities in general in support of RAMP, the impact of the initiative will be significantly enhanced. Also, these gains will be sustained and built upon beyond the life of the initiative through the institutionalization of RAMP's research-based methodologies within a context of strong reflective practice.

RAMP's effectiveness and impact were evaluated by means of, among other activities, midline (2017) and endline (2019) surveys. These surveys were used to gather data to measure progress toward the following RAMP goal-level indicators:

Ind-GL-01: Proportion of students who, by the end of two grades of primary schooling, demonstrate reading fluency and comprehension of grade-level text.

Ind-GL-02: Proportion of students who, by the end of two grades of primary schooling, demonstrate that they can do grade-level mathematics with understanding.

Ind-GL-03: Proportion of students who, by the end of three grades of primary schooling, demonstrate that they can read and understand grade-level text.

[3] Throughout the report and for the purpose of the RAMP indicators, *reading with comprehension* and *doing mathematics with understanding* are defined as follows:

- To *read with comprehension*, the student must score at least 80% on the EGRA reading comprehension subtask.
- To *do mathematics with understanding*, the student must score at least 80% on the EMGA addition and subtraction (L2) subtask and at least 70% on the EGMA missing number subtask.

Ind-GL-04: Proportion of students who, by the end of three grades of primary schooling, demonstrate that they can do grade-level mathematics with understanding.

Ind-GL-05: Proportion of students obtaining zero scores on oral reading fluency (ORF) at the end of grade 2 (G2).

Ind-GL-06: Proportion of G2 and G3 students who demonstrate that they can read grade-level text with comprehension.

Ind-GL-07: Proportion of G2 and G3 students who demonstrate that they can do grade-level mathematics with understanding.

RAMP has four intermediate results (IR), each with a range of indicators—the indicators for which data were also collected in the midline (2017) and endline (2019) surveys are included here:

IR 1. Improved early grade reading and mathematics learning materials are integrated into every K2–G3 classroom in Jordan.

IR 2. Teachers and administrators are better equipped to provide effective reading and mathematics instruction.

IR 3. Communities participate in the education of reading and mathematics for all children and hold schools accountable for results.

Ind.3.2.1:

Proportion of G2 and G3 students who report receiving parental support for learning to read and do mathematics in USAID RAMP-supported schools.

Ind. 3.2.2:

Percent of surveyed teachers reporting active parental engagement.

IR 4. The MoE institutionalizes early grade reading and mathematics policies, standards, and assessments.

1.4.2 RAMP activities

As the IRs listed in the previous section highlight, RAMP activities focus on materials development, in-service teacher training, community mobilization, and institutionalization of RAMP within MoE policies and practices.

RAMP materials development activities focus on the following:

1. Refining the teaching and learning materials for early grade reading and mathematics that were developed as part of the Intervention and Remedial Pilot Research Activities and the associated in-service training materials
2. Developing modules and training materials that address the RAMP cross-cutting themes (creating supportive classrooms for both genders, creating inclusive education environments for students with disabilities, and creating psychosocially supportive environments for students who have experienced trauma)

3. Developing materials (worksheets, training videos, and other resources) for a web-based community of practice
4. Developing pre-service and induction training materials on using the RAMP research-based pedagogies for developing early grade reading and mathematics skills.

The training activities focus on the following:

1. Providing in-service training for all active teachers teaching K2–G3 in all the public schools of the Kingdom
2. Developing an induction module on the RAMP methodologies to be integrated into the MoE's onboarding of teachers.

The community mobilization activities focus on the following:

1. Increasing parental involvement in the schools that their children attend and in their children's learning
2. Introducing volunteer-led, community-based reading groups.

Other initiatives the contributed to RAMP's progress are as follows:

1. Supporting the establishment of a reading incentive program
2. Providing a school readiness program.

The in-service training for teachers is the activity that likely had the greatest direct impact on students' reading and mathematics performance. This training activity is described in greater detail below.

The in-service training activity targets all K2–G3 teachers and takes the form of both workshop-based direct instruction and in-class coaching support. Workshop contents are based on direct instruction and are organized in three modules (Module 1 [M1], Module 2 [M2], and Module 3 [M3]). Broadly speaking, M1 covers assessing students' foundational skills in reading and providing effective instruction to develop these skills. M2 does the same for mathematics, and M3 addresses the cross-cutting themes (gender, disabilities, and psychosocial approaches). Each module is presented over 5 days, with M1 and M2 presented during the school holidays before the start of the first semester and M3 presented during the school holidays before the start of the second semester. Appropriately trained MoE Supervisors and RAMP coaches conduct the training. In addition to the workshop-based direct instruction, all participating teachers receive in-class coaching support from MoE Supervisors and specially appointed RAMP coaches. Optimally, each teacher will receive 6 coaching visits per semester (12 per year) during the year in which they start implementing RAMP.

RAMP is targeting almost 15,000 teachers (14,736 teachers at the time of writing this report). One of RAMP's early strategic decisions was to stagger the introduction in two ways. First, the Kingdom's governorates were divided into three groups: Cohort 1 (C1; approximately 26% of the teachers), Cohort 2 (C2; approximately 47% of the teachers), and Cohort 3 (C3; approximately 27% of the teachers). The staggering of the training for kindergarten 1 (K1)–G2 and G3 teachers was premised on two key points: (1) it ensures that when G3 teachers implement the RAMP methodologies, they do so with students who have already been exposed to the methodologies by their G2 teachers; and

(2) it allows new K2–G2 teachers who missed their training to be trained with the G3 teachers.

1.5 RAMP 2017 Midline Study

A midline study was conducted at the end of the 2016–2017 academic year—in May 2017—to measure RAMP's impact in its first 2 years. Because RAMP is fundamentally concerned with improving reading and mathematics, the RAMP midline survey included only EGRA and EGMA, in addition to student, teacher, and school-level interviews.

1.6 RAMP 2019 Endline Study

This endline study was conducted at the end of the 2018–2019 academic year—in May 2019—to measure RAMP's impact to date and the progress of the initiative toward the RAMP indicator targets.

Considering that (1) a National Survey was conducted at the end of the 2013–2014 academic year (in May 2014) to measure the impact of the Intervention Pilot Research Activity and (2) RAMP is, in many ways, an extension/expansion of the Intervention Pilot Research Activity, the MoE, understandably, felt that conducting a national baseline study specifically for RAMP at the start of the initiative in 2015 was not necessary. Accordingly, agreement was reached that the baseline values for the RAMP indicators would be based on the performance of the students in the control group of the 2014 National Survey. This decision does, however, create limitations on the levels of disaggregation at which comparisons can be made between the findings of this study and the 2014 findings. These limitations are discussed elsewhere in the report.

1.6.1 Endline Assessor Training

In line with RAMP's implementation principles—MoE capacity building, facilitating ownership, and ensuring sustainability—the endline training of assessors was co-led by RAMP and MoE staff. The training was held at the Ayass Hotel in Amman on April 6–17, 2019. In total, 14 RAMP and MoE trainers trained 113 assessors (including 32 supervisors), for the endline data collection across all 12 governorates. See **Annex 2** for more details on the training.

1.6.2 Early Grade Reading and Mathematics Assessments

To compare the results of the 2019 study with the results of the 2014 and 2017 studies, it is essential that the instruments are as psychometrically comparable as possible.

An instrument validation workshop was conducted in March 2019 to revise the midline EGRA instrument. The instrument adaptation workshop included representatives from a range of MoE departments. The following changes were made:

- The listening comprehension subtask was not changed because this subtask already showed relatively high scores at baseline and was not a focus of the

program. Instead of dropping the subtask because of its redundancy, it was retained for reasons of comparability.

- The reading passage and comprehension questions were minimally revised to keep the subtasks as similar as possible in terms of their length and level of difficulty while mitigating concerns about test leakage and memorization.
- For all other subtasks, the items in each line of the stimulus sheet were systematically rearranged. Items were not interchanged between the lines of the stimulus sheet.
- Finally, a new silent reading comprehension subtask was added at endline. In this subtask, students are given 2 minutes to read a passage silently and are then asked five comprehension questions about the story that they just read. This task was designed to align with the reading comprehension subtask administered over the past few years as a part of the lot quality assurance sampling (LQAS) activities.

Accordingly, the final EGRA subtasks included at endline were listening comprehension, letter sounds, syllable sounds, oral reading, reading comprehension, and silent reading comprehension. Subtasks on the EGMA instrument remained identical to those used at baseline and midline (i.e., number identification, quantity comparison, level 1 [L1] addition, L1 subtraction, level 2 [L2] addition, L2 subtraction, missing number, and word problems).

The EGRA and EGMA instruments used in the RAMP endline study (2019) are provided in **Annex 16**.

1.6.3 Additional Sources of Data

Data were gathered from a wide range of sources to gain a fuller understanding of the impact of RAMP's intervention and the variables that influence its effective implementation. These sources included the following:

- **Oral student questionnaire:** This questionnaire was completed by all students. In addition to asking students about their reading and general homework habits, the questionnaire included a range of questions related to typical wealth variables to allow the development of a wealth indicator for students in the data set. This indicator will serve as a control factor for any regressions calculated in the study. The wealth variable questions were revised since the 2017 version of this instrument to account for the contents of the most recent national household survey conducted in Jordan.
- **Oral teacher questionnaire:** This questionnaire was completed by the G2 and G3 teachers of the assessed students in all sampled schools. The oral questionnaire had two components: (1) questions about the teachers' background characteristics (e.g., experience, training) and (2) questions about their participation in and experience with RAMP. This questionnaire was revised from midline, and the number of questions was reduced.
- **School or principal questionnaire:** This was completed by the principal (or deputy principal) of each school. The questionnaire had three components: (1) questions about the school's characteristics (e.g., numbers of students and

teachers in each grade); (2) questions about student and teacher attendance characteristics, parent involvement, and whether quarterly Parent–Teacher Council meetings had taken place as scheduled; and (3) questions that dealt with the school's experience with RAMP. This questionnaire was revised from midline, and the number of questions was reduced.

The student, teacher, and school/principal questionnaires used in the RAMP endline study (2019) are provided in **Annex 16.**

2 Methodology

2.1 Sample

2.1.1 2014 Sample (Baseline)

The MoE and USAID decided that conducting a baseline study for RAMP was not necessary; instead, they agreed to use the data and results from the 110 control schools involved in the 2014 National Survey as the baseline for the RAMP indicators. A total of 2,159 students from the 110 control schools were assessed in 2014.

This decision does, however, have implications for this study. Key among these is that although the 2014 data allow national-level point estimates to be calculated reliably, and while the data can also be disaggregated at either the grade or gender level (but not both), no further disaggregation is meaningful. Thus, although the 2017 and 2019 data can be used for analyses at the governorate and cohort levels, the 2014 data do not enable such comparisons between 2014 and 2017/2019.

2.1.2 2017 Sample (Midline)

In 2017, the Jordan Education Management Information System (EMIS) unit provided a list of all public primary schools in the nation, totaling 2,538 schools. Of these, 348 schools were removed from the list because they did not have at least 20 G2 and G3 students combined. A further 240 schools were removed because they had been selected for and participated in the Impact Evaluation Baseline Study conducted by Management Systems International (MSI) in November and December 2016. Finally, an additional 99 schools were removed at the request of the MoE because they had been established recently. The final population consisted of 1,851 schools, from which a study sample was drawn.

The 2017 study involved 240 schools (20 per governorate) and 4,769 students (20 students per school). To account for the disproportionate sampling and make the sample representative of the national population, weights were calculated as the inverse of the selection probability for each student. All scores reported for this study were calculated using the student weights as noted.

2.1.3 2019 Sample

In 2019, the Jordan EMIS unit provided a list of the nation's 2,565 public primary schools. Of these, 347 schools were removed from the list because they did not have at least 20 G2 and G3 students combined. Twenty more schools were removed because they are military culture schools. Finally, an additional five schools were removed at the request of the MoE because they had been established recently. The final population consisted of 2,193 schools, from which a study sample was drawn. The study population (2,193 schools) contained an estimated 243,333 G2 and G3 students. ***Table 1*** summarizes the population from which the sample was drawn.

Table 1. Sample Population

Cohort	Region	Governorate	Schools	Excluded Schools	Non-Excluded Schools	Students in Non-Excluded Schools
1	North	Ajloun	86	16	70	5,258
1	North	Jerash	125	19	106	7,603
1	Central	Zarqa	256	43	213	39,035
1	South	Karak	176	22	154	12,433
2	North	Irbid	502	43	459	44,928
2	Central	Amman	499	57	442	71,130
2	South	Tafilah	76	9	67	4,492
2	South	Aqaba	58	21	37	5,325
3	North	Mafraq	390	76	314	25,291
3	Central	Balqa	166	24	142	15,113
3	Central	Madaba	102	23	79	5,944
3	South	Ma'an	129	19	110	6,781
		Total	**2,565**	**372**	**2,193**	**243,333**

In the first selection stage, the 2,193 schools were stratified by governorate to create 12 strata. Within each stratum, schools were sorted by school shift (morning shift, afternoon shift, and single shift) and the combined enrollment in G2 and G3. Twenty schools were then selected from each stratum with equal probability proportional to G2 and G3 enrollment.

The second selection stage involved sampling classes/teachers within each sampled school. One G2 class was randomly selected from the G2 classes with equal probability, and one G3 class was randomly selected from the G3 classes with equal probability.

The third selection stage involved random selection with equal probability of 10 students from each of the randomly selected classes. The second and third stages of selection were conducted by the trained assessors at schools on the assessment day.

Given the MoE's interest in better understanding the impact of the Syrian refugee situation on education in Jordan, the endline survey included a sample of students from "special" schools (i.e., War Child, Syrian only, and Refugee Camp schools).

Table 2 summarizes the sampling process, excluding special schools.

Table 2. Sampling Stages

Stage Number	Item Sampled	Stratified by…	Probability of Selection
Stage 1	Schools (240)	**Governorates (12)** 20 schools/governorate	Equal probability proportional to G2 and G3 enrollment
Stage 2	G2 classrooms (480) G3 classrooms (480)	**Grade (G2 and G3)** 1 G2 classroom per school 1 G3 classroom per school	Equal
Stage 3	G2 students (4,800) G3 students (4,800)	No stratification	Equal

2.2 Descriptive Statistics

Table 3 summarizes the data collected during the survey across governorates, grades, and gender (excluding special schools).

Table 3. Descriptive Statistics for the Final Sample

Cohort	Region	Governorate	Students					Teachers	Schools
			G2		G3				
			Female	Male	Female	Male	Total		
1	North	Ajloun	139	65	131	64	399	40	20
1	North	Jerash	118	80	132	70	400	40	20
1	Central	Zarqa	124	87	118	70	399	37	20
1	South	Karak	127	72	124	76	399	39	20
2	North	Irbid	119	84	120	75	398	40	20
2	Central	Amman	97	93	118	92	400	39	20
2	Central	Tafilah	108	88	109	92	397	38	20
2	South	Aqaba	122	65	134	79	400	37	20
3	North	Mafraq	106	95	116	83	400	40	20
3	Central	Balqa	118	76	119	87	400	39	20
3	Central	Madaba	132	65	129	68	394	38	20
3	South	Ma'an	110	79	129	77	395	39	20
		Total	**1,420**	**949**	**1,479**	**933**	**4,781**	**466**	**240**
			2,369		**2,412**				

Table 4 summarizes the data collected during the survey across special schools.

Table 4. Descriptive Statistics for Final Sample of Special Schools

School Type	Students					Teachers	Schools
	G2		G3		Total		
	Female	Male	Female	Male			
War Child	108	94	130	68	400	37	20
Syrian Only	110	83	125	79	397	40	20
Refugee Camp	117	86	115	82	400	40	20
	335	263	370	229	1,197	117	60
	598		599				

2.3 Weighting

To make the sample representative of the national population, weights were calculated as the inverse of the selection probability for each student. Three stages of weighting were applied (stratum, school, and student) so that the sample of student scores would be representative of student performance at the national level. All scores reported for this study were calculated using the student weights as noted.

One of the key advantages of weighting is that it accounts for disproportionate sampling.

2.4 Equating Procedures

The purpose of test equating is to calculate comparable scores on different forms of a test (in the case of this study, the 2014, 2017, and 2019 assessments). Equating is done to ensure that differences in scores are the result of differences in ability and not differences in test difficulty. Apart from the ORF calculated based on the oral reading passage (see the earlier discussion), there was no need to perform additional equating for these two assessment versions. Because the endline oral reading passage was designed to have an identical level of difficulty as the midline passage, the same equating formula was used in both 2017 and 2019 to equate back to the 2014 (baseline) passage.

2.5 Reliability Estimates

Both the EGRA and EGMA were tested for reliability to ensure that the assessment instruments were measuring their intended constructs. Cronbach's alpha values for both EGRA and EGMA indicated that the instruments showed good internal consistency on average ($\alpha \geq 0.85$ for EGRA and EGMA scales at all three time points). Overall, these reliability measures provide evidence that each assessment was measuring a single underlying construct: early grade reading ability for EGRA and early grade mathematics ability for EGMA.

3 Findings

Overall, RAMP successfully produced significant improvements in G2 and G3 student reading and mathematics performance from baseline to endline, across foundational and higher-order skills. The impacts were typically larger for G2 students. This finding is not surprising given the cohort-designed rollout of the program, which led to nearly one-third of G3 classrooms only implementing the RAMP methodology during the final year of the program. However, G3 students from C1 (i.e., the students and teachers with the greatest exposure to RAMP) exhibited some of the largest gains overall.

Throughout this section, wherever possible, comparisons are made among baseline, midline, and endline results. However, there are few important considerations to keep in mind:

- The 2014 data only allow for reliable reporting of point estimates at the Kingdom, grade, and gender levels. No further disaggregation is possible. As a result, no baseline comparisons can be made for governorates or cohorts.
- Similarly, because the 2014 data were the product of data collection that occurred prior to the start of the program, RAMP questionnaires are only available at midline and endline. Therefore, student, teacher, and school characteristics are primarily compared between midline and endline only.
- Because of the staggered implementation of RAMP (i.e., the cohort design), all schools were exposed to RAMP methodologies by the time of the endline data collection, but the length of their exposure varied. For example, G3 classes in C3 had only been exposed to RAMP for less than two semesters by the time of the endline survey. Conversely, G3 students in C1 were exposed to RAMP longer than any other group (i.e., 3 full years). Results by cohort are further explored in section 3.4.2.

3.1 EGRA

3.1.1. EGRA Results by Year

The results displayed in ***Table 5*** show positive gains since baseline for nearly all subtasks (with statistical significance provided for comparisons from baseline to endline). The gains are particularly large (and statistically significant) for all of the most basic reading skills except for listening comprehension (i.e., letter sounds, syllable sounds, and invented words). This is encouraging evidence for RAMP, seeing as these are all skills that are foundational for reading fluency and reading comprehension. Despite the overall positive trends, however, some of the gains did lessen between midline and endline. Such reduced gains may be a reflection of working at full scale (across KG2-G3 in all schools) after rolling out across all three cohorts. Overall, there was no real change in the ORF of students in Jordan during the course of the program, but their reading comprehension skills did improve. This differential stems in part from the difficulty in getting teachers to focus on ORF in the classroom, although they did emphasize reading comprehension. Accordingly, the impressively high full comprehension rate—nearly 61%—on the silent reading task indicates that although

students are struggling to read orally, those who are reading are comprehending what they read.

Table 5. Overall EGRA Results by Year (G2 and G3 Students)

Subtask	Measure	Baseline G2 and G3	Midline G2 and G3	Endline G2 and G3
Letter sound	fluency (correct letters per min.)	37 [±3.9]	47.9 [±2.0]	50.1* [±1.9]
	% correct of items attempted	64.9% [±5.4]	76.1% [±2.2]	79.1%* [±2.1]
Syllable sound	fluency (correct syllables per min.)	25.2 [±2.0]	31.7 [±1.7]	33.0* [±1.7]
	% correct of items attempted	63.8% [±3.7]	72.5% [±2.4]	71%* [±2.3]
Invented words	fluency (correct words per min.)	8.9 [±.9]	12.2 [±.7]	15.1* [±.8]
	% correct of items attempted	39.4% [±3.5]	49.3% [±2.3]	57%* [±2.1]
Listening comprehension	% correct	67.2%* [±2.2]	60.6% [±2.0]	58% [±2.1]
Oral reading	ORF	26.6 [±2.2]	25.9 [±1.4]	27.1 [±1.7]
	% correct of items attempted	63.7% [±4.4]	60.4% [±2.6]	60.1% [±2.7]
Reading comprehension	% correct of items attempted	45.3% [±4.4]	49.8% [±2.8]	49.3% [±2.9]
	% correct	33.8% [±3.5]	36.7% [±2.4]	38.6%* [±2.9]
	% of students with 80% comp.	17.9% [±3.5]	21.3% [±2.4]	23.6%* [±3.0]
	% of students with 80% comp on silent reading			60.5% [±3.8]

* p<.05 (baseline vs. endline)

3.1.2. EGRA Results by Grade

The trends are nearly identical between G2 and G3. In both cases, there were significant gains in letters, syllables, and invented words from baseline to endline but no gains in ORF. The main difference between the two grades is that while they both showed positive gains in reading comprehension, only the G2 gains were statistically significant, nearly doubling from 8% at baseline to 14% at endline. Once again, the silent reading comprehension measures tell an important story about the comprehension levels of students when not tied explicitly to ORF scores, with nearly half of G2 students and three-quarters of G3 students showing proficiency in the silent comprehension task.

These results are summarized in ***Table 6*** for G2 and ***Table 7*** for G3. For the sake of brevity, not all results are reflected here.

Table 6. EGRA Results for G2 at Baseline, Midline, and Endline

Subtask	Measure	Baseline G2	Midline G2	Endline G2
Letter sound	fluency (correct letters per min.)	38.2 [±4.3]	47.3 [±2]	48.0* [±2.4]
Syllable sound	fluency (correct syllables per min.)	22.1 [±2.4]	29.8 [±2.2]	29.8* [±1.9]
Invented words	fluency (correct words per min.)	7.1 [±.9]	10.8 [±.9]	13.3* [±.9]
	% correct of items attempted	34.5% [±4.3]	48.2% [±3.3]	55.3%* [±2.8]
Oral reading	ORF	19.1 [±2.2]	20.7 [±1.5]	20.2 [±1.5]
	% correct of items attempted	56.7% [±6.7]	56.2% [±3.6]	53.0% [±3.1]
Reading comprehension	% of students with 80% comp.	7.9% [±3.3]	11.4% [±2.2]	13.5%* [±2.9]
	% of students with 80% comp on silent reading			46.6% [±4.4]

* p<.05 (baseline vs. endline)

Table 7. EGRA Results for G3 at Baseline, Midline, and Endline

Subtask	Measure	Baseline G3	Midline G3	Endline G3
Letter sound	fluency (correct letters per min.)	35.7 [±4.1]	48.6 [±3.1]	52.1* [±2.2]
Syllable sound	fluency (correct syllables per min.)	28.7 [±2.1]	33.8 [±2.3]	36.1* [±2]
Invented words	fluency (correct words per min.)	10.9 [±1.0]	13.7 [±.9]	16.9* [±.0]
	% correct of items attempted	44.9% [±3.6]	50.4% [±2.6]	58.6%* [±2.4]
Oral reading	ORF	35.0 [±2.8]	31.2 [±2.1]	33.6 [±2.3]
	% correct of items attempted	71.6%* [±3.2]	64.7% [±3.3]	66.9% [±3.0]
Reading comprehension	% of students with 80% comp.	29.0% [±5.4]	31.5% [±3.7]	33.3% [±4.3]
	% of students with 80% comp on silent reading			73.7% [±4.4]

* p<.05 (baseline vs. endline)

3.2 EGMA

3.2.1. EGMA Results by Year

Statistically significant, positive gains were found on all mathematics subtasks from baseline to endline. While the gains were relatively small for the more foundational skills (i.e., because the baseline scores were quite high), the gains for conceptual skills, which is the focus of RAMP, were larger. The gains for both addition and subtraction L2 and word problems were statistically significant from midline to endline, showing improved gains over the last 2 years of the program.

Table 8 summarizes the overall data for the EGMA subtasks from baseline to endline for G2 and G3 combined. Statistical significance is indicated for comparisons from baseline to endline.

Table 8. Overall EGMA Results by Year (G2 and G3 Students)

Subtask	Measure	Baseline G2 and G3	Midline G2 and G3	Endline G2 and G3
Number Identification	% correct of items attempted	88.1% [±1.9]	92.6% [±.8]	92.2%* [±1]
Quantity Comparison	% correct	78.9% [±2.3]	83.7% [±1.4]	85.1%* [±1.4]
Addition L1	fluency (correct items per min.)	11.9 [±.5]	12.5 [±.3]	13.1* [±.4]
Subtraction L1	fluency (correct items per min.)	9.5 [±.5]	10 [±.3]	10.4* [±.3]
Addition and Subtraction L2	% correct	41.9% [±2.8]	45.9% [±2.2]	52.1%* [±2.5]
Missing Number	% correct	60.3% [±2.9]	64.5% [±1.9]	64.1%* [±2.2]
Word Problems	% correct	57.6% [±2.7]	59.9% [±2.0]	63.6%* [±2.4]

* p<.05 (baseline vs. endline)

3.2.2. EGMA Results by Grade

Because "doing mathematics with understanding" is RAMP's goal, only measures that focus on mathematical conceptual understanding are included in this disaggregation (complete grade-level EGMA results can be found in **Annex 11**). Overall, positive results were found for conceptual mathematics skills in G2 and G3, as shown in ***Table 9*** and ***Table 10***. The results are particularly strong for addition and subtraction L2, with statistically significant positive gains from baseline to endline AND from midline to endline for both grades. While G3 students showed an improvement in missing number from baseline to endline, there was actually a small reduction in the endline results for G2 students since midline; this change was not statistically significant.

Table 9. G2 EGMA Results at Baseline, Midline, and Endline

Subtask	Measure	Baseline G2	Midline G2	Endline G2
Addition and Subtraction L2	% correct	36.8% [±3.1]	41.6% [±3.1]	47.3%* [±3.2]
Missing Number	% correct	54.3% [±3.6]	60.2% [±2.9]	56.7% [±2.6]

Table 10. G3 EGMA Results at Baseline, Midline, and Endline

Subtask	Measure	Baseline G3	Midline G3	Endline G3
Addition and Subtraction L2	% correct	47.5% [±3.4]	50.3% [±2.7]	56.7%* [±2.7]
Missing Number	% correct	66.9% [±3.1]	68.9% [±2.3]	71.2%* [±2.3]

3.3 Performance on RAMP Indicators

A key purpose of the RAMP endline (2019) study was to establish whether RAMP was having an impact on early grade reading and mathematics performance across a range of preset indicators. Accordingly, this report aims to establish whether there has been a shift in the proportion of students achieving each of the RAMP goal-level indicators as published in the RAMP AMELP.

These results show that RAMP was able to meet two endline targets (based on 95% confidence intervals)—GL_02 and GL_07—both of which show impressive improvements in the mathematics proficiency of students in the early grades (***Table 11***). After relatively flat results between baseline and midline for mathematics, RAMP provided teachers with booster training focused on improving their conceptual understanding of mathematics and providing students with opportunities to solve problems and discuss different solutions. Additionally, all students were provided with math workbooks beginning in the 2017/2018 school year. While not all gains can be attributed to these two measures, it does appear that RAMP efforts did have an impact on improved mathematics performance across indicators.

Table 11. Summary of Performance on RAMP Indicators by Year

Indicator Number	Indicator[4]	Baseline Value	Midline Value	Endline Target	Endline Value
GL_01	***Proportion of learners who, by the end of two grades of primary schooling,***	7.9%	11.4%	22%	13.5%*

[4] For the purposes of these indicators and throughout the report, *reading with comprehension* and *doing mathematics with understanding* are defined as follows:

- To *read with comprehension*, the student must score at least 80% on the EGRA reading comprehension subtask.

Indicator Number	Indicator[4]	Baseline Value	Midline Value	Endline Target	Endline Value
	demonstrate reading fluency and comprehension of grade-level text.				
GL_02	***Proportion of students who, by the end of two grades of primary schooling, demonstrate that they can do grade-level mathematics with understanding.***	9.8%	11.7%	17%	18.7%*
GL_03	***Proportion of students who, by the end of three grades of primary schooling, demonstrate that they can read and understand grade-level text.***	29.0%	31.5%	50%	33.3%
GL_04	***Proportion of students who, by the end of three grades of primary schooling, demonstrate that they can do grade-level mathematics with understanding.***	19.5%	21.2%	34%	29.2%*
GL_05	***Proportion of students obtaining zero scores on ORF at the end of G2.***	12.5%*	12.6%	6%	20.7%
GL_06	***Proportion of G2 and G3 students who demonstrate that they can read grade-level text with comprehension.***	17.9%	21.3%	35%	23.6%*
GL_07	***Proportion of G2 and G3 students who demonstrate that they can do grade-level mathematics with understanding.***	14.4%	16.0%	25%	24.1%*

* p<.05 (baseline vs. endline)

While the remaining five targets were not met, statistically significant improvement occurred from baseline to endline for three targets: GL_01 (G2 reading proficiency), GL_04 (G3 math proficiency), and GL_06 (G2/3 reading proficiency). Although failing to reach targets is disappointing, the progress made on these indicators should not be overlooked or undervalued. Shifting entire distributions of student performance is very difficult and takes a long time. When few students are reaching a standard (and most students are far from the standard) at baseline, it is difficult to produce large gains in just a few years.[5] Accordingly, these small gains must be examined in relative terms. Using GL_01 as an example, the 5.6 percentage point increase in G2 students reading with fluency and comprehension actually represents a 71% increase in standard-meeting performance from baseline to endline.

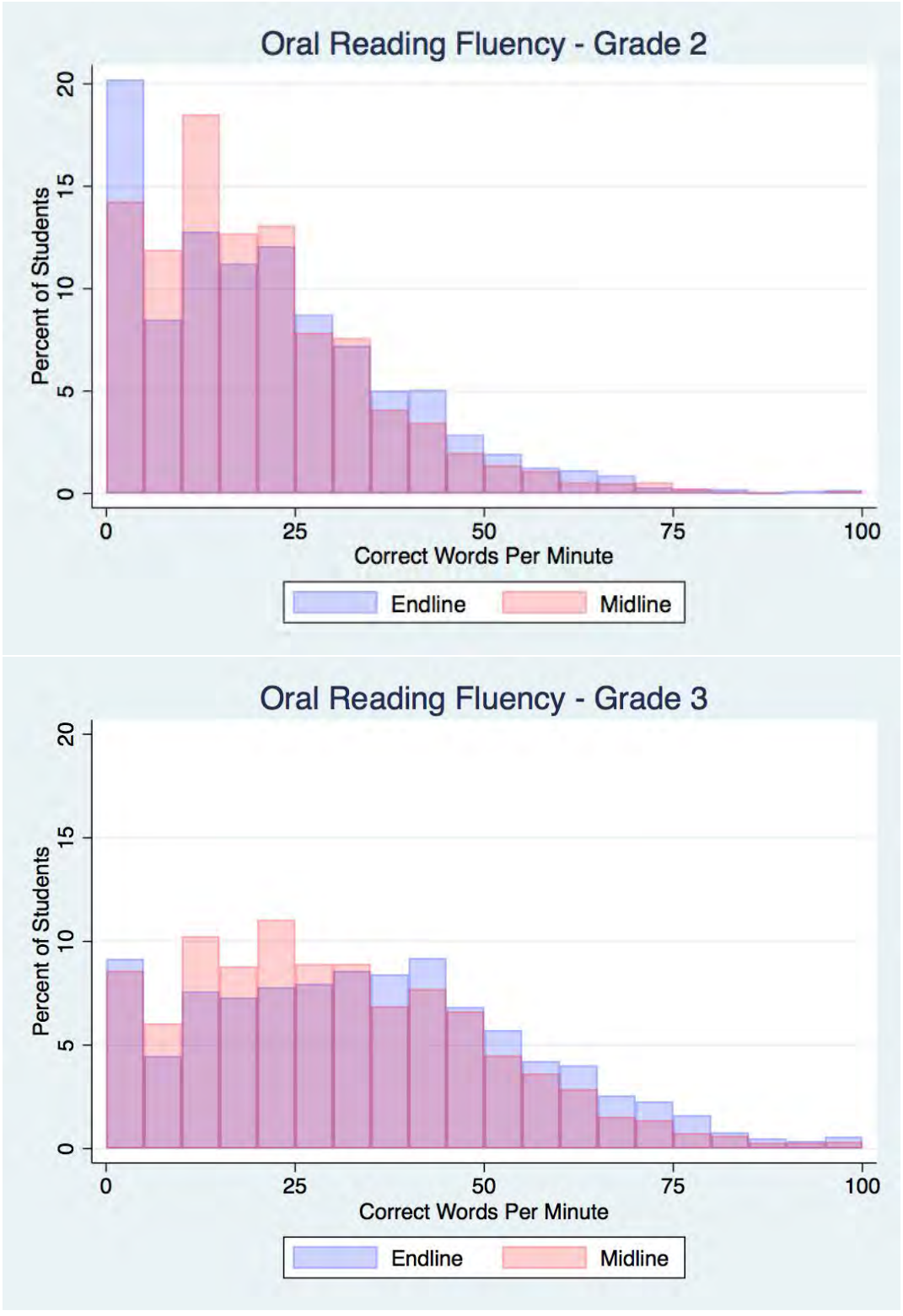
Small gains in indicator GL_03 (G3 reading comprehension) were achieved, but they were not statistically significant. The confounding finding from ***Table 11*** is the increase in oral reading zero scores for students at the end of G2 (i.e., GL_05). It is interesting to note that this increase in zero scores occurs in the same population of G2 students among whom overall reading fluency and comprehension (GL_01) increased significantly.

- To *do mathematics with understanding*, the student must score at least 80% on the EMGA addition and subtraction (L2) subtask and at least 70% on the EGMA missing number subtask.

[5] Stern, J. M. B., & Piper, B. (2019). *Resetting targets: Examining large effect sizes and disappointing benchmark progress.* Research Triangle Park, NC: RTI Press. https://rtipress.scholasticahq.com/article/8870.pdf

This somewhat counter-intuitive finding is explained in part by ***Figure 1***, which displays the distributions of ORF scores on the x-axis and the percent of students reaching those scores on the y-axis. Where the red bars are taller, the proportion of students was higher at midline; where the blue bars are taller, the proportion of students was higher at endline. This figure shows that there was an increase in zero scores for G2 and G3. However, while a larger proportion of students scored between 1 word and approximately 25–30 words for both grades at midline, the proportion of readers reading more than 25–30 words was higher at endline. In other words, while the weakest students are not improving (and this group is apparently growing slightly), the mid-range and higher-performing students are continuing to improve. This trend could be a sign that the weakest students are still not being fully involved in the learning process, but it may also be a result of the influx of refugee students, the reduction of teaching hours because of double shifting, and the recent reduction in the number of Arabic language class lessons taught per week, which limits the use of reading resources used in the classroom, as all time is focused on completing the textbook. All of these factors are likely to most adversely impact struggling students.

Figure 1. Distribution of ORF Scores by Grade



3.4 Performance by Key Indicators and Subgroups

To better understand the performance of students under RAMP, it is necessary to examine the results disaggregated across subgroups. With 15 subtasks and a wide range of student, teacher, and school characteristics, it was essential to narrow down the focus of this section to the most critical performance indicators and subgroups. Ultimately, four performance indicators and six subgroups were selected for closer examination as these indicators and subgroups were determined to be central to conversations about RAMP among the program team, USAID, and the MoE. The four selected indicators represent high-level reading and mathematics skills that are goals for the Jordanian primary education system. All of the indicators will be presented separately for G2 and G3 and are as follows:

- **Reading Proficiency:** The percent of students who are able to correctly answer at least 80% of the reading comprehension questions based on the oral reading passage.
- **ORF Benchmark:** The percent of students meeting or exceeding the ORF benchmark of 46 correct words per minute.
- **Silent Reading:** The percent of students who are able to correctly answer at least 80% of the reading comprehension questions on the silent reading subtask (endline only).
- **Math Proficiency:** The percent of students who are able to score at least 80% on addition and subtraction L2 and at least 70% on missing numbers.

3.4.1. Gender

At baseline (and midline), the reading performance of female students was higher than that of male students, whereas male students performed slightly better in mathematics. ***Table 12*** shows that these gaps were smaller at endline for G2 and that the difference was no longer statistically significant for reading or math proficiency. However, a large difference was found in silent reading comprehension, with girls significantly outperforming boys (53% vs. 40%).

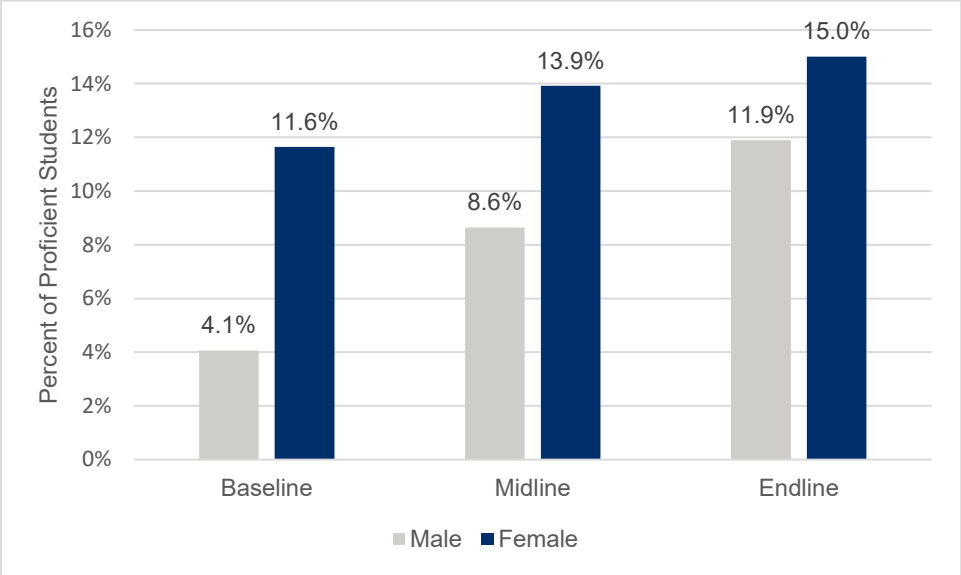
Table 12. Summary of G2 Performance by Gender and Year

Indicator	Baseline		Endline	
	Male	Female	Male	Female
Reading Proficiency	4.1% [±2.6]	11.6%* [±5.8]	11.9% [±3.6]	15% [±3.7]
ORF Benchmark	4.9% [±3.3]	11.9%* [±5.2]	5.3% [±2.3]	9.4%* [±2.7]
Silent Reading			39.9% [±6.1]	52.8%* [±4.9]
Math Proficiency	11.8% [±6.8]	7.8% [±3.6]	18.1% [±4.7]	19.3% [±4.0]

* p<.05 (male vs. female)

The overall trend for G2 is displayed in ***Figure 2*** and clearly shows the shrinking of the reading proficiency gap between male and female students from baseline (7%) to midline (5%) to endline (3%) (where the light blue bars represent male students, and the dark blue bars represent female students).

Figure 2. G2 Reading Proficiency by Gender and Year



Similar results were found in G3, as shown in ***Table 13***. A significant difference in silent reading scores remains, but the gap is much smaller than in G2. The main difference between the two grades is that the difference in math proficiency actually grew from baseline to endline, with male students significantly outperforming female students at endline (34% vs. 25%).

Table 13. Summary of G3 Performance by Gender and Year

Indicator	Baseline		Endline	
	Male	Female	Male	Female
Reading Proficiency	25.1% [±9.1]	32.2% [±6.7]	32.6% [±5.5]	33.8% [±4.6]
ORF Benchmark	20.8% [±6.5]	35.5%* [±7.6]	27.3% [±4.7]	28.6% [±4.8]
Silent Reading			70.2% [±5.9]	76.8%* [±4.8]
Math Proficiency	19.1% [±5.8]	19.9% [±4.7]	33.8%* [±4.9]	25.1% [±3.7]

* p<.05 (male vs. female)

3.4.2. Cohort

Because of the staggered rollout of RAMP trainings, students across the country had exposure to RAMP for different periods of time. At one extreme, G3 classes in C3 were only exposed to the program for one school year (with teacher training occurring in August 2018 and coaching beginning in September 2018). At the other end of the spectrum, G3 students in C1 have been exposed to RAMP methodologies since the beginning of their primary schooling career. In addition to RAMP exposure, there are differences in school-level characteristics across cohorts. For example, according to endline estimates, C3 schools are significantly smaller, averaging 127 students per school, as compared to 188 and 189 students for C1 and C2 schools, respectively. Additionally, C3 had the largest proportion of all-day schools (82%); in comparison, 76% of C1 schools and only 63% of C2 schools were all day. All these factors are likely to impact performance.

The G2 midline and endline results are presented by cohort in ***Table 14*** (note: cohort-level data are only available at midline and endline). These results show that C3 students performed significantly below C1 and C2 students at midline for all indicators. By endline, the introduction of RAMP appears to have narrowed the gap slightly in reading, with C3 students showing the largest gains. In mathematics, however, C3 students remained behind, with the largest gains being seen in C2. One hypothesis for this finding is that the mathematics booster program was introduced shortly after the C1 training ended but before the C3 training began. Therefore, it is possible that C2 benefitted most from the booster program.

Table 14. Summary of G2 Performance by Cohort and Year

Indicator	Midline			Endline		
	C1	C2	C3	C1	C2	C3
Reading Proficiency	12.1% [±3.5]	13.1% [±3.5]	6.3% [±2.8]	14.5% [±3.4]	14.7% [±5.0]	9.5% [±3.4]
ORF Benchmark	7.2% [±3.0]	7.1% [±2.6]	3.1% [±1.6]	7.1% [±2.4]	7.8% [±3.0]	6.7% [±3.6]
Silent Reading				46.5% [±5.0]	49.0% [±7.5]	40.7% [±5.9]
Math Proficiency	13.7% [±4.8]	11.6% [±4.0]	8.9% [±3.6]	17.7% [±4.0]	21.8% [±6.6]	12.3% [±2.9]

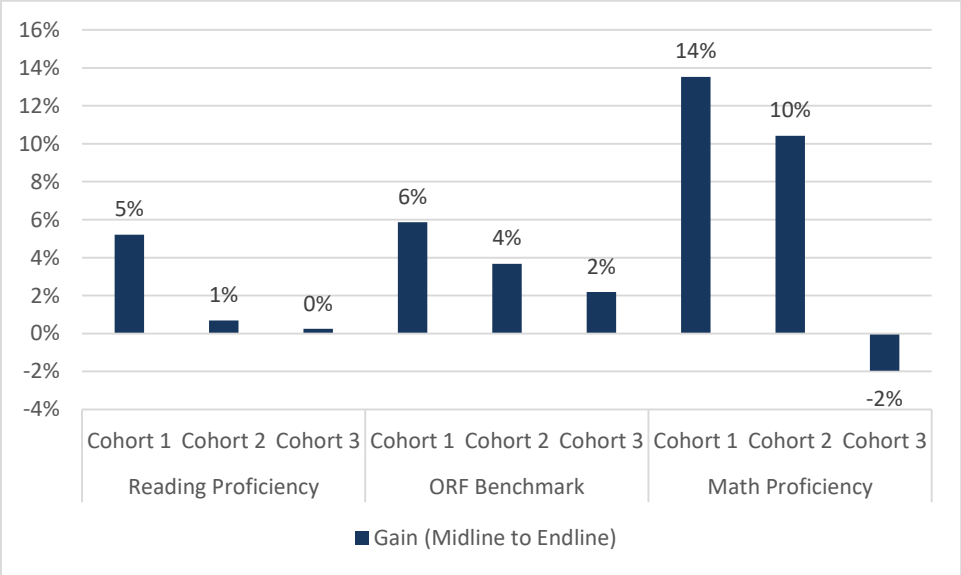
The results for C3 G3 students were quite different than those for G2, as shown in ***Table 15***. In G3, C3 students showed little to no improvement across the indicators (and scored well below their C1 and C2 counterparts in silent reading). This is likely a result of C3 teachers only being introduced to the RAMP methodology at the beginning of the school year and, therefore, having little time to implement the approach and effect change. In contrast, the results for G3 students in C1 were the strongest of any grade/cohort combination.

Table 15. Summary of G3 Performance by Cohort and Year

Indicator	Midline			Endline		
	C1	C2	C3	C1	C2	C3
Reading Proficiency	30.9% [±5.6]	35.4% [±6.3]	22.6% [±5.6]	36.1% [±6.8]	36.1% [±7.3]	22.8% [±4.0]
ORF Benchmark	24.3% [±4.5]	26.7% [±5.5]	17.4% [±4.4]	30.1% [±5.8]	30.3% [±6.5]	19.6% [±3.9]
Silent Reading				72.7% [±5.8]	77.8% [±7.7]	65.0% [±4.9]
Math Proficiency	18.9% [±5.8]	22.4% [±4.8]	18.2% [±3.8]	32.5% [±5.3]	32.8% [±5.6]	16.2% [±3.6]

The overall gains for G3 performance (by indicator and cohort) are displayed in ***Figure 3***. Clearly, G3 students in C1 showed the largest gains in reading proficiency, ORF benchmark, and math proficiency. Because G3 students and teachers in C1 had the longest exposure to RAMP methodologies, it is not surprising that they showed the largest gains. This finding indicates that longer exposure to the program may be essential for translating into larger learning gains, once both students and teachers are familiar with the teaching and learning process.

Figure 3. G3 Performance by Indicator, Cohort, and Year

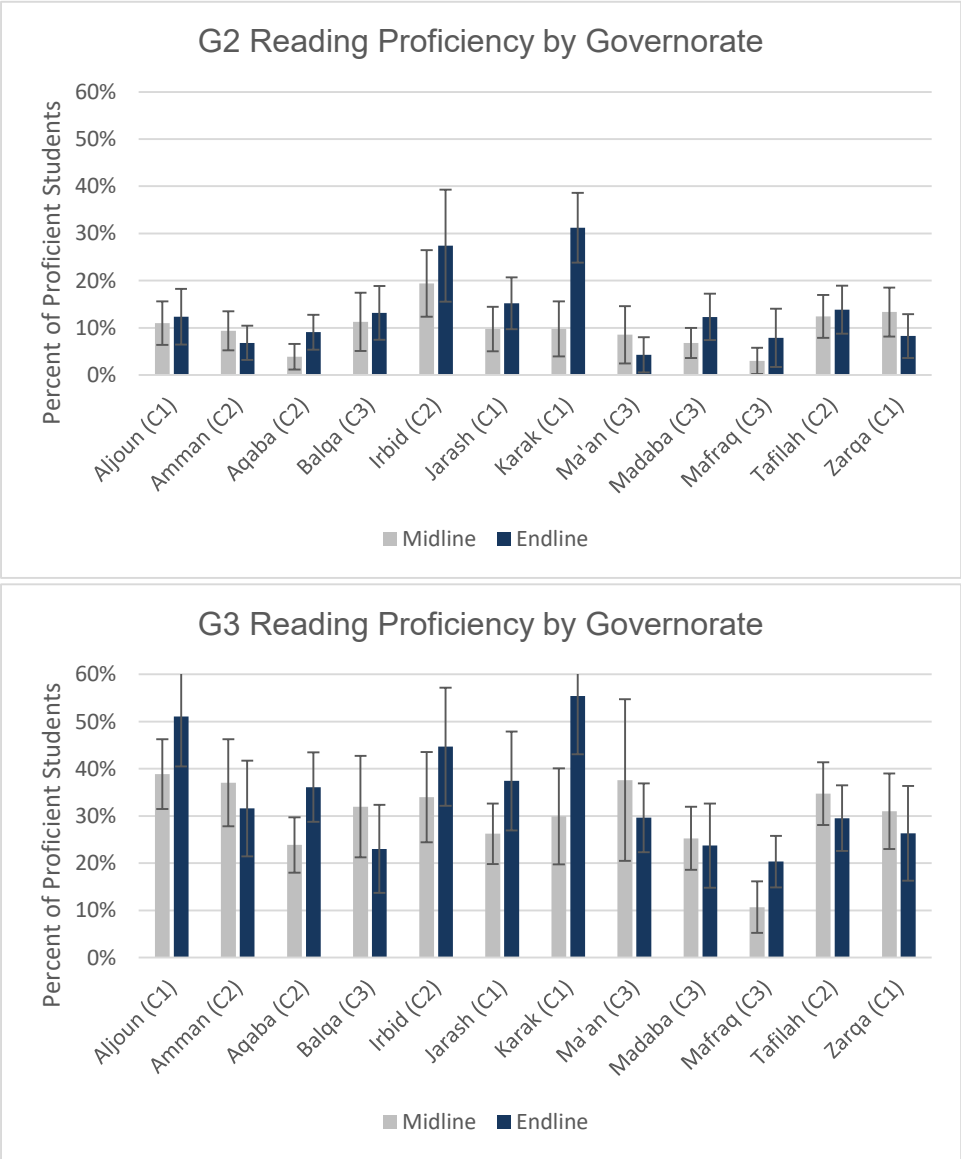


3.4.3. Governorate

Because there are 12 governorates in Jordan, this section is organized differently than other subgroup analyses, with the goal of clearly displaying trends without requiring cumbersome tables or an overabundance of figures. Accordingly, the results in this section are presented first for reading proficiency, followed by math proficiency, and then some additional results on silent reading and oral reading zero scores.

Regarding reading proficiency, there is a large amount of variation across governorates. As shown in ***Figure 4***, several governorates increased their proportion of students reading with proficiency from midline to endline, while others' results remained the same (note: none of the reductions were statistically significant). Only Aqaba and Karak showed statistically significant gains in both G2 and G3, with Karak achieving the largest gains overall. Mafraq produced significant gains in G3 only (from 11% to 20%). Madaba showed marginally significant gains in G2, while Aljoun and Jarash had marginally significant gains in G3.

Figure 4. Reading Proficiency by Governorate, Grade, and Year



Black bars represent 95% confidence intervals.

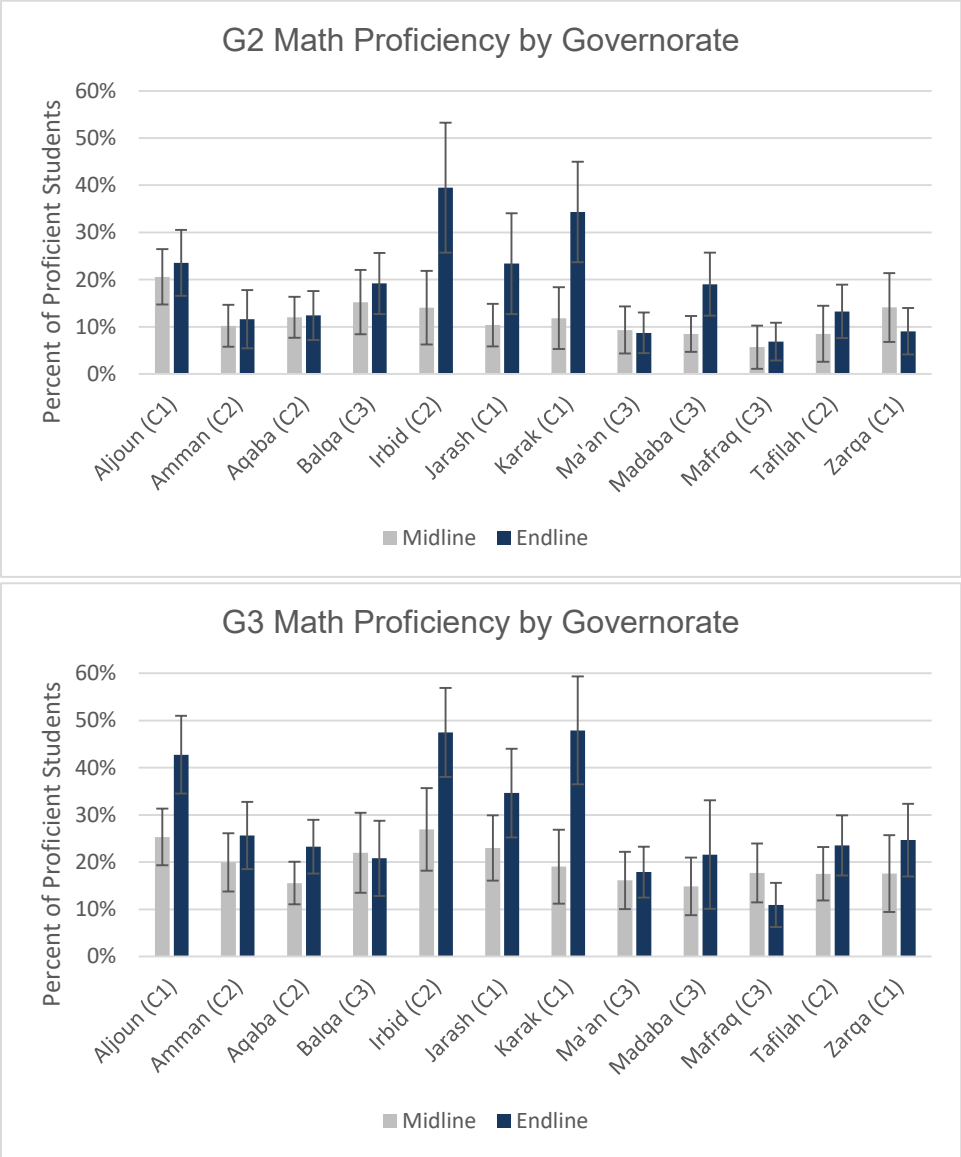
Similar variation is seen for the math proficiency results (***Figure 5***), with similar governorates leading the performance gains. Irbid, Jarash, and Karak all showed statistically significant improvement in G2 and G3. Aljoun and Aqaba also produced

significant math proficiency gains in G3, while Madaba did so in G2. Conversely, Mafraq showed a marginally significant decline from midline to endline in G3.

Overall, across both reading and math proficiency, Karak and Aqaba produced the most consistently significant gains, and Jarash, Madaba, Irbid, and Aljoun also showed impressive progress.

Karak, Aljoun, and Jarash are all C1 governorates; thus, their performance may be a result of their longer exposure to RAMP. The results in Aqaba, in contrast, are attributed to the focused effort there. Based on poor LQAS results in Aqaba and Madaba, RAMP added two coaches to these governorates; these coaches were charged with spending 2 days in each school to provide classroom support and lead community of practice meetings. Moving forward, this targeted approach may be advisable for other low-performing governorates.

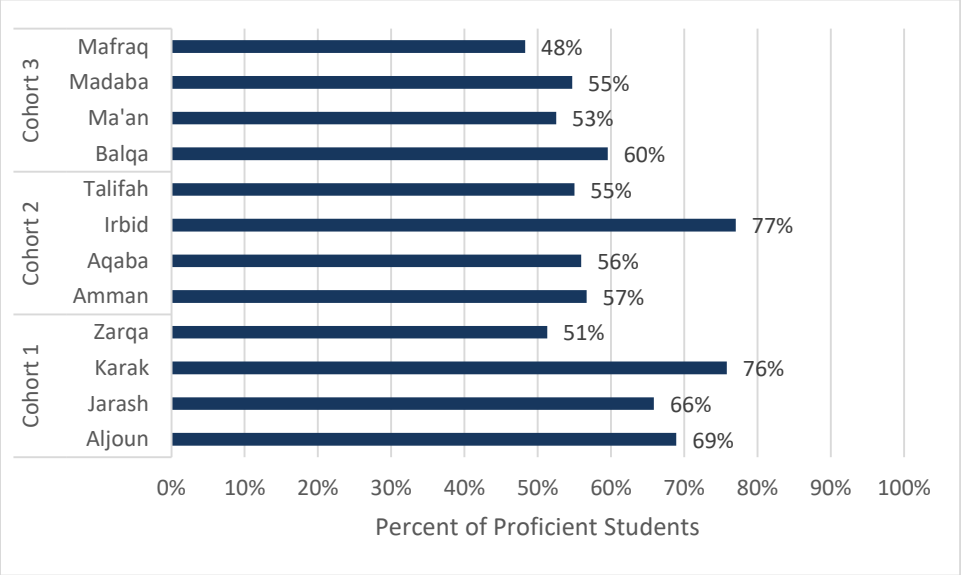
Figure 5. Math Proficiency by Governorate, Grade, and Year



Black bars represent 95% confidence intervals.

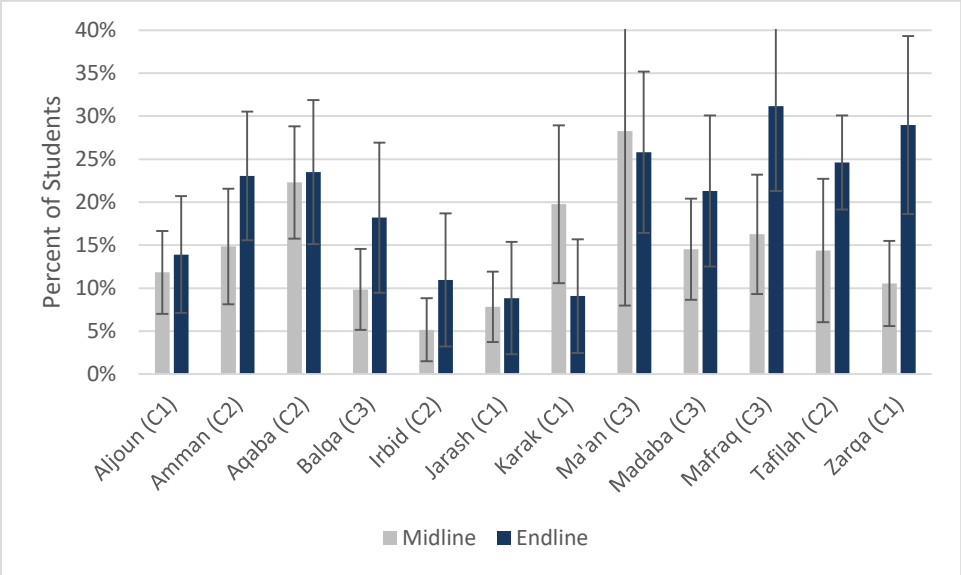
One of the overall goals of RAMP and the Jordanian education system is to work toward 55% of G2 and G3 students meeting reading comprehension standards. Although that goal has not yet been met according to the traditional EGRA reading comprehension subtask, nationally, the percent of students meeting the reading comprehension benchmark on the silent reading task is 60% at endline. As with all other performance indicators, this value varies by governorate. ***Figure 6*** shows that Aljoun, Irbid, Jarash, and Karak far exceeded the 55% goal, including their 95% confidence intervals. As Irbid is a C2 governorate, it would be interesting to further investigate the factors leading to its relatively high scores. Amman, Aqaba, Balqa, Madaba, and Tafilah just met the measure. However, Zarqa, Mafraq, and Ma'an remain below the 55% mark. This is likely attributable in part to the fact that Zarqa and Mafraq have relatively large proportions of camp schools and that Ma'an has a high number of remote schools and only two supervisors covering kindergarten and the early grades. Additionally, Mafraq and Ma'an are both C3 governorates, and their results may have been impacted by short exposure to RAMP, less desirable placements for teachers, and wide geographical coverage, reducing the amount and quality of support that could be provided efficiently.

Figure 6. Silent Reading Proficiency by Governorate and Grade (Endline Only)



Lastly, given the noted increase in G2 oral reading zero scores nationally, it is important to try to understand if certain governorates may be at the heart of this issue. The governorate-level results for oral reading zero scores in G2 are displayed in ***Figure 7***. Clearly, the majority of governorates showed small changes in their oral reading zero scores from midline to endline, with very few differences being statistically significant. On the positive end, Karak produced a marginally significant decrease in zero scores from 20% to 9%; this result is not surprising because the supervisors can easily cover all schools in this governorate, and Karak has well-established communities of practice. At the other end, the overall increase in zero scores appears to be strongly driven by poor performance in Tafilah, Mafraq, and Zarqa, which showed statistically significant increases of 10%, 15%, and 18%, respectively.

Figure 7. G2 ORF Zero Scores by Governorate and Year



Black bars represent 95% confidence intervals.

All of these governorate-level results (reading proficiency gains, math proficiency gains, silent reading proficiency, and oral reading zero scores) are closely aligned. Karak stands out as the best-performing governorate in terms of gains since midline. Overall, the results indicate that both gain scores and overall performance are currently concentrated in approximately half of the governorates, with a few consistently struggling to keep up. Furthermore, clear, significant performance improvement was achieved in C1 governorates, except for Zarqa. This result is attributed to the following:

- This cohort's long duration of exposure to RAMP
- The existence of Senior Teachers in schools in Jarash and Aljoun
- Small class sizes (except for Zarqa)
- Low proportions of two-shift schools (except for Zarqa).

Details of these latter two points are displayed in ***Table 16***, which shows that Zarqa has significantly higher student:teacher ratios and a greater proportion of two-shift schools than its C1 counterparts. These two factors are likely to negatively impact results, with more students per classroom and less time available for reading lessons.

Table 16. Summary of G2 Performance by Nationality and Year

Governorate	Student:Teacher Ratio	Proportion (%) of Two-Shift Schools
Jarash	21.8	2%
Aljoun	20.7	7%
Karak	21.4	8%

Governorate	Student:Teacher Ratio	Proportion (%) of Two-Shift Schools
Zarqa	30.8	47%

3.4.4. Nationality

The demographics of students in Jordanian schools have changed in recent years, partly because of the increasing numbers of Syrian refugees into the country. Two survey questions used at midline and endline were designed to capture this information. First, head teachers were asked to provide the total number of KG2–G3 students in their school who were Syrian refugees. Additionally, in the student questionnaire, students were asked about where their parents were from. Both measures pointed to increases in the Syrian population in schools, with head teachers noting that the number of Syrian refugee students in KG2–G3 increased, on average, from 17 in 2017 to 25 in 2019. Similarly, in 2017, 11.6% of students said their parents were from Syria, whereas in 2019, this proportion increased to 15.2%. Accordingly, one question worth investigating is whether differences in performance exist between Jordanian and Syrian students.

A comparison of Jordanian and Syrian G2 student performance in regular public schools is presented in ***Table 17*** (student performance data for special schools, such as those in refugee camps, are provided in section 3.4.6). Scores across all three indicators at midline were strikingly similar among Jordanian and Syrian students. Counterintuitively, at endline, Syrian students achieved slightly larger gains than Jordanian students, although none of the differences observed were statistically significant.

Table 17. Summary of G2 Performance by Nationality and Year

Indicator	Midline		Endline	
	Jordanian	Syrian	Jordanian	Syrian
Reading Proficiency	10.7% [±2.5]	10.9% [±6.1]	13.2% [±2.8]	16.4% [±11.1]
ORF Benchmark	6.3% [±1.8]	5.7% [±6.6]	7.1% [±1.8]	9.8% [±5.8]
Silent Reading			46.0% [±4.3]	51.6% [±15.2]
Math Proficiency	11.7% [±3.0]	13.2% [±5.0]	17.1% [±3.3]	30.2% [±16.7]

The trend for G3 is slightly different than that for G2, as shown in ***Table 18***. In this case, Jordanian students outperformed Syrian students at midline (significantly in reading proficiency), but Syrian students produced slightly larger gains from midline to endline, eliminating the gap between the two groups.

Table 18. Summary of G3 Performance by Nationality and Year

Indicator	Midline		Endline	
	Jordanian	Syrian	Jordanian	Syrian
Reading Proficiency	33.3%* [±3.8]	22.8% [±7]	32.9% [±4.9]	33.8% [±8.5]
ORF Benchmark	25.2% [±3.4]	16.8% [±8.9]	28.1% [±4.1]	27.6% [±10.4]
Silent Reading			74.1% [±5.0]	70.7% [±11.7]
Math Proficiency	21.9% [±3.0]	13.8% [±10.4]	28.0% [±3.7]	33.6% [±10.4]

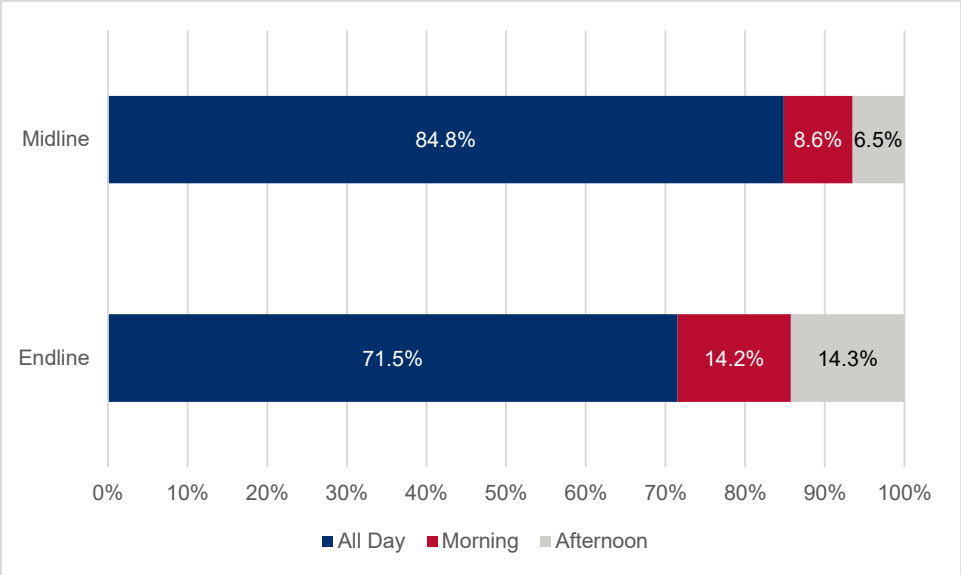
* p<.05 (Jordanian vs. Syrian)

These results provide evidence that while the proportion of Syrian students is increasing in Jordanian schools, the performance of Syrian students does not appear to be driving smaller overall gains (at least not based on students reporting their parents' country of origin, which may or may not reliably capture Syrian refugee status). It is still possible, however, that the increase in Syrian refugees is impacting the system as a whole, resulting in larger class sizes and the need for more dual-shift schools and placing additional burden on the existing teaching force.

3.4.5. School Status (Full Day vs. Two Shift)

Jordan has a primarily single-shift school system, but in recent years, partly because of increased enrollments of Syrian refugee children without increases in the number of classrooms, more schools have adopted the two-shift approach. Specifically, as displayed in ***Figure 8***, the proportion of single-shift schools dropped from 85% at midline (2017) to 72% at endline (2019). Because multi-shift schools have less time for instruction (by design), the increase in multi-shift schools may be adversely affecting average performance in the system.

Figure 8. Percentages of Schools by Type and Year



The performance of G2 students across all three school types (for midline and endline) is displayed in ***Table 19***, which shows that differences in performance across school types are small and inconsistent. While all-day and afternoon-shift schools showed small improvements in reading and math proficiency, morning-shift schools actually showed decreases for both indicators, though the two-shift schools did show small gains in the proportion of students reaching the ORF benchmark.

Table 19. Summary of G2 Performance by School Status and Year

Indicator	All Day		Morning Shift		Afternoon Shift	
	Midline	Endline	Midline	Endline	Midline	Endline
Reading Proficiency	12.1% [±2.7]	14.5% [±3.4]	11.3% [±5.1]	10.1% [±5.9]	7.8% [±6.4]	13.3% [±8.8]
ORF Benchmark	7.2% [±2.1]	7.3% [±1.7]	3.9% [±3.7]	6.6% [±5.4]	4.9% [±3.5]	8.1% [±5.9]
Silent Reading		48.8% [±5.0]		42.6% [±10.3]		43.9% [±12.3]
Math Proficiency	11.8% [±3.1]	19.6% [±3.8]	13.5% [±7.4]	10.2% [±5.7]	8.4% [±3.8]	22.2% [±12.0]

G3 students exhibited greater differences among school types, as shown in ***Table 20***. Similar to G2 there was little difference in performance across indicators at midline. However, all-day schools produced significant increases in the proportion of students achieving the ORF benchmark and math proficiency, with a smaller, non-significant increase in reading proficiency. In contrast, morning shifts showed decreases in both reading indicators and a smaller increase in math proficiency. Interestingly, afternoon shift students' results were more similar to those of all-day students.

Table 20. Summary of G3 Performance by School Status and Year

Indicator	All Day		Morning Shift		Afternoon Shift	
	Midline	Endline	Midline	Endline	Midline	Endline
Reading Proficiency	32.5% [±4.5]	35.7% [±5.6]	27.8% [±9.1]	22.1% [±8.4]	31.6% [±6.5]	35.2% [±8.0]
ORF Benchmark	25.6% [±3.6]	32.1% [±4.6]	22.4% [±8.2]	18.3% [±8.2]	18.2% [±8.4]	24.6% [±8.8]
Silent Reading		73.9% [±6.1]		73.4% [±8.4]		73.5% [±9.5]
Math Proficiency	21.5% [±3.1]	30.2% [±4.4]	16.6% [±10.5]	24.2% [±7.5]	20.4% [±7.3]	30.2% [±8.2]

It is worth noting that these differences in performance by school shift were observed at a time when the total number of students in two-shift schools decreased by 25%, while the average number of students in all-day schools increased by nearly 20%.

3.4.6. School Type (Traditional vs. Special Schools)

During the endline data collection, an additional 60 special schools were purposefully sampled to examine the reading and mathematics performance of students in non-traditional public schools. This additional school sample was made up of three groups: (1) refugee camp schools, (2) Syrian day schools, and (3) War Child Program schools. Because data for the special school population are only available at endline, comparisons must be made with the understanding that baseline-level performance information is not available. Therefore, the results represent current performance status only, and no claims can be made about gains/improvement for these schools.

The results of G2 students in traditional, refugee camp, Syrian day, and War Child Program schools are displayed in ***Table 21***. Unsurprisingly, the lowest-performing students across all four indicators are in the refugee camp schools; the other three school types showed somewhat similar performance, except for the lower reading proficiency mark in War Child Program schools.

Table 21. Summary of G2 Performance by School Type

	Traditional Schools	Refugee Camp Schools	Syrian Day Schools	War Child Program Schools
Reading Proficiency	13.5% [±2.9]	4.1%* [±2.9]	11.4% [±8.2]	8.3%* [±3.4]
ORF Benchmark	7.4% [±1.9]	1.9%* [±2.1]	6.8% [±5.8]	4.6% [±2.5]
Silent Reading	46.6% [±4.4]	30.5%* [±10.7]	50.5% [±9.7]	45.0% [±7.5]

	Traditional Schools	Refugee Camp Schools	Syrian Day Schools	War Child Program Schools
Math Proficiency	18.7% [±3.7]	11.3% [±8.3]	25.8% [±10.2]	14.5% [±4.6]

* p<.05 (relative to traditional schools)

The results for G3 were very similar trend, as seen in ***Table 22***. Refugee camp students once again underperformed relative to their peers, while all other groups produced similar results.

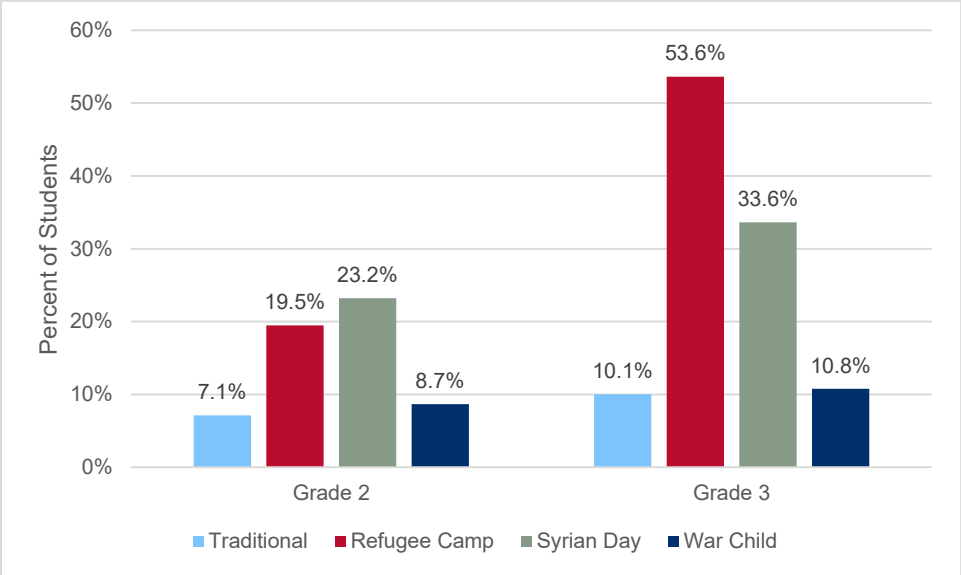
Table 22. Summary of G3 Performance by School Type

	Traditional Schools	Refugee Camp Schools	Syrian Day Schools	War Child Program Schools
Reading Proficiency	33.3% [±4.3]	19.2%* [±11.5]	26.3% [±11.5]	32.7% [±6.9]
ORF Benchmark	28.0% [±3.8]	11.9%* [±5.6]	27.9% [±9.9]	24.8% [±6.1]
Silent Reading	73.7% [±4.4]	59.8%* [±11.6]	77.5% [±7.5]	63.7% [±7.0]
Math Proficiency	29.2% [±3.3]	22.9% [±15.0]	32.0% [±10.4]	37.4%* [±6.8]

* p<.05 (relative to traditional schools)

One factor that likely impacted these results was the age of the students in the different types of schools. Assuming that students in G2 should be 8 years old (on average) and that those in G3 should be 9 years old, we estimated the proportion of students in each grade who were deemed to be 'overage'. This metric is simply relative to the average and is not meant to describe a specific problem, per se. However, as shown in ***Figure 9***, the percentages of overage students in traditional and War Child schools were very similar in G2 and G3, whereas refugee camp and Syrian day schools had significantly larger proportions of overage students, with G3 at refugee camp schools showing the highest proportion of 54%. In the case of refugee camp students, these results are likely driven by children who missed a year (or more) of schooling, while in Syrian day schools, the high proportions may be attributable to students who have repeated a grade (or part of one) during their transition to the new system. This could explain the relative impact of the older students on school performance in these two school types (i.e. older students in Syrian day schools improve performance, while older students in refugee camp schools have a negative impact on performance).

Figure 9. Overage Students by School Type and Grade



3.5 Key Student, Teacher, and School-Level Characteristics

Examining changing demographics and teacher and head teacher perceptions of RAMP is essential for understanding RAMP's successes and continuing challenges.

3.5.1. School Characteristics

The school characteristics displayed in ***Table 23*** provide evidence that the system is becoming increasingly challenging. While the number of two-shift schools has increased to accommodate the growing student population (and increased proportion of Syrian refugees), school sizes have continued to grow among all-day schools, which account for 72% of schools (a 13% decrease since midline). The growing student body in single-shift schools and the increased proportion of morning and afternoon shift schools (the former of which produce lower achievement) are likely factors affecting the overall performance of the system.

Table 23. School Characteristics by Year

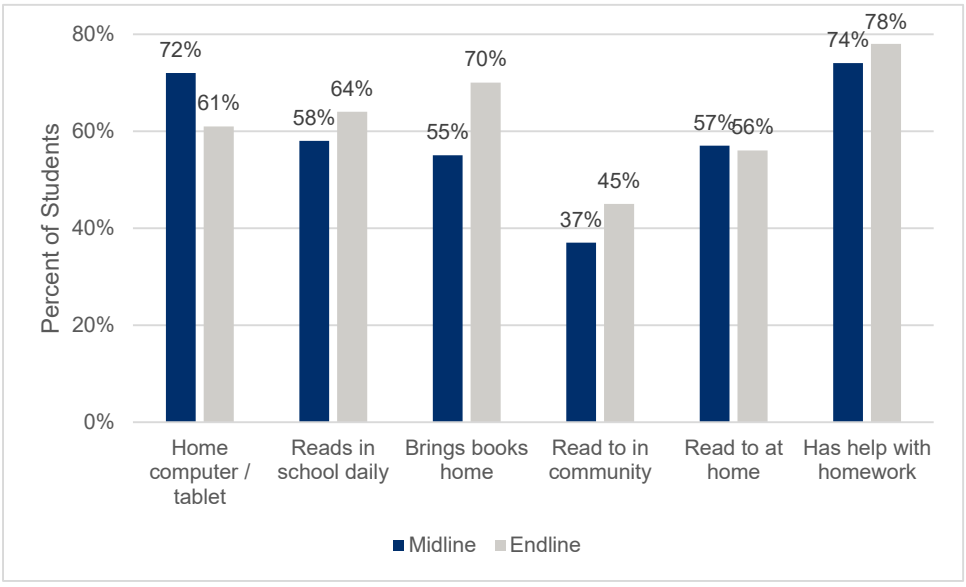
		Midline	Endline
Type of School (%)	All Day	85%	72%
	Two Shift	15%	28%
School Size (Number of Students in KG2–G3)	All Day	119	142
	Two Shift	324	243
Syrian Students (% of Total)		11%	15%

3.5.2. Student Characteristics

Many student-level characteristics remained unchanged from midline to endline. However, several characteristics are worth examining, as displayed in ***Figure 10***. Interestingly, the proportion of students reporting that they had a computer, laptop, or tablet in their home decreased from 72% at midline to 61% at endline, signifying a potential reduction in access to technology at home, though mobile/smartphones are still ubiquitous.

Regarding RAMP initiatives outside the classroom, there was no change in the proportion of students who were read to at home and only a small increase in the proportion of students who receive help with homework (up to 78%). In contrast, larger increases were found in the proportion of students who have time to read books in the classroom or library every day (from 58% to 64%), the proportion of students who bring books home from school (from 55% to 70%), and the proportion of students who meet with others in the community and listen to someone read (from 37% at midline to 45% at endline). All these results indicate increased access to print materials for students, but the percentages remain below the expected levels. In terms of prioritization, regression models for all outcome indicators showed that the most consistently predictive student-level characteristic was bringing books home from the classroom or school library. This variable showed particularly strong associations with increased silent reading comprehension and decreased zero scores on ORF.

Figure 10. Student Characteristics by Year

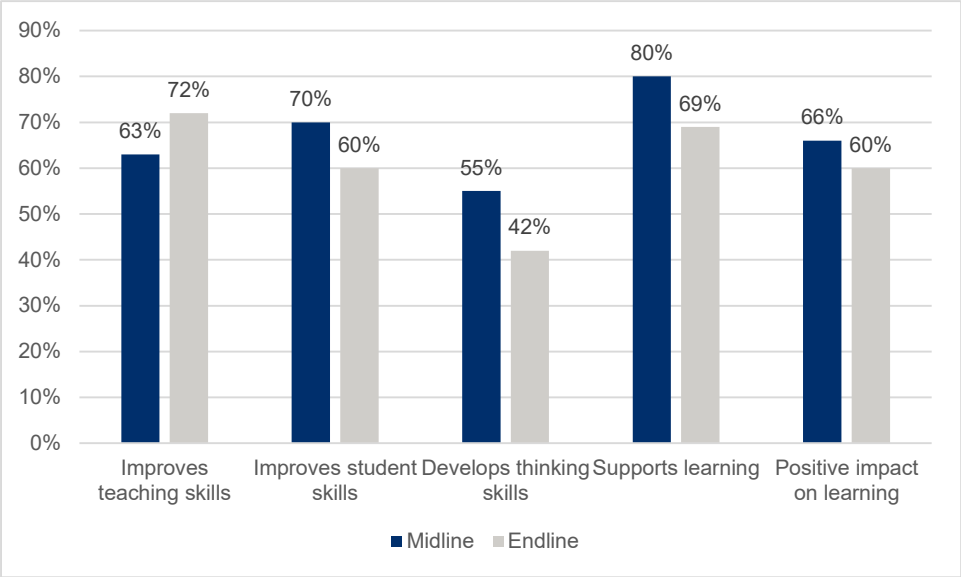


3.5.3. Head Teacher Characteristics

Head teachers generally have positive perceptions of RAMP, with the majority claiming that the program's training, materials, and support are sufficient and that their teachers are implementing the initiative as expected. However, it appears that the

strength of some of these claims may be diminishing over time. As shown in ***Figure 11***, the proportion of head teachers believe that RAMP has improved teaching skills increased, whereas the perception that RAMP positively impacted student performance (i.e., skills, thinking skills, and learning) decreased since midline. As it appears counterintuitive that teaching would be improving but learning would not, this is an areas requiring further investigation. Having head teachers on board, believing in, and supporting a program is ultimately essential for the program's long-term success.

Figure 11. Head Teacher Perceptions of RAMP by Year



3.5.4. Teacher Characteristics

Teachers are at the heart of RAMP. The majority of the G2 and G3 teachers interviewed for this endline survey felt strongly that RAMP was having a positive impact in their classrooms and that they were being provided with the necessary supports to implement the program faithfully. However, there were some changes to the teaching force and implementation issues that may explain some of the obstacles affecting RAMP's performance. Based on the teachers sampled in this study, it was estimated that the proportion of substitute teachers more than doubled from midline to endline (from 7% to 19%), as shown in ***Table 24***. Additionally, the proportion of classrooms that had a single teacher for the duration of the year decreased, and the number of classrooms that had three or more teachers for the year increased by 5%. Lastly, the proportion of teachers with pre-service training in early grade reading or mathematics decreased by more than half from 2017 to 2019. This reduction in the number of consistent, highly trained, permanent teachers in the classroom was undoubtedly a challenge for RAMP and may be one of the reasons for the performance gains that were smaller than anticipated.

Table 24. Teacher Characteristics by Year

	Midline	Endline
Substitute teacher	7%	19%
Only teacher for the class this year	92%	86%
Three or more teachers for the class this year	0%	5%
Pre-service training on early grade reading	29%	13%
Pre-service training on early grade mathematics	25%	14%

While the factors in ***Table 24*** were outside RAMP's control, there were also some implementation challenges that likely impacted student performance. As clearly shown in ***Table 25***, the proportion of teachers who attended all days of all three RAMP teacher trainings increased significantly, but 10% of teachers in G2 and G3 classrooms did not attend any RAMP training at all; as a result, less than two-thirds of teachers received the RAMP certificate. Additionally, although 60% of teachers reported receiving support visits more than once a month at midline, that proportion was reduced to 43% by endline. This reduction is partly attributable to the more targeted approach toward coaching (i.e., focusing on those teachers/classrooms with the greatest need), but additional analyses did not show any clear pattern of increased/decreased coaching visits in terms of student performance. Lastly, only 1 in 10 teachers reported that the RAMP materials (e.g., teacher's guides, student activity books) arrived on time. Clearly, a program cannot be implemented with full fidelity and all performance targets cannot be met, if significant proportions of teachers are not trained on the methodology and/or do not have the necessary materials in their classroom at the start of the school year. These should be major focuses of programming moving forward.

Table 25. RAMP Implementation by Year

	Midline	Endline
Teacher attended all days of all RAMP trainings	31%	67%
Teacher attended none of the trainings	11%	10%
Teacher received RAMP certificate		59%
Supervisor support visits (more than once per month)	60%	43%
RAMP materials arrived on time		10%

Despite these challenges, as shown in ***Figure 12***, more than 80% of teachers felt that RAMP improved math performance, reading performance, and student enthusiasm for learning.

Figure 12. Teacher Perceptions of RAMP's Impact on Students by Year

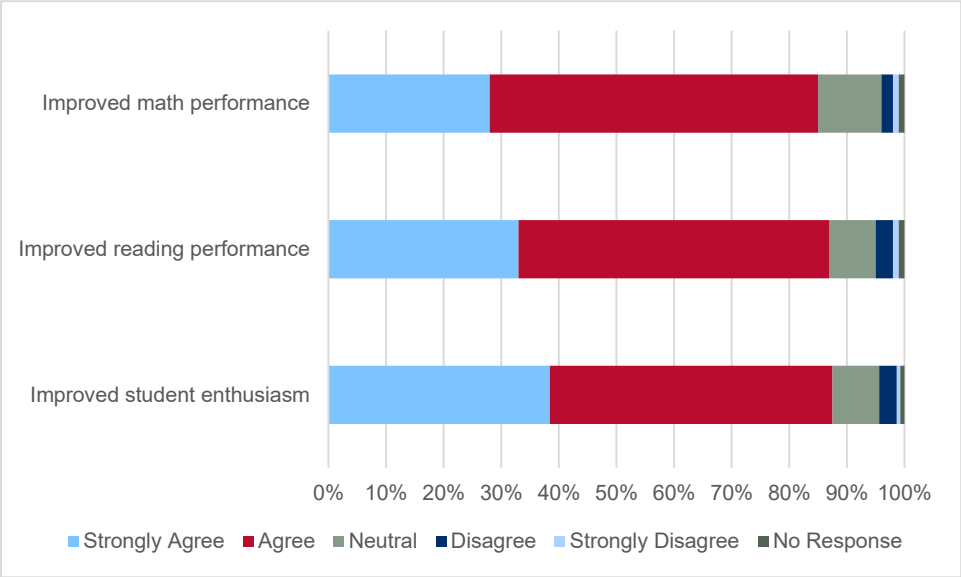
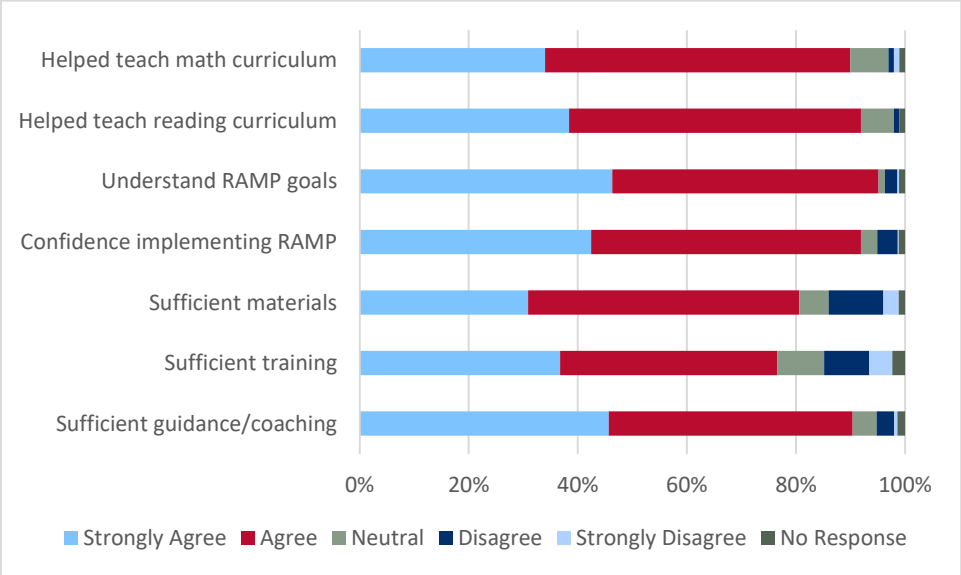


Figure 13 shows that more than 90% of teachers reported that they understood RAMP's goals and that they had confidence in implementing the methodology as intended. However, while nearly all teachers noted receiving sufficient guidance/coaching during the program, smaller proportions of teachers believed that they had been provided with sufficient materials and/or training (supporting the results shown in ***Table 25***). It is clear from the evidence throughout this section that RAMP has been doing a very good job with implementation overall but that the delivery, training, and use of materials requires additional work.

Figure 13. Teacher Perceptions of RAMP by Year



4 Conclusions and Recommendations

The conclusions and recommendations included in this report were jointly developed by the RAMP team and MoE during an analysis and review workshop held in July 2019. The workshop was attended by 26 participants and focused on reviewing the RAMP endline survey results, identifying gains and major accomplishments, discussing challenges, and proposing solutions and next steps for both research and implementation.

4.1 Conclusions

RAMP successfully produced reading and mathematics performance gains for students in both G2 and G3 over the life of the program (2015–2019). The largest reading gains were achieved in the more foundational letter sound, syllable sound, and invented word subtasks. This result is very encouraging as that these foundational skills are the building blocks for reading fluency and comprehension, and improvements are a sign of progress that is expected to lead to continued gains in higher-order skills as teachers become more comfortable with the RAMP methodology and students increase their more basic skills in earlier grades. Interestingly, although reading comprehension scores also improved, there were no changes in ORF scores over the course of the program. The MoE noted that this discrepancy may be related the Arabic language itself and the use of diacritics. Requiring students to read with diacritics, and marking words incorrect based on the absence of the diacritic being read correctly (even if the word is pronounced correctly without the diacritic), are likely to lead to poor fluency rates. This is especially problematic given that students commonly read without diacritics. Therefore, the nature of the ORF task, as administered and scored, is arguably producing lower scores and slower progress than may be the reality. This issue would also explain why fluency rates are stagnating, but students are continuing to improve their comprehension and perform so well on the silent reading comprehension measure (i.e., an estimated 60% of students are reading with comprehension based on the silent reading task).

Nevertheless, progress in reading has been slower than anticipated and below the target levels set by the program. MoE officials noted that one possible explanation for the slow progress was the fact that the number of periods (or lessons) allocated to Arabic language per week was recently reduced from nine to seven. While this change was made to promote "integrated learning," it is likely that untrained (or undertrained) teachers have had some difficulty implementing this approach and, therefore, have simply spent less time on Arabic language and reading.

Regarding USAID performance indicators, RAMP was able to meet two endline targets, both of which relate to mathematics. After a slow start and lack of progress from baseline to midline, RAMP introduced a mathematics booster training to teachers, which focused on their conceptual understanding of early grade mathematics. Additionally, the program provided teachers and students with reinforcement resources (e.g., student mathematics workbooks), which provided additional opportunities to solve problems and discuss solutions to conceptual problems. These initiatives and

RAMP's renewed focus on mathematics led to impressive gains in student performance in the final 2 years of the program.

The performance of male students improved from baseline to endline and narrowed the gender gap considerably. The MoE attributes these changes to increased focus on male students by female teachers' (resulting from psychosocial and capacity building training sessions), increased parental awareness and involvement in their children's learning, and the MoE-led school femininization approach. This latter approach is focused on recruiting exclusively female teachers to teach male students in the early grades based on the belief that female teachers are better at teaching children at the primary level.

The impact of the cohort-designed rollout of RAMP can clearly be seen when examining the cohort- and governorate-level estimates. Generally, G2 classes showed greater progress than G3 classes because G2 classrooms had, on average, longer exposure to the RAMP methodology. Similarly, C1, which was the first cohort to receive training and, therefore, had the longest teacher and student exposure to the program, showed the strongest performance. In contrast, C3 had the lowest results overall, which is attributed to the short duration of RAMP exposure (less than 1 year), the fact that these locations were considered less desirable for teacher placements, and the large geographical areas of the four governorates (Mafraq, Ma'an, Madaba, and Balqa), which are obstacles to providing consistent, timely, high-quality support.

It has also been hypothesized that the recent influx of Syrian refugee students had an impact on the education system and may be a contributing factor to the slower-than-intended progress in student performance. Although schools did report larger numbers of Syrian students at endline than midline, the performance levels of Syrian students were not below those of Jordanian students. However, this increased student population size did decrease the proportion of single-shift schools from 85% at midline to 72% at endline. Because two-shift schools provide students with reduced class time, it is reasonable to assume that this change has adversely affected student reading and mathematics progress. In addition, morning-shift schools had significantly lower results than all-day schools. In contrast, the results of afternoon-shift schools were relatively similar to those of all-day schools, likely because afternoon-shift schools have been targeted by various educational interventions implemented by international organizations seeking to serve the refugee students enrolled in those schools.

Overall, teachers and head teachers view RAMP favorably. The overwhelming majority of head teachers believe that the program provided sufficient training, materials, and support to their teachers and noted that RAMP successfully improved teaching. Similarly, more than 90% of teachers reported that they understood RAMP's goals, while more than 80% noted that the program improved math performance, reading performance, and student enthusiasm for learning in their classrooms.

Despite the support for RAMP within schools, a few system- and program-level factors likely had a negative impact on progress in performance. Most notably, the majority of teachers claimed that they did not receive materials on time for the start of the school year. This has clear implications for teachers' ability to implement the methodology as intended. Additionally, the proportion of classrooms with a single teacher from midline

to endline decreased, while the proportion of classrooms with at least three teachers throughout the year increased. The increases in the proportions of non-permanent teachers and those without pre-service training in early grade reading or mathematics, and the constant percentage of teachers who did not receive RAMP training, in addition to those issues mentioned above, are all likely to have been obstacles the consistent and faithful implementation of the new methodology and pedagogical approach in the classroom. Thus, these areas should be a continued focus for the system in the coming years.

4.2 Recommendations

There was consensus among endline results analysis workshop participants that the high-performing students are continuing to improve while the low-performing students are remaining behind and not progressing as intended. This hypothesis is supported by the analysis of reading fluency results, which showed simultaneous increases in both zero scores and the proportion of students reading 30 or more correct words per minute. Typically, teachers tend to try to cover the whole curriculum and the entire textbook during the school year, without verifying whether students are understanding and keeping up with the learning pace. This practice is likely to have the greatest impact on poorer-performing students who fall behind early and never catch up. Accordingly, the system must make particular effort to ensure that teachers apply differentiated instruction in all grades and subjects. It is recommended that the MoE provide additional guidance on how to employ differentiated instruction approaches in the classroom. Examples of such guidance include providing training and support to teachers on how to adapt the teaching pace to that of student learning; reframing the goal from finishing the textbook to focusing on student understanding, with a focus on weak readers; developing tools to facilitate the implementation of differentiated instruction, such as remedial activity books; and equipping teachers with a wider variety of instruction techniques to develop students' foundational skills in reading.

The issue of pace has become even more problematic because of the reduction in the number of Arabic language lessons per week and the shortened school day in two-shift schools. As a result, the MoE should increase the time allocated to reading in the curriculum and should instruct teachers to give students time to read a wider variety of text and make sure they are involved in more engaged reading activities. Additionally, it is important for the MoE to identify strategies to provide equal treatment and learning opportunities to all students, particularly those in two-shift schools. Furthermore, it is recommended that the MoE conduct a time-on-task study to assess the amount of time in a school day that students actually learn and work actively and the amount of time students spend reading and engaged with print materials throughout the day. This study would facilitate understanding how time is used during the day and where students could be provided additional opportunities to read and engage with print materials.

Relatedly, there are two learning periods that are not allocated to a specific topic (i.e., "extra-hours") and are usually used for extra-curricular activities. Although extra-curricular activities are important, it is recommended that the actual use of these periods be reviewed to ensure that they are being used effectively. If they are not, it may be

advisable to repurpose some of this time for reading, with a particular focus on supporting low-achieving students. For example, during this time, teachers could work exclusively with remedial groups, while other students engage in independent or group work/activities. This would be an ideal opportunity to use different strategies (e.g., practical exercises, concrete examples, games) that can be difficult to implement with a whole class and that can help children develop or reinforce their foundational skills by applying them in more engaging, focused activities.

Furthermore, classroom observations have shown that teachers tend to focus on and interact with only a small group of students during their lessons. Not surprisingly, these are typically the students who are highest performing and need the least support. Therefore, many children only minimally participate in lessons, and maintaining their focus is difficult. Accordingly, the next training program should focus on the following:

- Differentiated instruction techniques and remedial activities/remedial group management
- Classroom management techniques designed to ensure the involvement of all students in the learning process
- Formative evaluation, which will help teachers to adjust instruction to the actual learning pace of students and differentiate instruction as needed.

The data reflect a positive association between reading outside the classroom and student performance. This relationship was particularly evident in the predictive power of students bringing books home from schools for increased reading scores. Therefore, it is recommended that the Reading Incentive Program be a focus of additional support, including examinations of the fidelity of implementation and use by teachers and schools.

Lastly, it is clear from the endline results that the RAMP model works better in some governorates than in others. This variation is partly attributable to the level of fidelity with which the program is being implemented (e.g., how much coaching is occurring, the quality of coaching support, access to and use of learning materials). High fidelity of implementation will ultimately only be achieved with similarly high levels of accountability and support. Therefore, principals should play an increasing role in monitoring teachers and supporting them to ensure that they apply new teaching methodologies, including differentiated instruction. Additionally, supervisors should be held accountable for their coaching activities and the quality of the support they provide to teachers, and field directorates should be held accountable for monitoring performance and implementation fidelity.

The education system in Jordan is strong, and the RAMP methodology is sound. Working closely with the MoE at all stages, the program was able to improve instruction in the classroom and produce student performance gains in both reading and mathematics. The recommendations laid out in this section are intended to provide the MoE with guidance on how best to continue to build upon the successes of RAMP and how to overcome the obstacles encountered to ensure that all students in the country are provided with the opportunities to learn and succeed they deserve.

Annex 1: EGRA and EGMA Subtask Overview

Table 1-1. Early Grade Reading Assessment (EGRA) Instrument Subtasks Used at Baseline, Midline, and Endline

EGRA Subtask	2014	2017	2019	Skill	Description: The child is asked to…
Listening comprehension (*untimed*)	✓	✓	✓	Oral language comprehension and vocabulary	…listen to a story that the assessor reads out loud, and then verbally answer five questions about the story.
Letter-sound identification (*timed*)	✓	✓	✓	Alphabetic principle—letter-sound correspondence	…say the sound each letter makes, while looking at a printed page of 100 letters of the alphabet in random order, upper and lower case.
Syllable sounds (*timed*)	✓	✓	✓	Beginning decoding skills and identifying syllables from the language	…read a list of 50 syllables presented in random order.
Non-word reading (*timed*)	✓	✓	✓	Alphabetic principle—letter-sound correspondence and fluency (automatic decoding)	…read a list of 50 non-words printed on a page. Words are constructed from actual orthography but are not real words.
Oral reading (*timed*)	✓	✓	✓	Fluency (automatic word reading in context)	…read out loud a grade-level-appropriate short story printed on a page.
Reading comprehension (*untimed*)	✓	✓	✓	Comprehension	…verbally respond to five questions that the assessor asks about the short story.
Silent reading comprehension *(untimed)*			✓	Comprehension	…verbally respond to five questions that the assessor asks about a short story that was read silently

Table 1-2. Early Grade Mathematics Assessment (EGMA) Instrument Subtasks Used at Baseline, Midline, and Endline

EGMA Subtask	Timing	Skill	Description: The child is asked to…
Subtasks that assess procedural (recall) knowledge			
Number identification	Timed 60 sec.	The ability to identify written number symbols. If students cannot identify numbers, they cannot do mathematics.	…say the names of numbers presented on a page with 20 numbers. The numbers range from one- to two- and three-digit numbers.
Addition and subtraction Level 1 (L1) (basic facts)	Timed 60 sec.	Knowledge of and confidence with basic addition and subtraction. It is expected that students should develop some level of automaticity/ fluency with mathematics facts such as these since they are foundational mathematics skills.	…solve addition/subtraction problems, with sums/ differences below 20, without the aid of paper and pencil. The items range from problems with only single digits to problems that involve bridging the ten.[6] (10 items per addition and subtraction subtask)
Subtasks that assess conceptual (applied) knowledge			
Quantity discrimination (number comparison)	Not timed	The ability to make judgments about differences by comparing quantities, represented by numbers	…identify the larger of a pair of numbers. The number pairs used range from a pair of single-digit numbers to five pairs of double-digit numbers and four pairs of three-digit numbers. (10 items)
Missing number (number patterns)	Not timed	The ability to discern and complete number patterns	…determine the missing number in a pattern of four numbers, one of which is missing. Patterns used include counting forward and backward by ones, by fives, by tens, and by twos. (10 items)

[6] "Bridging the ten" refers to addition and subtraction situations where the addition and subtraction involves moving from one decade to the next. For example, 8 + 6 and 28 + 6 both involve "bridging the ten." A common strategy that may be adopted by students when bridging the ten mentally is first to "make" or "complete the ten"—e.g., 8 + 6 = 8 + 2 + 4 = 10 + 4 = 14, and 28 + 6 = 28 + 2 + 4 = 30 + 4 = 34.

EGMA Subtask	Timing	Skill	Description: The child is asked to…
Addition and subtraction Level 2 (L2)[7]	Not timed	The ability to use and apply the procedural addition and subtraction knowledge assessed in the L1 subtask to solve more complicated addition and subtraction problems	…solve addition/subtraction problems that involve the knowledge and application of the basic addition and subtraction facts assessed in the L1 subtask. Students could use any strategy that they wanted, including the use of paper and pencil supplied by the assessor. The problems extended to the addition and subtraction of two-digit numbers involving bridging. (Five items per addition and subtraction subtask)
Word problems	Not timed	The ability to interpret a situation (presented orally to the student), devise a plan, and solve the problem.	…solve problems presented orally, using any strategy that they wanted, including the use of paper and pencil and/or counters supplied by the assessor. The numerical values involved in the problem were deliberately small to allow for the targeted skills to be assessed without confounding problems with calculation skills that might otherwise impede performance. The problem situations used were designed to evoke different mathematical situations and operations. (Six items)

[7] The addition and subtraction L2 subtasks were more conceptual than the addition and subtraction L1 subtasks because the students had to understand what they were doing and apply the L1 skills. In other words, although the L2 subtasks were not purely conceptual—because with time and practice, students will develop some automaticity with the types of items in these subtasks—they were more conceptual than the L1 subtasks, especially for grade 2 students.

Annex 2: Assessor Training Overview

The Ministry of Education (MoE) and Early Grade Reading and Mathematics Initiative (RAMP) team nominated seven trainers for this activity. These included supervisors and MoE personnel who were selected based on their prior success as assessors; experience administering the Early Grade Reading Assessment (EGRA), Early Grade Mathematics Assessment (EGMA), and lot quality assurance sampling (LQAS); and experience working with primary-age children. These trainers received a 2-day training from RAMP to strengthen their familiarity with EGRA and EGMA protocols and other questionnaire tools administered as part of the endline survey. This was done to ensure that all trainers had a level of comfort with and a common understanding of the various subtask protocols and administration procedures.

In total, 113 assessors were trained from the 12 governorates, including 32 supervisors. The training sessions were designed to provide introductions to and ample time to practice with the EGRA, EGMA and survey questionnaire protocols to ensure that all head teacher, teacher, and student responses and scores were reliably collected and accurately recorded using Prodigy software on tablets.

To achieve the above mentioned outcomes of training, assessors were distributed into three groups to create more opportunities for assessors to work with each other and practice survey and assessment administration. Furthermore, training sessions ensured deeper engagement and better outcomes by having variety of activities that included:

- Trainer demonstrations of EGRA and EGMA administration
- Videos (filmed ahead of time and simulated children answering EGRA and EGMA tasks): Assessors watched the videos, marked the children's responses, and then discuss the assessors' inputs to reach a common understanding.
- Whole-group practice on protocols (using a "round-robin" approach to practicing items, particularly for the EGRA instrument)
- Small-group practice
- Pairs practice
- Simulation in which one of the assessors or trainers played the role of an assessor (in other cases, the role of a student) making mistakes or not following protocols, while all other assessors were asked to provide feedback on issues/mistakes and ways to improve.
- School visits: Two school visits were conducted during the training to provide assessors with an opportunity to practice administering EGRA and EGMA to children and using tablets in conditions similar to those they would encounter during actual data collection.
- Assessor Accuracy Measuring (AAM): Two sessions were held to evaluate the accuracy of assessors and the degree to which the assessors agreed in their scoring of the same observation.

The second AAM showed that the average of accuracy among assessors was 90%. All 113 assessors passed the benchmark of 85% accuracy and were, therefore, eligible to collect data.

Annex 3: Oral Reading Passage Equating Report

Five new reading passages were created and piloted alongside the 2014 Early Grade Reading Assessment (EGRA) reading passage. The pilot design consisted of five forms of the assessment, with each form consisting of the 2014 passage and two new passages. The order of administration of passages within each form was randomized to reduce testing effects. The following list displays the passages in each form:

- Form 1: 2014 Passage; Passage A; Passage B
- Form 2: 2014 Passage; Passage C; Passage D
- Form 3: 2014 Passage; Passage E; Passage A
- Form 4: 2014 Passage; Passage B; Passage C
- Form 5: 2014 Passage; Passage D; Passage E

Each form was administered to approximately 100 students; thus, all 500 students read the 2014 passage, and approximately 200 students read each of the new passages. The means and standard deviations of oral reading fluency (ORF) scores for each of the passages are displayed in ***Table 2-1***. Despite efforts to create passages of similar difficulty, the mean fluency score for each passage differed from that of the 2014 passage.

Table 2-1. Mean ORF Scores by Passage

Passage	N	2014 passage	New passage
A	199	24.1 (20.3)	19.8 (19.7)
B	201	23.1 (19.8)	17.2 (19.9)
C	202	23.1 (18.7)	19.1 (19.1)
D	199	22.8 (17.9)	14.2 (14.2)
E	199	24.2 (18.9)	23.5 (19.5)

Note: Standard deviations in parentheses.

Based on the mean ORF rates, the results show that the 2014 passage was significantly easier than passages A, B, C, and D. The difference between the mean fluency score on passage E and the 2014 passage, on the other hand, was minimal (0.7 correct words per minute). The general rule of thumb states that if the difference in scores is less than one tenth of a standard deviation (i.e., approximately 1.8 to 2 words per minute in these data), the error associated with equating is higher than the difference accounted for by equating. In other words, there is no clear advantage to equating in these situations. Equating formulas are provided for passage E for

consistency, but the recommendation is that this passage does not need to be adjusted in future administrations.

There are many viable approaches to equating ORF scores using a common-persons approach (i.e., the same set of students taking multiple forms of the assessment). Because initial analyses showed a linear relationship between the scores on the passages, linear equating was used as the preferred approach (for simplicity). Although both a classical test theory linear equating approach and a linear regression scaling approach were initially examined, the regression approach provided smaller amounts of bias and was, therefore, chosen for the final approach. The results of the equating are displayed in ***Table 2-2***.

Table 2-2. Mean Equated ORF Scores by Passage

Passage	N	2014 Passage	Equated Passage
A	199	24.1 (20.3)	23.7 (19.2)
B	201	23.1 (19.8)	22.5 (18.7)
C	202	23.1 (18.7)	22.8 (17.8)
D	199	22.8 (17.9)	22.6 (16.9)
E	199	24.2 (18.9)	24.0 (17.4)

Note: Standard deviations in parentheses.

The equated scores all show significant reductions in the difference across passages, with 0.6 correct words per minute as the largest difference between any given passage and the 2014 passage.

For future administrations of the newly created passages, the following equations should be used to scale the ORF scores to be comparable to those on the 2014 assessment (with the caveat about not needing to equate passage E):

- Passage A: Equated_A = Passage_A*0.9511093 + 5.252185
- Passage B: Equated_B = Passage_B*0.9093058 + 7.426361
- Passage C: Equated_C = Passage_C*0.9110216 + 5.692733
- Passage D: Equated_D = Passage_D*1.168624 + 6.217576
- Passage E: Equated_E = Passage_E*0.878969 + 3.51767

Annex 4: Overall EGRA Results, Including Zero Scores

Subtask	Measure	2014 G2 and G3	2017 G2 and G3	2019 G2 and G3
Listening comprehension	% correct	67.2 [±2.2]	60.6 *** [±2.0]	58 [±2.1]
	% of students with zero scores	4.8 [±1.7]	5.9 [±1.1]	6.9 [±1.2]
Letter sound	fluency (correct letters per min.)	37 [±3.9]	47.9 *** [±2.0]	50.1 [±1.9]
	% correct of items attempted	64.9 [±5.5]	76.1 *** [±2.2]	79.1 [±2.1]
	% of students with zero scores	21.2 [±4.9]	7.3 *** [±1.4]	7.6 [±1.7]
Syllable sound	fluency (correct syllables per min.)	25.2 [±2.0]	31.7 *** [±1.7]	33 [±1.7]
	% correct of items attempted	63.8 [±3.7]	72.5 *** [±2.4]	71 [±2.3]
	% of students with zero scores	10.8 [±3.0]	4.8 *** [±1.3]	8.8 *** [±1.6]
Invented words	fluency (correct words per min.)	8.9 [±.9]	12.2 *** [±.7]	15.1 *** [±.8]
	% correct of items attempted	39.4 [±3.6]	49.3 *** [±2.3]	57 *** [±2.1]
	% of students with zero scores	30.7 [±4.7]	15.1 *** [±1.7]	10.1 *** [±1.9]
Oral reading	oral reading fluency	26.6 [±2.2]	25.9 [±1.4]	27.1 [±1.7]
	% correct of items attempted	63.7 [±4.3]	60.4 [±2.6]	60.1 [±2.7]
	% of students with zero scores	9.1 [±2.9]	10.4 [±1.8]	14.8 ** [±2.2]
Reading comprehension	% correct of items attempted	45.3 [±4.3]	49.8 [±2.8]	49.3 [±2.9]
	% correct	33.7 [±3.5]	36.7 [±2.4]	38.6 [±2.9]
	% of students with 80% comp.	17.9 [±3.5]	21.3 [±2.4]	23.6 [±3.0]
	% of students with zero scores	34.1 [±4.9]	30.5 [±3.5]	32.5 [±3.4]

Annex 5: EGRA Results by Grade, Including Zero Scores

Subtask	Measure	2014 G2	2017 G2	2019 G2	2014 G3	2017 G3	2019 G3
Listening comprehension	% correct	62.5 [±2.8]	56.9 ** [±2.6]	52.8 * [±2.8]	72.5 [±2.7]	64.4 *** [±2.2]	62.9 [±2.4]
	% of students with zero scores	6.6 [±2.8]	8.2 [±1.6]	8.9 [±1.9]	2.9 [±1.1]	3.6 [±1.1]	4.9 [±1.5]
Letter sound	fluency (correct letters per min.)	38.2 [±4.3]	47.3 *** [±2]	48 [±2.4]	35.7 [±4.1]	48.6 *** [±3.1]	52.1 [±2.2]
	% correct of items attempted	67.4 [±6.3]	77.8 ** [±1.9]	79.4 [±2.4]	62.2 [±5.4]	74.3 *** [±3.5]	78.7 * [±2.5]
	% of students with zero scores	19.6 [±5.7]	6.7 *** [±1.6]	7.4 [±2.0]	23.1 [±5.0]	7.8 *** [±2.3]	7.8 [±2.1]
Syllable sound	fluency (correct syllables per min.)	22.1 [±2.4]	29.8 *** [±2.2]	29.8 [±1.9]	28.7 [±2.1]	33.8 *** [±2.3]	36.1 [±2]
	% correct of items attempted	60.1 [±5.4]	72.3 *** [±3.1]	68.1 [±3.0]	68 [±3.5]	72.7 [±3.3]	73.8 [±2.5]
	% of students with zero scores	13.2 [±4.3]	4.8 *** [±1.5]	11.7 *** [±2.5]	8.1 [±2.8]	4.7 [±2.4]	6 [±1.5]
Invented words	fluency (correct words per min.)	7.1 [±.0]	10.8 *** [±.0]	13.3 *** [±.9]	10.9 [±1.0]	13.7 *** [±.9]	16.9 *** [±.0]
	% correct of items attempted	34.5 [±4.3]	48.2 *** [±3.3]	55.3 *** [±2.8]	44.9 [±3.6]	50.4 * [±2.6]	58.6 *** [±2.4]
	% of students with zero scores	37.2 [±6.4]	16.8 *** [±2.6]	13.1 [±2.8]	23.4 [±4.4]	13.4 *** [±2.4]	7.2 *** [±2.1]
Oral reading	oral reading fluency	19.1 [±2.2]	20.7 [±1.5]	20.2 [±1.5]	35 [±2.8]	31.2 * [±2.1]	33.6 [±2.3]
	% correct of items attempted	56.7 [±6.7]	56.2 [±3.6]	53 [±3.1]	71.6 [±3.2]	64.7 ** [±3.3]	66.9 [±3.0]
	% of students with zero scores	12.5 [±4.4]	12.6 [±2.5]	20.7 *** [±3.3]	5.3 [±2.3]	8.2 [±2.5]	9.2 [±2.3]
Reading comprehension	% correct of items attempted	33.3 [±5.3]	42.7 ** [±4.0]	41.2 [±3.2]	58.6 [±3.9]	57.3 [±3.4]	57 [±3.5]
	% correct	21.7 [±3.9]	27 * [±2.9]	28.4 [±2.9]	47.1 [±3.7]	46.8 [±3.4]	48.3 [±3.7]
	% of students with 80% comp.	7.9 [±3.3]	11.4 [±2.2]	13.5 [±2.9]	29 [±5.4]	31.5 [±3.7]	33.3 [±4.3]
	% of students with zero scores	47.2 [±7.4]	40.4 [±5.2]	43 [±3.8]	19.5 [±4.4]	20.2 [±4.1]	22.6 [±3.7]

Annex 6: EGRA Results by Gender, Including Zero Scores

Subtask	Measure	2014 Male G2 and G3	2014 Female G2 and G3	2017 Male G2 and G3	2017 Female G2 and G3	2019 Male G2 and G3	2019 Female G2 and G3
Listening comprehension	% correct	65.3 [±3.4]	68.9 [±2.4]	60.7 [±2.5]	60.5 [±2.6]	57.9 [±2.8]	58 [±2.3]
	% of students with zero scores	5.5 [±2.5]	4.3 [±1.8]	5.7 [±1.5]	6.1 [±1.6]	6.9 [±1.6]	6.9 [±1.8]
Letter sound	fluency (correct letters per min.)	36.2 [±5.2]	37.7 [±3.4]	47.8 *** [±2.2]	48.1 *** [±3.0]	48.5 [±2.4]	51.6 [±2.2]
	% correct of items attempted	63.5 [±6.9]	66.3 [±5.1]	77.8 *** [±2.1]	74.6 ** [±3.4]	77.7 [±3.0]	80.3 ** [±2.4]
	% of students with zero scores	23.8 [±6.5]	19 [±4.6]	5.8 *** [±1.4]	8.5 *** [±2.3]	7.9 [±2.1]	7.4 [±2.0]
Syllable sound	fluency (correct syllables per min.)	22.4 [±2.5]	27.8 [±2.3]	28.8 *** [±2.2]	34.2 *** [±2.3]	30.8 [±2.2]	35.1 [±1.8]
	% correct of items attempted	59.4 [±4.8]	67.8 [±3.5]	68.8 *** [±2.8]	75.6 *** [±3.3]	68.1 [±3.4]	73.7 [±2.4]
	% of students with zero scores	14.5 [±4.1]	7.5 [±2.9]	5.5 *** [±1.5]	4.2 [±2.1]	10.9 *** [±2.7]	6.8 [±1.6]
Invented words	fluency (correct words per min.)	7.6 [±.9]	10.1 [±1.1]	10.9 *** [±1.0]	13.4 *** [±.9]	14.2 *** [±1.1]	15.9 *** [±.8]
	% correct of items attempted	35.3 [±4.2]	43.1 [±3.8]	45.7 *** [±3.2]	52.3 *** [±2.6]	54.4 *** [±3.0]	59.3 *** [±2.4]
	% of students with zero scores	33.2 [±6.7]	28.4 [±5.4]	18.2 *** [±2.8]	12.5 *** [±2.1]	13 [±2.9]	7.4 *** [±1.9]
Oral reading	oral reading fluency	22.5 [±2.5]	30.3 [±2.8]	23 [±1.9]	28.2 [±1.8]	25.4 [±2.4]	28.5 [±1.8]
	% correct of items attempted	58.3 [±5.6]	68.6 [±3.7]	55.4 [±3.5]	64.5 [±3.3]	57.1 [±3.7]	62.9 [±2.9]
	% of students with zero scores	12.9 [±4.1]	5.7 [±2.3]	12.8 [±2.5]	8.5 [±2.5]	17.8 [±3.3]	12 [±2.8]
Reading comprehension	% correct of items attempted	41.2 [±5.3]	48.9 [±4.5]	47.3 [±4.1]	51.9 [±3.3]	48.2 [±4.2]	50.4 [±3.2]
	% correct	28.8 [±4.0]	38.1 [±3.9]	33.2 [±3.4]	39.7 [±3.0]	36.9 [±4.0]	40.2 [±3.0]
	% of students with 80% comp.	13.6 [±4.4]	21.7 [±4.4]	18.3 [±3.2]	23.7 [±3.0]	22.3 [±3.9]	24.8 [±3.2]
	% of students with zero scores	41.6 [±5.9]	27.4 [±5.3]	34.9 [±5.3]	26.9 [±4.0]	35.7 [±4.7]	29.7 [±3.7]

Annex 7: EGRA Results by Nationality (Jordanian versus Syrian refugee)

Subtask	Measure	Syrian 2017 G2 and G3	Jordanian 2017 G2 and G3	Syrian 2019 G2 and G3	Jordanian 2019 G2 and G3
Listening comprehension	% correct	65.1 [±6.9]	59.9 [±2.0]	58.8 [±7.4]	57.9 [±2.1]
	% of students with zero scores	3.1 [±2.6]	6.5 [±1.2]	6.7 [±3.1]	6.8 [±1.2]
Letter sound	fluency (correct letters per min.)	39 [±6.9]	49.3 [±1.7]	49.9 * [±6.3]	50.1 [±1.9]
	% correct of items attempted	68 [±10.2]	77.5 [±1.6]	77.8 [±5.4]	79.2 [±2.2]
	% of students with zero scores	10.1 [±5.8]	6.7 [±1.2]	9.1 [±4.0]	7.4 [±1.9]
Syllable sound	fluency (correct syllables per min.)	23.9 [±3.4]	32.7 [±1.7]	32.2 * [±6.2]	33.1 [±1.7]
	% correct of items attempted	61.5 [±5.8]	73.9 [±2.1]	66.5 [±7.1]	71.7 [±2.5]
	% of students with zero scores	9 [±7.3]	4.3 [±1.0]	12.1 [±4.0]	8.2 *** [±1.7]
Invented words	fluency (correct words per min.)	9.5 [±1.3]	12.6 [±.7]	15.4 *** [±3.0]	15.1 *** [±.8]
	% correct of items attempted	40.6 [±4.0]	50.4 [±2.2]	53.8 *** [±7.5]	57.4 *** [±2.3]
	% of students with zero scores	21.9 [±5.0]	14.2 [±1.6]	12.9 * [±5.2]	9.6 ** [±2.2]
Oral reading	oral reading fluency	21.1 [±3.1]	26.3 [±1.4]	27.5 * [±5.6]	27 [±1.9]
	% correct of items attempted	48.5 [±6.0]	61.8 [±2.4]	57.3 [±8.4]	60.6 [±3.0]
	% of students with zero scores	19.2 [±5.8]	9.4 [±1.6]	16.6 [±6.4]	14.6 *** [±2.5]
Reading comprehension	% correct of items attempted	43.2 [±5.6]	50.3 [±3.1]	48.2 [±8.5]	49.5 [±3.2]
	% correct	31.8 [±5.3]	36.9 [±2.7]	40.6 [±8.1]	38.2 [±3.1]
	% of students with 80% comp.	18 [±7.3]	21.4 [±2.6]	26.4 [±8.1]	23.1 [±3.1]
	% of students with zero scores	38.8 [±8.2]	30 [±3.8]	34.1 [±8.6]	32.4 [±3.8]

* p<.05; ** p<.01; *** p<.001

Annex 8: EGRA Results by Governorate, Including Zero Scores

Subtask	Measure	2017 Ajloun G2	2019 Ajloun G2	2017 Ajloun G3	2019 Ajloun G3	2017 Amman G2	2019 Amman G2	2017 Amman G3	2019 Amman G3
Listening comprehension	% correct	52.1 [±4.1]	50.8 [±5.4]	61.7 [±4.4]	70.3 ** [±3.8]	56.4 [±4.5]	48.4 * [±5.8]	67 [±4.4]	60.4 [±5.1]
	% of students with zero scores	5.7 [±3.2]	13.2 [±7.5]	4.6 [±2.8]	1.8 [±1.9]	10.6 [±3.6]	7.3 [±3.9]	1.9 [±2.2]	5.5 [±3.7]
Letter sound	fluency (correct letters per min.)	34.9 [±3.4]	40.6 [±5.3]	40.4 [±5.0]	45.3 [±3.7]	42.8 [±4.0]	41.2 [±4.3]	45 [±4.9]	49.1 [±4.8]
	% correct of items attempted	64.8 [±5.4]	66.1 [±7.4]	64.3 [±6.0]	69 [±4.8]	73.8 [±3.3]	76.1 [±4.9]	71 [±5.6]	79 [±5.9]
	% of students with zero scores	15.3 [±5.3]	19.6 [±8.2]	12.1 [±4.9]	11.9 [±5.2]	7.5 [±4.5]	5.8 [±4.7]	11.1 [±4.8]	5.4 [±4.9]
Syllable sound	fluency (correct syllables per min.)	28 [±3.0]	37.6 *** [±3.7]	35.7 [±3.8]	44.7 ** [±4.1]	29.4 [±4.6]	24.2 [±3.6]	33.9 [±4.1]	33.1 [±4.5]
	% correct of items attempted	73.4 [±5.3]	78.9 [±5.6]	76.6 [±5.7]	83.4 [±4.5]	74.4 [±7.4]	60.9 ** [±6.6]	74.7 [±5.0]	73.7 [±6.5]
	% of students with zero scores	6.2 [±3.5]	5 [±5.5]	6.7 [±4.5]	2.1 [±3.6]	4.5 [±3.2]	15.6 *** [±6.3]	3.2 [±3.4]	3.6 [±2.4]
Invented words	fluency (correct words per min.)	9.8 [±1.2]	16 *** [±2.0]	14.9 [±2.2]	21 *** [±2.2]	10.7 [±2.4]	10.6 [±1.7]	14.3 [±2.2]	15.1 [±2.3]
	% correct of items attempted	48.4 [±5.5]	63.5 *** [±5.6]	53.5 [±6.0]	67.6 *** [±4.5]	49.9 [±8.6]	47.3 [±5.7]	52.9 [±6.1]	55.5 [±5.6]
	% of students with zero scores	18.2 [±5.5]	6.5 ** [±5.2]	14.6 [±6.5]	3.3 ** [±3.7]	16.8 [±7.4]	18.4 [±7.3]	12.6 [±5.3]	6 [±5.4]
Oral reading	oral reading fluency (ORF)	21.3 [±2.4]	24.8 [±3.6]	35.5 [±3.3]	46 *** [±4.9]	20.5 [±3.0]	16.5 [±2.7]	32.2 [±3.8]	30.5 [±5.5]
	% correct of items attempted	60.3 [±5.5]	64.9 [±6.1]	71.9 [±5.0]	81.5 ** [±3.9]	56.2 [±8.4]	45 * [±6.4]	65.6 [±5.3]	63.2 [±7.6]
	% of students with zero scores	11.8 [±5.0]	13.9 [±7]	4 [±3.1]	3.3 [±3.4]	14.8 [±6.9]	23 [±7.6]	6.1 [±3.7]	8.4 [±5.6]
Reading comprehension	% correct of items attempted	43.9 [±7.6]	45.9 [±8.1]	65.1 [±6.0]	70.9 [±4.3]	45.6 [±6.8]	33.3 ** [±4.4]	63.7 [±5.8]	55.4 [±8.3]
	% correct	27.3 [±5.3]	31.9 [±6.3]	54.9 [±5.8]	65 * [±5.8]	27.9 [±5.1]	21.7 [±4.1]	52.4 [±7.1]	46.2 [±9.0]
	% of students with 80% comp.	11 [±4.7]	12.3 [±6.1]	38.9 [±7.5]	51 [±10.5]	9.4 [±4.3]	6.8 [±3.8]	37 [±9.3]	31.6 [±10.2]
	% of students with zero scores	40 [±9.1]	32.4 [±9.6]	14.4 [±5.6]	8.2 [±4.7]	36.6 [±8.8]	50.6 * [±7.6]	14.3 [±5.2]	26.8 * [±9.6]

* p<.05; ** p<.01; *** p<.001

Subtask	Measure	2017 Aqaba G2	2019 Aqaba G2	2017 Aqaba G3	2019 Aqaba G3	2017 Balqa G2	2019 Balqa G2	2017 Balqa G3	2019 Balqa G3
Listening comprehension	% correct	52.3 [±4.3]	50.7 [±4.9]	57 [±3.9]	66.7 ** [±4.7]	58 [±6.7]	49.6 [±7.1]	63.3 [±5.6]	58.7 [±6.5]
	% of students with zero scores	9.1 [±4.0]	10.1 [±5.2]	7.1 [±3.6]	2.9 [±2.4]	9.4 [±5.5]	14.2 [±6.9]	5.2 [±4.1]	8.6 [±6.4]
Letter sound	fluency (correct letters per min.)	46.2 [±3.5]	50.5 [±4.6]	49.7 [±3.2]	58 ** [±4.8]	54.3 [±3.5]	53.8 [±4.9]	57.8 [±4.0]	53.3 [±5.1]
	% correct of items attempted	80.9 [±3.8]	83.1 [±4.1]	77.7 [±3.8]	85.7 ** [±4.2]	89 [±3.1]	85.3 [±4.2]	87.7 [±2.9]	79.4 * [±5.8]
	% of students with zero scores	6 [±3.4]	4.5 [±3.5]	8.7 [±3.5]	2.2 ** [±2.3]	2.2 [±2.4]	4.5 [±4.0]	1.7 [±2.2]	10.2 *** [±4.6]
Syllable sound	fluency (correct syllables per min.)	21 [±2.4]	27.8 ** [±4.2]	28 [±2.4]	36.9 *** [±4.3]	26.8 [±4.3]	30.7 [±4.5]	31 [±4.2]	33 [±5.5]
	% correct of items attempted	59.5 [±4.9]	64.2 [±7.7]	67.1 [±3.7]	74.9 * [±5.1]	70.3 [±6.1]	71.3 [±7.2]	72.9 [±5.8]	69.7 [±9.1]
	% of students with zero scores	11.5 [±4.8]	12.6 [±7.0]	5.2 [±2.7]	4.4 [±3.8]	3.4 [±3.6]	6.3 [±4.0]	3.1 [±2.8]	9.3 [±9.0]
Invented words	fluency (correct words per min.)	7.6 [±1.1]	12.9 *** [±1.8]	10.3 [±1.4]	17.3 *** [±2.0]	10.4 [±2.2]	13.1 [±1.9]	13.1 [±2.1]	15.1 [±2.5]
	% correct of items attempted	38.5 [±4.8]	55.5 *** [±6.2]	39.9 [±3.9]	60.7 *** [±5.0]	48.4 [±5.9]	54.6 [±5.7]	51.3 [±5.4]	53.3 [±6.9]
	% of students with zero scores	27.8 [±6.2]	14 ** [±6.7]	19.2 [±4.6]	6.5 ** [±4.7]	8.5 [±5.4]	7.7 [±6.1]	7.5 [±4.6]	9.9 [±8.8]
Oral reading	ORF	14.2 [±1.9]	18.4 * [±2.8]	24.7 [±2.3]	29.6 * [±3.3]	20.3 [±4.1]	21.8 [±3.8]	30.5 [±5.6]	31.7 [±6.0]
	% correct of items attempted	44 [±5.0]	50.7 [±6.8]	56.1 [±3.9]	67 ** [±6.3]	57 [±7.6]	54.3 [±7.5]	66.3 [±7.8]	63.8 [±9.1]
	% of students with zero scores	22.3 [±6.7]	23.5 [±8.5]	7.6 [±3.3]	8 [±5.1]	9.8 [±4.9]	18.2 [±8.9]	6.4 [±5.1]	11.8 [±8.2]
Reading comprehension	% correct of items attempted	29.4 [±5.4]	40.3 * [±6.7]	51.7 [±5.2]	55.9 [±6.5]	38.8 [±11.5]	41.5 [±7.2]	53.2 [±10.9]	48.2 [±8.8]
	% correct	15.6 [±3.5]	24.9 ** [±4.8]	39.1 [±5.0]	46.5 [±6.1]	24.1 [±8.9]	29.4 [±6.1]	44.3 [±11.3]	40.6 [±9.1]
	% of students with 80% comp.	3.9 [±3.0]	9.1 * [±3.8]	23.9 [±5.9]	36.1 ** [±7.4]	11.3 [±6.4]	13.2 [±5.9]	32 [±10.8]	23 [±9.4]
	% of students with zero scores	57 [±7.2]	48.3 [±8.2]	27.4 [±5.6]	28.1 [±7.8]	47.7 [±13.5]	39.5 [±10.3]	27.4 [±12.2]	26.7 [±10.5]

* p<.05; ** p<.01; *** p<.001

Subtask	Measure	2017 Irbid G2	2019 Irbid G2	2017 Irbid G3	2019 Irbid G3	2017 Jarash G2	2019 Jarash G2	2017 Jarash G3	2019 Jarash G3
Listening comprehension	% correct	61.9 [±6.3]	61.5 [±9.4]	65.2 [±4.2]	66.2 [±7.0]	50 [±4.9]	46.1 [±4.0]	57 [±4.4]	63.1 [±5.8]
	% of students with zero scores	6.4 [±3.5]	8.7 [±7.0]	1.3 [±1.7]	2.5 [±2.3]	8.7 [±4.9]	9.1 [±5.8]	4.2 [±3.3]	5.1 [±5.8]
Letter sound	fluency (correct letters per min.)	54.3 [±5.5]	54.9 [±7.7]	55.8 [±5.7]	57.4 [±5.3]	49.5 [±5.2]	56.5 [±5.4]	53.2 [±4.7]	62.3 ** [±3.7]
	% correct of items attempted	77.8 [±5.2]	82.9 [±7.4]	79.4 [±4.7]	81 [±6.4]	78.4 [±5.5]	82 [±7.1]	75.7 [±5.2]	84.1 * [±4.8]
	% of students with zero scores	7 [±3.6]	9.4 [±6.8]	1.3 [±1.8]	9.6 *** [±6.5]	8.6 [±3.9]	9.4 [±6.3]	8.8 [±3.8]	5.1 [±5.4]
Syllable sound	fluency (correct syllables per min.)	36.1 [±3.6]	39 [±5.3]	40.6 [±5]	47.7 [±5.5]	28 [±3.2]	35.4 * [±4.5]	34.3 [±3]	41.5 * [±4.9]
	% correct of items attempted	74.5 [±5.7]	79.1 [±7.1]	76.8 [±5.9]	85.8 * [±4.6]	68.8 [±5.8]	73.3 [±7.1]	72.3 [±4.2]	76.5 [±6.0]
	% of students with zero scores	1.2 [±1.6]	9 *** [±6.1]	1.1 [±2.0]	2.4 [±2.7]	3.4 [±4.0]	4.5 [±3.6]	5 [±4.1]	1 * [±1.4]
Invented words	fluency (correct words per min.)	14.4 [±2.1]	18.1 * [±2.7]	16.7 [±2.2]	23.1 *** [±2.1]	9.1 [±1.5]	15.9 *** [±2.0]	12.8 [±1.6]	19.5 *** [±2.7]
	% correct of items attempted	54.9 [±6.5]	68.9 ** [±6.8]	55.1 [±5.4]	73.6 *** [±4.3]	41 [±5.0]	56.9 *** [±5.5]	45.9 [±4.6]	60.9 *** [±6.3]
	% of students with zero scores	9.3 [±4.8]	7.4 [±4.5]	5.6 [±4.3]	2.3 [±2.9]	15.4 [±6.3]	5.7 * [±5.1]	16.2 [±7.8]	2.8 *** [±3.1]
Oral reading	ORF	25.6 [±3.6]	27.5 [±5.4]	35.8 [±4.5]	43.7 * [±6.0]	21 [±3.1]	26.1 * [±3.5]	31.4 [±2.6]	41.6 ** [±6.2]
	% correct of items attempted	63.3 [±7.3]	67.1 [±9.0]	70.2 [±7.3]	81 * [±5.2]	53.9 [±5.7]	58.7 [±6.7]	65 [±4.3]	73.2 * [±5.0]
	% of students with zero scores	5.2 [±4.0]	10.9 [±8.2]	5.7 [±5.3]	3.2 [±3.9]	7.8 [±4.3]	8.8 [±7]	4.9 [±3.6]	1 ** [±1.2]
Reading comprehension	% correct of items attempted	48.5 [±6.0]	54.8 [±12.5]	57.5 [±7.1]	65.5 [±8.9]	41.3 [±5.4]	53.2 ** [±6.3]	61.9 [±4.2]	69.7 * [±5.4]
	% correct	35.9 [±6.6]	42.4 [±11.4]	50.8 [±7.6]	59 [±9.6]	25.5 [±4.7]	37.6 ** [±6.0]	48 [±4.3]	60.3 ** [±6.4]
	% of students with 80% comp.	19.4 [±7.2]	27.4 [±11.9]	34 [±9.6]	44.7 [±12.4]	9.8 [±4.9]	15.2 [±5.6]	26.2 [±6.5]	37.4 [±10.5]
	% of students with zero scores	29.2 [±7.3]	29.1 [±12.3]	12 [±6.2]	11.1 [±5.1]	39.7 [±7.8]	22.8 ** [±7.1]	11.6 [±5]	4 ** [±2.7]

* p<.05; ** p<.01; *** p<.001

Subtask	Measure	2017 Karak G2	2019 Karak G2	2017 Karak G3	2019 Karak G3	2017 Ma'an G2	2019 Ma'an G2	2017 Ma'an G3	2019 Ma'an G3
Listening comprehension	% correct	53.6 [±7.0]	72.8 *** [±5.2]	62.6 [±7.2]	77.3 * [±8.3]	53 [±6.5]	44.2 * [±5.5]	56.5 [±4.6]	58.8 [±5.7]
	% of students with zero scores	11.7 [±6.7]	3.3 [±3.6]	8 [±4.4]	5 [±6.3]	11.9 [±6.4]	17 [±5.1]	7.4 [±5.1]	7.8 [±3.6]
Letter sound	fluency (correct letters per min.)	36.3 [±5.6]	52.2 ** [±7.8]	45.6 [±6.2]	56.7 * [±7.7]	37.8 [±8.2]	48.6 * [±5.9]	50.3 [±12.0]	55.6 [±7.7]
	% correct of items attempted	62.6 [±7.2]	83.5 *** [±7.5]	69.1 [±7.7]	84.3 ** [±7.9]	69.8 [±11.3]	80.1 [±5.9]	74.2 [±9.2]	78.4 [±8.6]
	% of students with zero scores	16.9 [±6.6]	4.3 ** [±4.1]	12.5 [±5.8]	3.8 [±5.8]	8.8 [±4.3]	7.2 [±3.7]	11.1 [±6.7]	11.8 [±8.2]
Syllable sound	fluency (correct syllables per min.)	26.1 [±5.6]	40.8 *** [±5.0]	33 [±4.7]	46.5 ** [±7.1]	21.6 [±6.6]	24.5 [±5.0]	32.7 [±6.0]	30.7 [±5.0]
	% correct of items attempted	61.6 [±9.5]	83.3 *** [±5.6]	71 [±6.7]	84.6 * [±7.5]	58.1 [±13.7]	59 [±9.6]	71.9 [±7.8]	64.7 [±8.2]
	% of students with zero scores	13.7 [±9.0]	3.1 * [±4.9]	4.5 [±4.2]	5.3 [±8.1]	17.3 [±11.4]	16.5 [±9.7]	7.1 [±5.7]	11.2 [±6.2]
Invented words	fluency (correct words per min.)	8.5 [±2.4]	18.9 *** [±2.6]	12.7 [±2.4]	22.8 *** [±3.6]	7.6 [±2.8]	10.8 [±2.1]	12.3 [±2.6]	14.4 [±2.4]
	% correct of items attempted	37.9 [±8.6]	71.3 *** [±6.1]	46 [±6]	71.8 *** [±6.4]	34 [±11.1]	47.2 [±7.8]	47.1 [±9.5]	48.2 [±6.0]
	% of students with zero scores	29.3 [±10.8]	4 *** [±3.5]	14.3 [±5.6]	2.5 ** [±4.4]	43.3 [±16.5]	18.2 * [±10.2]	26.8 [±9.5]	12.3 ** [±5.2]
Oral reading	ORF	20.5 [±4.9]	28.9 ** [±3.8]	30.8 [±5.0]	41.4 * [±6.9]	16 [±5.4]	16.5 [±3.6]	34 [±9.1]	30.4 [±3.5]
	% correct of items attempted	52.1 [±9.3]	73.2 *** [±7.4]	63.8 [±7.3]	80.7 ** [±8.8]	45.8 [±14.1]	45.2 [±8.5]	66.6 [±8.2]	61.6 [±5.5]
	% of students with zero scores	19.8 [±9.3]	9.1 [±7.1]	7.9 [±6.2]	3.9 [±7.3]	28.3 [±19.8]	25.8 [±9.5]	10.8 [±5.9]	11.1 [±4.3]
Reading comprehension	% correct of items attempted	37.6 [±9.4]	65.4 *** [±7.6]	52.4 [±9.7]	73.6 ** [±11.7]	31.7 [±13.0]	33.9 [±8.0]	59.3 [±9.7]	53.6 [±5.7]
	% correct	25.7 [±7.8]	48.2 *** [±7.0]	44.5 [±9.6]	65.6 ** [±11.3]	18.7 [±8.1]	20.9 [±5.7]	48.1 [±12.0]	46 [±5.5]
	% of students with 80% comp.	9.8 [±6.2]	31.2 *** [±7.4]	29.9 [±10.2]	55.4 ** [±12.3]	8.5 [±6.5]	4.3 [±4.2]	37.6 [±16.8]	29.6 [±7.4]
	% of students with zero scores	44.7 [±12.8]	18.1 *** [±8.9]	23.8 [±10.5]	13.4 [±12.7]	57.6 [±15.0]	48.1 [±10.3]	21.5 [±8.2]	24.2 [±6.0]

* p<.05; ** p<.01; *** p<.001

Subtask	Measure	2017 Madaba G2	2019 Madaba G2	2017 Madaba G3	2019 Madaba G3	2017 Mafraq G2	2019 Mafraq G2	2017 Mafraq G3	2019 Mafraq G3
Listening comprehension	% correct	57.9 [±4.4]	55.1 [±6.1]	58.9 [±4.8]	61.2 [±5.9]	43 [±7.5]	49.6 [±5.7]	52.3 [±5.4]	58.2 [±6.7]
	% of students with zero scores	5.4 [±3.6]	9 [±4.3]	7.4 [±4.6]	5.5 [±3.9]	12.4 [±6.1]	8.6 [±4.8]	9 [±6.7]	3 [±3.8]
Letter sound	fluency (correct letters per min.)	52.9 [±2.6]	55.9 [±5.6]	55.8 [±3.5]	56.9 [±5.2]	43.2 [±6.0]	45.8 [±6.3]	43.1 [±6.1]	52.9 * [±5.9]
	% correct of items attempted	87.8 [±3.7]	83.3 [±5.4]	83.3 [±3.9]	79.1 [±6.9]	82.4 [±5.2]	76.3 [±5.4]	73.3 [±6.4]	77.4 [±5.2]
	% of students with zero scores	1.4 [±1.8]	5.9 * [±4.4]	3.9 [±2.9]	9.2 [±7.6]	4.6 [±4.5]	8.6 [±3.9]	10.6 [±5.2]	8.3 [±3.8]
Syllable sound	fluency (correct syllables per min.)	26 [±2.4]	27.1 [±3.6]	31.6 [±2.7]	33.8 [±5.2]	20.4 [±8.7]	24.5 [±4.4]	25.9 [±5.1]	27.1 [±3.3]
	% correct of items attempted	72.3 [±3.8]	59.7 ** [±6.7]	72 [±4.6]	65.8 [±6.4]	62.6 [±10.9]	59.4 [±7.9]	64.2 [±8.8]	59.5 [±5.1]
	% of students with zero scores	2.8 [±2.8]	17.8 *** [±8.4]	4.3 [±3.2]	8.6 [±6.2]	7.9 [±6.9]	15.5 [±7.8]	8 [±7.1]	13.2 [±4.9]
Invented words	fluency (correct words per min.)	8.8 [±1.3]	12.9 *** [±1.6]	11 [±1.5]	16.5 *** [±2.8]	6.5 [±2.5]	10.4 * [±1.9]	8.9 [±2.3]	13.5 ** [±1.9]
	% correct of items attempted	44.2 [±5.3]	53.9 * [±5.7]	42.6 [±5.1]	53 * [±6.3]	34.7 [±6.4]	45.9 * [±6.2]	38.1 [±7.4]	47.3 [±5.5]
	% of students with zero scores	21.3 [±6.8]	13.3 [±6.3]	19.3 [±6.0]	8.6 * [±5.2]	23 [±7.3]	12.5 * [±5.0]	24.4 [±9.4]	11.5 * [±5.4]
Oral reading	ORF	17.3 [±2.4]	19.6 [±2.5]	28.8 [±3.1]	34.6 [±6.0]	13.7 [±4.6]	15.2 [±4.0]	22.3 [±6.0]	25.2 [±3.4]
	% correct of items attempted	53.9 [±5.1]	52.3 [±6.5]	62.9 [±4.8]	63.7 [±7.8]	41.3 [±12.0]	40.5 [±7.7]	51.7 [±10.4]	53 [±5.8]
	% of students with zero scores	14.5 [±6.0]	21.3 [±8.9]	10.2 [±4.7]	14.5 [±6.8]	16.3 [±7.1]	31.2 * [±9.9]	16.1 [±9.9]	15.8 [±5.8]
Reading comprehension	% correct of items attempted	40.4 [±7.0]	35.5 [±7.2]	55.6 [±6.4]	48.6 [±7.4]	22.9 [±15.6]	28.9 [±9]	42.1 [±8.6]	48.8 [±8.1]
	% correct	21.1 [±4.6]	25.4 [±5.0]	41.4 [±5.5]	42.2 [±7.4]	13.3 [±9.2]	19.8 [±7.9]	30 [±7.5]	38.2 [±6.5]
	% of students with 80% comp.	6.8 [±3.3]	12.3 [±5.0]	25.3 [±6.8]	23.7 [±9.0]	3 [±3.2]	7.9 [±6.7]	10.7 [±5.7]	20.3 * [±5.5]
	% of students with zero scores	46.6 [±9.5]	50.1 [±8.8]	22.7 [±6.7]	25.4 [±8.9]	65.1 [±22.5]	59 [±10.4]	34.1 [±13.6]	30.8 [±8.5]

* p<.05; ** p<.01; *** p<.001

Subtask	Measure	2017 Tafilah G2	2019 Tafilah G2	2017 Tafilah G3	2019 Tafilah G3	2017 Zarqa G2	2019 Zarqa G2	2017 Zarqa G3	2019 Zarqa G3
Listening comprehension	% correct	55.8 [±5.4]	45.9 ** [±4.7]	63 [±3.1]	58.4 [±4.8]	63.4 [±5.3]	50.7 *** [±5.0]	69.7 [±5.6]	63.6 [±6.0]
	% of students with zero scores	10.2 [±6.0]	13.8 [±6.8]	4.2 [±3.4]	8.7 [±5.1]	2.5 [±1.9]	9.3 [±3.2]	2.5 [±3.0]	5.9 [±5.0]
Letter sound	fluency (correct letters per min.)	45.6 [±7.3]	47.3 [±4]	49 [±6.1]	57.5 * [±4.2]	53.2 [±5]	47.1 [±5.7]	46.9 [±11.4]	44.9 [±6.9]
	% correct of items attempted	71.7 [±10.8]	78.8 [±5.1]	71.6 [±8.7]	83.7 * [±4.1]	85.1 [±4.8]	79.3 [±7.3]	72.9 [±13.4]	72.5 [±7.5]
	% of students with zero scores	12.5 [±8.6]	8.7 [±4.1]	11.7 [±6.9]	4.3 * [±2.8]	2.9 [±3.2]	8 [±5.9]	7.1 [±9.7]	11.2 [±6.5]
Syllable sound	fluency (correct syllables per min.)	34.4 [±3.7]	30.1 [±4.7]	36.5 [±4.0]	35.9 [±3.7]	33.2 [±4.8]	27.1 [±4.1]	32 [±7.4]	30.4 [±4.7]
	% correct of items attempted	78.3 [±4.1]	66.6 *** [±5.7]	75.6 [±4.5]	72.5 [±4.6]	78.7 [±4.2]	67.9 * [±7.7]	70.5 [±11.6]	67.8 [±7.0]
	% of students with zero scores	5.4 [±3.9]	10.7 [±4.7]	6.1 [±3.3]	3.6 [±2.6]	2.7 [±2.7]	11.5 ** [±6.7]	8.1 [±9.7]	10 [±6.8]
Invented words	fluency (correct words per min.)	11.4 [±2.0]	12.2 [±2.2]	13.6 [±2.2]	15.3 [±1.8]	12 [±1.8]	11.7 [±1.8]	13.5 [±2.5]	13.1 [±2.0]
	% correct of items attempted	47.9 [±5.7]	50.1 [±6.5]	47.7 [±6.1]	54.5 [±4.2]	53.9 [±5.2]	54.8 [±7.7]	51.8 [±7.6]	52.6 [±6.9]
	% of students with zero scores	18 [±7.7]	11.3 [±4.5]	20 [±6.3]	4.1 *** [±3.2]	13.3 [±4.1]	17.7 [±8.1]	13.7 [±7.4]	14.3 [±8.0]
Oral reading	ORF	24.2 [±4.4]	20.8 [±4.1]	33.6 [±3.9]	32.5 [±4.1]	21.1 [±3.6]	16.3 * [±2.5]	29.3 [±6.2]	28.5 [±4.4]
	% correct of items attempted	63.3 [±7.0]	51.9 * [±6.9]	66.2 [±5.6]	66.3 [±5.2]	59.9 [±6.5]	49.5 * [±7.9]	63.2 [±11.4]	60.8 [±7.9]
	% of students with zero scores	14.4 [±8.6]	24.6 [±5.5]	11.6 [±5.4]	10.3 [±4.3]	10.5 [±5.1]	29 *** [±10.4]	10.1 [±8.4]	16.3 [±8.5]
Reading comprehension	% correct of items attempted	40.4 [±8.4]	33.3 [±7.0]	58.5 [±5.8]	55.6 [±5.9]	48.7 [±9.9]	38.3 [±7.4]	55.8 [±11.6]	50.9 [±9.1]
	% correct	28.1 [±6.5]	24.4 [±6.0]	49.2 [±6.3]	45.6 [±5.3]	28.2 [±7.5]	21.8 [±3.7]	43.9 [±9.4]	40.3 [±8]
	% of students with 80% comp.	12.4 [±4.7]	13.9 [±5.2]	34.7 [±6.7]	29.5 [±7.0]	13.4 [±5.3]	8.3 [±4.9]	31 [±8.1]	26.3 [±10.1]
	% of students with zero scores	39.9 [±11.5]	50.6 [±9.4]	21 [±6.5]	24.5 [±7.0]	37 [±11.5]	49.1 [±7.3]	27.4 [±13.6]	28.3 [±9.3]

* p<.05; ** p<.01; *** p<.001

Annex 9: EGRA Results by Cohort, Including Zero Scores

Subtask	Measure	Cohort 1 Grade 2 2017	Cohort 1 Grade 2 2019	Cohort 1 Grade 3 2017	Cohort 1 Grade 3 2019
Listening comprehension	% correct	59.1 [±3.7]	54.9 [±3.4]	66.6 [±4.4]	67 [±3.8]
	% of students with zero scores	5.2 [±1.8]	8.3 [±2.2]	3.8 [±2.1]	5.3 [±3.0]
Letter sound	fluency (correct letters per min.)	48.2 [±3.6]	48.8 [±3.6]	46.9 [±7.4]	49.7 [±4.2]
	% correct of items attempted	78.4 [±3.6]	79.4 [±4.5]	71.9 [±8.5]	76.2 [±4.5]
	% of students with zero scores	7.2 [±2.3]	8.4 [±3.5]	8.6 [±5.4]	8.9 [±3.8]
Syllable sound	fluency (correct syllables per min.)	30.8 [±3.3]	32.1 [±2.5]	32.7 [±5.0]	36.5 [±3.2]
	% correct of items attempted	73.9 [±3.5]	72.9 [±4.3]	71.2 [±7.7]	73.8 [±4.2]
	% of students with zero scores	5.2 [±2.4]	8.2 [±3.7]	7 [±6.3]	7.2 [±4]
Invented words	fluency (correct words per min.)	10.8 [±1.3]	14.2 *** [±1.1]	13.4 [±1.7]	16.7 ** [±1.5]
	% correct of items attempted	48.9 [±3.8]	59.4 *** [±4.4]	50.2 [±4.8]	59.1 ** [±4.2]
	% of students with zero scores	17 [±3.3]	12.2 [±4.6]	14.1 [±4.8]	9.4 [±4.5]
Oral reading	oral reading fluency (ORF)	21 [±2.4]	21.1 [±1.8]	30.2 [±4.3]	34.5 [±3.2]
	% correct of items attempted	57.8 [±4.5]	57.2 [±4.6]	64.1 [±7.6]	68.4 [±5.0]
	% of students with zero scores	12.1 [±3.7]	20.7 * [±6.0]	8.7 [±5.7]	10.6 [±5.1]
Reading comprehension	% correct of items attempted	45.4 [±6.5]	46.8 [±4.7]	56.5 [±7.7]	59.9 [±5.9]
	% correct	27.4 [±4.9]	30.5 [±3.1]	45.2 [±6.5]	50.4 [±5.5]
	% of students with 80% comp.	12.1 [±3.5]	14.5 [±3.4]	30.9 [±5.6]	36.1 [±6.8]
	% of students with zero scores	39 [±7.6]	37.4 [±4.9]	24.3 [±9.6]	20.3 [±5.9]

* p<.05; ** p<.01; *** p<.001

Subtask	Measure	Cohort 2 Grade 2 2017	Cohort 2 Grade 2 2019	Cohort 2 Grade 3 2017	Cohort 2 Grade 3 2019
Listening comprehension	% correct	58.4 [±3.6]	53.1 [±4.8]	65.9 [±2.9]	62.6 [±3.8]
	% of students with zero scores	9 [±2.5]	8.2 [±3.2]	1.9 [±1.3]	4.5 * [±2.2]
Letter sound	fluency (correct letters per min.)	47.2 [±2.9]	46.8 [±3.9]	49.4 [±3.3]	52.6 [±3.3]
	% correct of items attempted	75.3 [±2.7]	79 [±3.9]	74.4 [±3.5]	80.2 [±4.0]
	% of students with zero scores	7.5 [±2.8]	7.2 [±3.4]	7.3 [±2.7]	6.6 [±3.5]
Syllable sound	fluency (correct syllables per min.)	32 [±2.7]	30 [±3.1]	36.4 [±2.9]	38.3 [±3.4]
	% correct of items attempted	74.4 [±4.7]	67.8 [±4.9]	75.3 [±3.5]	77.8 [±4.1]
	% of students with zero scores	3.4 [±1.8]	12.9 *** [±4.1]	2.6 [±1.8]	3.2 [±1.6]
Invented words	fluency (correct words per min.)	12 [±1.5]	13.5 [±1.5]	15.1 [±1.4]	17.9 * [±1.7]
	% correct of items attempted	51.5 [±5.3]	55.6 [±4.5]	53.2 [±3.9]	61.9 ** [±3.9]
	% of students with zero scores	14.2 [±4.3]	13.9 [±4.6]	10.4 [±3.3]	4.7 * [±3.2]
Oral reading	ORF	22.4 [±2.1]	20.7 [±2.6]	33.5 [±2.6]	35 [±4.0]
	% correct of items attempted	58.9 [±5.2]	53.5 [±5.2]	67.2 [±4.0]	69.5 [±5.0]
	% of students with zero scores	11.4 [±3.9]	18.7 * [±5.2]	6.3 [±2.7]	6.7 [±3.4]
Reading comprehension	% correct of items attempted	46.2 [±4.4]	41.4 [±5.4]	60.8 [±4.2]	58.8 [±5.7]
	% correct	30.7 [±3.7]	29.5 [±5.0]	51.3 [±4.8]	50.5 [±6.2]
	% of students with 80% comp.	13.1 [±3.5]	14.7 [±5.0]	35.4 [±6.3]	36.1 [±7.3]
	% of students with zero scores	34.3 [±5.7]	42.7 [±6.4]	14 [±3.7]	21.5 * [±6.1]

* p<.05; ** p<.01; *** p<.001

Subtask	Measure	Cohort 3 Grade 2 2017	Cohort 3 Grade 2 2019	Cohort 3 Grade 3 2017	Cohort 3 Grade 3 2019
Listening comprehension	% correct	49.9 [±5.7]	49.5 [±3.4]	56.6 [±3.1]	58.8 [±3.7]
	% of students with zero scores	10.8 [±3.6]	11.5 [±3.0]	7.5 [±3.2]	5.7 [±2.5]
Letter sound	fluency (correct letters per min.)	46.2 [±4.0]	49.8 [±3.2]	49.6 [±3.6]	53.9 [±3.2]
	% correct of items attempted	82.6 [±3.3]	80.4 [±2.8]	78.3 [±3.3]	78.4 [±3.2]
	% of students with zero scores	4.3 [±2.1]	6.9 [±2.1]	7.6 [±2.6]	9.5 [±2.5]
Syllable sound	fluency (correct syllables per min.)	22.8 [±5.0]	26.7 [±2.4]	29.1 [±2.7]	30.2 [±2.4]
	% correct of items attempted	64.9 [±6.3]	62.9 [±4.3]	68.7 [±4.4]	64.1 [±3.8]
	% of students with zero scores	7.7 [±3.9]	13.2 [±3.9]	6.1 [±3.3]	11.2 * [±3.5]
Invented words	fluency (correct words per min.)	7.9 [±1.6]	11.6 *** [±1.0]	10.8 [±1.3]	14.5 *** [±1.2]
	% correct of items attempted	39 [±4.6]	49.7 *** [±3.4]	43.6 [±4.0]	49.9 * [±3.4]
	% of students with zero scores	22.5 [±5.0]	11.9 *** [±3.2]	19.9 [±4.7]	10.7 ** [±3.6]
Oral reading	ORF	16 [±3.0]	17.9 [±2.1]	27.2 [±3.4]	29 [±2.5]
	% correct of items attempted	47.2 [±7.6]	46.7 [±4.2]	59.4 [±5.2]	58.8 [±3.9]
	% of students with zero scores	16.4 [±5.1]	25.3 * [±5.2]	12 [±4.7]	13.8 [±3.6]
Reading comprehension	% correct of items attempted	30.2 [±9.9]	34.2 [±4.7]	49.5 [±5.3]	49.3 [±4.6]
	% correct	17.6 [±6.0]	23.5 [±4.0]	38.2 [±5.1]	40.5 [±4.1]
	% of students with 80% comp.	6.3 [±2.8]	9.5 [±3.4]	22.6 [±5.6]	22.8 [±4.0]
	% of students with zero scores	57.6 [±13.6]	50.5 [±5.8]	28.8 [±7.1]	28 [±5.0]

* p<.05; ** p<.01; *** p<.001

Annex 10: Overall EGMA Results, Including Zero Scores

Subtask	Measure	2014 G2 and G3	2017 G2 and G3	2019 G2 and G3
Number Identification	fluency (correct items per min.)	31.8 [±2.1]	36.8 [±1.2]	40.8 *** [±1.5]
	% correct of items attempted	88.1 [±1.9]	92.6 [±.8]	92.2 *** [±1]
	% of students with zero scores	.3 [±.5]	.1 [±.1]	.1 [±.1]
Quantity Comparison	% correct	78.9 [±2.3]	83.7 [±1.4]	85.1 *** [±1.4]
	% of students with zero scores	2 [±1.1]	.5 [±.3]	.7 *** [±.3]
Addition L1	fluency (correct items per min.)	11.9 [±.5]	12.5 [±.3]	13.1 *** [±.4]
Subtraction L1	fluency (correct items per min.)	9.5 [±.5]	10 [±.3]	10.4 ** [±.3]
Addition and Subtraction L1	% correct	53.1 [±2.1]	55.7 [±1.3]	57.2 ** [±1.6]
	% of students with zero scores	2.3 [±1.2]	1 [±.4]	1.2 [±.6]
Addition and Subtraction L2	% correct	40.4 [±2.9]	45.9 [±2.2]	52.1 *** [±2.5]
	% of students with zero scores	13.5 [±2.8]	7.8 [±1.9]	8 *** [±1.6]
Missing Number	% correct	60.3 [±2.9]	64.5 [±1.9]	64.1 * [±2.2]
	% of students with zero scores	3.1 [±1.5]	1.7 [±.5]	2.5 [±.8]
Word Problems	% correct	57.6 [±2.7]	59.9 [±2.0]	63.6 *** [±2.4]
	% of students with zero scores	7.3 [±2.2]	6.6 [±1.1]	5.9 [±1.1]

* p<.05; ** p<.01; *** p<.001

Annex 11: EGMA Results by Grade, Including Zero Scores

Subtask	Measure	2014 Grade 2	2017 Grade 2	2019 Grade 2	2014 Grade 3	2017 Grade 3	2019 Grade 3
Number Identification	fluency (correct items per min.)	27.5 [±2.2]	32.8 [±1.6]	35.5 *** [±1.5]	36.6 ** [±2.2]	41 *** [±1.7]	45.9 *** [±1.8]
	% correct of items attempted	84.4 [±2.3]	91.1 [±1.2]	89.8 [±1.4]	92.2 [±1.7]	94.1 [±.9]	94.6 * [±1]
	% of students with zero scores	.5 [±.9]	.1 [±.1]	.1 [±.1]	0 [±.1]	.1 [±.1]	.1 [±.2]
Quantity Comparison	% correct	73.5 [±2.8]	80.1 [±2.0]	80.8 *** [±1.8]	84.9 ** [±2.5]	87.4 *** [±1.2]	89.1 ** [±1.6]
	% of students with zero scores	2.7 [±1.8]	.6 [±.4]	1 * [±.6]	1.2 [±.8]	.4 * [±.3]	.4 * [±.3]
Addition L1	fluency (correct items per min.)	11.3 [±.6]	11.7 [±.5]	12 [±.5]	12.7 ** [±.5]	13.2 *** [±.4]	14.1 *** [±.5]
Subtraction L1	fluency (correct items per min.)	9.1 [±.7]	9.5 [±.4]	9.6 [±.4]	10 [±.5]	10.6 *** [±.3]	11.1 *** [±.4]
Addition and Subtraction L1	% correct	50.4 [±2.7]	52.9 [±2.0]	53.2 [±1.9]	56.1 * [±2.2]	58.7 *** [±1.5]	61.0 *** [±1.7]
	% of students with zero scores	2.3 [±1.5]	1.2 [±.6]	1.7 [±.9]	2.2 [±1.3]	.8 * [±.5]	.8 * [±.5]
Addition and Subtraction L2	% correct	35.3 [±3.2]	41.6 [±3.1]	47.3 *** [±3.2]	46 [±3.5]	50.3 [±2.7]	56.7 *** [±2.7]
	% of students with zero scores	16.9 [±3.6]	9.4 [±3.0]	10.4 *** [±2.3]	9.7 [±2.6]	6.1 * [±2.3]	5.6 ** [±1.8]
Missing Number	% correct	54.3 [±3.6]	60.2 [±2.9]	56.7 [±2.6]	66.9 *** [±3.1]	68.9 *** [±2.3]	71.2 * [±2.3]
	% of students with zero scores	4.1 [±2.2]	2.1 [±.8]	3.9 [±1.5]	2 [±1.2]	1.3 *** [±.6]	1.3 [±.7]
Word Problems	% correct	52.6 [±4.1]	53.6 [±2.9]	57.3 [±2.6]	63 *** [±2.7]	66.5 *** [±2.4]	69.7 *** [±2.7]
	% of students with zero scores	9.7 [±3.9]	8.1 [±2.0]	7.9 [±1.6]	4.6 * [±1.9]	5 ** [±1.1]	4.1 [±1.2]

* p<.05; ** p<.01; *** p<.001

Annex 12: EGMA Results by Gender, Including Zero Scores

Subtask	Measure	2014 Male G2 and G3	2014 Female G2 and G3	2017 Male G2 and G3	2017 Female G2 and G3	2019 Male G2 and G3	2019 Female G2 and G3
Number Identification	fluency (correct items per min.)	31.7 [±2.8]	31.9 [±2.1]	37.2 [±1.7]	36.4 [±1.5]	41.8 *** [±2.0]	40 [±1.5]
	% correct of items attempted	86.8 [±2.7]	89.3 [±1.8]	92.3 [±1.0]	92.8 [±.0]	92.2 [±1.5]	92.3 [±1.1]
	% of students with zero scores	.6 [±1.0]	0 [±.1]	.1 * [±.2]	0 [±.1]	.2 [±.2]	.1 [±.1]
Quantity Comparison	% correct	78.9 [±3.3]	78.9 [±2.6]	84 * [±2.2]	83.4 ** [±1.4]	87 * [±1.8]	83.3 [±1.5]
	% of students with zero scores	2 [±1.1]	2 [±1.5]	.7 ** [±.4]	.4 *** [±.3]	.8 [±.6]	.6 [±.4]
Addition L1	fluency (correct items per min.)	12 [±.6]	11.9 [±.6]	12.7 [±.5]	12.3 [±.4]	13.3 [±.6]	12.8 [±.4]
Subtraction L1	fluency (correct items per min.)	9.5 [±.6]	9.5 [±.4]	10.2 [±.4]	9.9 [±.3]	10.7 [±.5]	10.1 [±.3]
Addition and Subtraction L1	% correct	53.1 [±3.0]	53.1 [±2.2]	56.3 [±1.9]	55.2 [±1.4]	58.2 [±2.3]	56.2 [±1.6]
	% of students with zero scores	2.9 [±1.7]	1.7 [±1.7]	1.6 [±.7]	.5 * [±.4]	1.6 [±.8]	.9 [±.6]
Addition and Subtraction L2	% correct	40.3 [±4.2]	40.4 [±3.3]	45.6 * [±3.1]	46.1 ** [±2.7]	53 *** [±3.2]	51.3 ** [±2.8]
	% of students with zero scores	13.9 [±3.5]	13.1 [±4.7]	9.1 * [±3.2]	6.7 ** [±2.2]	8.5 [±2.4]	7.5 [±1.8]
Missing Number	% correct	60.6 [±4.3]	60 [±3.1]	64.8 [±2.7]	64.3 * [±2.2]	66.1 [±3.3]	62.2 [±2.1]
	% of students with zero scores	4.8 [±2.4]	1.6 [±1.3]	1.5 *** [±.6]	1.8 [±.7]	3 * [±1.5]	2.1 [±.9]
Word Problems	% correct	59.1 [±4.7]	56.1 [±2.4]	61.8 [±3.1]	58.4 [±2.4]	66.2 * [±3.1]	61.3 [±2.8]
	% of students with zero scores	7.7 [±3.4]	6.9 [±2.2]	6.9 [±1.6]	6.4 [±1.4]	5.2 [±1.4]	6.6 [±1.7]

Annex 13: EGMA Results by Nationality (Jordanian versus Syrian refugee)

Subtask	Measure	Syrian 2017 G2 and G3	Jordanian 2017 G2 and G3	Syrian 2019 G2 and G3	Jordanian 2019 G2 and G3
Number Identification	fluency (correct items per min.)	35.4 [±3.1]	36.8 [±1.2]	43.9 [±3.9]	40.3 [±1.7]
	% correct of items attempted	92.2 [±1.8]	92.5 [±.8]	93.9 [±2]	92 [±1.1]
	% of students with zero scores	0 [±.]	.1 [±.1]	.1 [±.4]	.1 [±.1]
Quantity Comparison	% correct	86.1 [±3.1]	83.1 [±1.5]	87.8 [±3.5]	84.6 [±1.6]
	% of students with zero scores	.4 [±.6]	.5 [±.3]	.6 [±.7]	.6 [±.4]
Addition L1	fluency (correct items per min.)	12.5 [±.8]	12.4 [±.3]	14.2 * [±1.3]	12.9 [±.5]
Subtraction L1	fluency (correct items per min.)	9.9 [±.5]	10.1 [±.3]	11.1 * [±1.1]	10.3 [±.4]
Addition and Subtraction L1	% correct	55.4 [±2.8]	55.5 [±1.4]	60.9 * [±5.0]	56.6 [±1.7]
	% of students with zero scores	1.5 [±1.8]	1 [±.4]	.9 [±.9]	1.1 [±.6]
Addition and Subtraction L2	% correct	44 [±5.6]	45.9 [±2.3]	56.7 ** [±7.8]	51.3 ** [±2.8]
	% of students with zero scores	9.8 [±6.5]	7.7 [±2.0]	5.6 [±2.7]	8.4 [±1.8]
Missing Number	% correct	62.7 [±4.1]	64.2 [±2.0]	69.9 [±6.3]	63.2 [±2.4]
	% of students with zero scores	2.1 [±2.1]	1.7 [±.5]	2.1 [±1.6]	2.4 [±.9]
Word Problems	% correct	62 [±4.4]	59.6 [±2.3]	68.7 [±7.2]	62.9 [±2.5]
	% of students with zero scores	4.2 [±3.4]	7 [±1.2]	4.5 [±2.7]	6.1 [±1.2]

* p<.05; ** p<.01; *** p<.001

Annex 14: EGMA Results by Governorate, Including Zero Scores

Subtask	Measure	2017 Ajloun G2	2019 Ajloun G2	2017 Ajloun G3	2019 Ajloun G3	2017 Amman G2	2019 Amman G2	2017 Amman G3	2019 Amman G3
Number Identification	fluency (correct items per min.)	34.4 [±2.0]	40.8 ** [±3.8]	44.8 [±3.0]	53.6 *** [±3.6]	32.8 [±4.0]	31.7 [±3.5]	41.3 [±3.5]	43.8 [±4.8]
	% correct of items attempted	93.6 [±2.0]	93.9 [±2.8]	96.6 [±1.3]	98.6 [±.5]	91.9 [±3.1]	87 [±3.2]	94.1 [±1.9]	94.1 [±2.7]
	% of students with zero scores	0 [±.]	0 [±.]	0 [±.]	0 [±.]	0 [±.]	0 [±.]	0 [±.]	0 [±.]
Quantity Comparison	% correct	81.6 [±3.0]	83.9 [±4.1]	88.9 [±3.0]	95.6 *** [±1.4]	83.2 [±3.1]	79 [±3.8]	89.8 [±2.4]	89.3 [±4.4]
	% of students with zero scores	.8 [±1.4]	.5 [±1.8]	0 [±.]	0 [±.]	.5 [±1.7]	.7 [±2.2]	0 [±.]	0 [±.]
Addition L1	fluency (correct items per min.)	11.4 [±.4]	12.7 ** [±.9]	13.6 [±.9]	15.8 *** [±.8]	12.2 [±1.2]	11 [±.8]	13.7 [±.7]	13.5 [±1.1]
Subtraction L1	fluency (correct items per min.)	9.6 [±.4]	10.2 [±.9]	10.8 [±.6]	12.6 *** [±.6]	9.6 [±.8]	8.6 [±.9]	10.8 [±.6]	10.7 [±.8]
Addition and Subtraction L1	% correct	52.5 [±1.9]	56.5 [±4.2]	59.8 [±2.5]	69.4 *** [±2.7]	54.4 [±5.0]	48.8 [±3.9]	60.4 [±2.7]	58.9 [±4.1]
	% of students with zero scores	1.2 [±1.5]	1.6 [±4.2]	.4 [±1.4]	0 [±.]	.9 [±1.8]	.9 [±1.7]	0 [±.]	.6 [±2.1]
Addition and Subtraction L2	% correct	47 [±5.1]	50.1 [±6.1]	53.4 [±4.4]	65.6 *** [±5.3]	41.6 [±5.6]	40.9 [±7.5]	51 [±5.1]	52.9 [±6.5]
	% of students with zero scores	8.7 [±3.4]	6.6 [±4.5]	3.4 [±2.7]	.9 [±1.3]	8.1 [±4.5]	13.1 [±6.1]	4.6 [±3.5]	6.9 [±5.1]
Missing Number	% correct	61.8 [±3.5]	62.4 [±5.3]	74.1 [±3.3]	81.7 ** [±3.6]	64.4 [±6.0]	50.2 *** [±5.6]	72.7 [±4.3]	71.8 [±6.1]
	% of students with zero scores	1.2 [±1.5]	2.3 [±3.1]	.4 [±1.3]	.4 [±1.3]	1.8 [±2.6]	5.1 [±3.6]	.8 [±1.4]	.7 [±2.3]
Word Problems	% correct	56.9 [±4.5]	67.4 * [±7.3]	66.6 [±3.8]	83.9 *** [±4.4]	49.8 [±6.2]	50.3 [±5.6]	67 [±5.9]	68.8 [±6.7]
	% of students with zero scores	8 [±3.1]	6.1 [±6.6]	4.6 [±2.7]	.3 *** [±.0]	8.5 [±5.6]	9.4 [±3.8]	6.2 [±2.8]	4.3 [±2.9]

* p<.05; ** p<.01; *** p<.001

Subtask	Measure	2017 Aqaba G2	2019 Aqaba G2	2017 Aqaba G3	2019 Aqaba G3	2017 Balqa G2	2019 Balqa G2	2017 Balqa G3	2019 Balqa G3
Number Identification	fluency (correct items per min.)	26.2 [±2.1]	29.5 [±3.2]	34.8 [±1.9]	39.9 * [±3.6]	29.1 [±3.8]	37.7 ** [±3.8]	36.9 [±4.1]	43.8 * [±4.7]
	% correct of items attempted	82.7 [±3.3]	84.7 [±3.9]	90.6 [±1.8]	92.7 [±2.6]	87 [±4.1]	91.8 [±3.0]	93.9 [±2.6]	92.4 [±4.0]
	% of students with zero scores	0 [±.]	0 [±.]	.2 [±.4]	0 * [±.]	0 [±.]	.8 [±1.4]	0 [±.]	.6 [±1.9]
Quantity Comparison	% correct	72 [±3.5]	77.5 * [±3.7]	80.3 [±2.9]	87.6 *** [±3.1]	79.7 [±4.8]	81.1 [±4.3]	86.3 [±3.3]	84.9 [±5.4]
	% of students with zero scores	.7 [±2.2]	1.1 [±3.4]	.9 [±1.2]	0 ** [±.]	.5 [±1.6]	1.2 [±1.5]	0 [±.]	2.3 [±4.4]
Addition L1	fluency (correct items per min.)	10.1 [±.6]	11 [±1.1]	11.9 [±.5]	13.7 ** [±1.1]	11.1 [±.0]	12.5 [±1.1]	12.6 [±1.1]	13.6 [±1.4]
Subtraction L1	fluency (correct items per min.)	8.1 [±.6]	9 [±1.0]	9.9 [±.6]	10.9 [±1.1]	9 [±.9]	10.1 [±.9]	10.2 [±.9]	10.5 [±1.1]
Addition and Subtraction L1	% correct	45.5 [±2.7]	48.4 [±4.3]	54 [±2.5]	59.3 [±3.8]	50.3 [±4.4]	55.8 [±4.3]	56.3 [±4.5]	58.3 [±5.3]
	% of students with zero scores	2.9 [±2.9]	2.4 [±3.3]	2 [±1.9]	0 *** [±.]	3.7 [±3.0]	1.1 [±2.1]	2.3 [±2.4]	1.2 [±2.1]
Addition and Subtraction L2	% correct	38.4 [±4.5]	39.7 [±4.9]	45.5 [±4.4]	54.8 ** [±4.9]	45.9 [±6.4]	46.8 [±6.9]	55 [±6.0]	52.3 [±6.9]
	% of students with zero scores	15.7 [±5.0]	18.4 [±6.6]	7 [±3.2]	6.7 [±4.3]	5.8 [±3.7]	8.9 [±5.2]	6 [±4.1]	6.5 [±7.6]
Missing Number	% correct	47.1 [±3.7]	49.4 [±5.0]	60.6 [±3.4]	67.1 * [±4.5]	55.2 [±6.8]	60.3 [±6.1]	64.8 [±6.1]	68 [±7.4]
	% of students with zero scores	3.1 [±2.6]	5.4 [±3.7]	1.6 [±2.0]	.3 [±.8]	3.6 [±2.9]	2 [±2.3]	1 [±1.7]	4.1 [±6.8]
Word Problems	% correct	50.7 [±4.7]	54.4 [±5.5]	57.8 [±3.9]	71.4 *** [±4.1]	56.4 [±5.5]	62.6 [±5.3]	67.6 [±4.8]	64.6 [±6.8]
	% of students with zero scores	13.5 [±4.9]	9.3 [±4.6]	3.8 [±2.1]	2.2 [±2]	6.8 [±3.9]	4 [±3.2]	4.6 [±3.6]	7.7 [±6.7]

* p<.05; ** p<.01; *** p<.001

Subtask	Measure	2017 Irbid G2	2019 Irbid G2	2017 Irbid G3	2019 Irbid G3	2017 Jarash G2	2019 Jarash G2	2017 Jarash G3	2019 Jarash G3
Number Identification	fluency (correct items per min.)	37.2 [±3.0]	42.9 * [±4.2]	46.7 [±3.4]	54.3 ** [±3.4]	35.3 [±2.4]	45.7 *** [±2.9]	43.2 [±2.1]	56.7 *** [±3.5]
	% correct of items attempted	93.4 [±1.9]	93.9 [±4.3]	96.1 [±2.6]	97.8 [±.0]	92.4 [±1.9]	95.4 [±1.6]	96.2 [±.0]	97.9 * [±1.1]
	% of students with zero scores	0 [±.]	0 [±.]	0 [±.]	0 [±.]	0 [±.]	0 [±.]	0 [±.]	0 [±.]
Quantity Comparison	% correct	84.8 [±2.4]	87.4 [±5.8]	89 [±2.9]	94.3 ** [±1.6]	81.3 [±2.9]	86.9 ** [±3.0]	87.8 [±2.3]	92.9 ** [±2.6]
	% of students with zero scores	0 [±.]	1.3 [±1.9]	0 [±.]	0 [±.]	0 [±.]	0 [±.]	.5 [±1.6]	0 [±.]
Addition L1	fluency (correct items per min.)	12.3 [±.8]	13.9 [±1.8]	14.1 [±.7]	16.2 ** [±1.1]	12.3 [±.9]	13.9 ** [±.0]	14.9 [±1]	16.2 [±1.2]
Subtraction L1	fluency (correct items per min.)	10 [±.6]	11.7 * [±1.3]	11 [±.7]	12.8 ** [±1.1]	10 [±.7]	10.8 [±.8]	11.2 [±.8]	12.5 * [±.9]
Addition and Subtraction L1	% correct	55.3 [±3.0]	62.1 [±6.0]	61.6 [±3.2]	69.1 ** [±3.5]	55 [±3.5]	60.4 * [±3.3]	63.3 [±3.2]	67.8 [±3.6]
	% of students with zero scores	.4 [±1.3]	2.6 [±5.3]	.6 [±2.2]	0 [±.]	0 [±.]	0 [±.]	.5 [±1.6]	.2 [±.8]
Addition and Subtraction L2	% correct	45 [±5.0]	62.8 *** [±8.8]	55.4 [±6.1]	71 *** [±4.6]	43.9 [±4.9]	55.5 * [±7.3]	53.1 [±4.7]	62.6 ** [±4.8]
	% of students with zero scores	4.8 [±4.1]	8.1 [±6.1]	3.6 [±5.1]	1 [±2.0]	8.1 [±3.7]	4.3 [±4.1]	3.8 [±2.9]	2.9 [±3.2]
Missing Number	% correct	64.5 [±5.3]	69.2 [±8.8]	71.3 [±4.2]	81.4 *** [±3.9]	58.6 [±5.1]	64.5 [±5.0]	69.4 [±3.7]	77.4 ** [±4.2]
	% of students with zero scores	.5 [±1.6]	4.3 * [±5.3]	1.7 [±2.1]	.4 [±1.2]	1.3 [±2.9]	1.5 [±2.1]	.9 [±1.6]	.7 [±1.4]
Word Problems	% correct	58.2 [±6.8]	70 * [±8.4]	73.3 [±4.4]	77.9 [±7.3]	55.2 [±4]	61.4 [±8.6]	67.6 [±4.0]	73.9 * [±4.6]
	% of students with zero scores	6.8 [±4.6]	4.6 [±4.1]	3.1 [±2.5]	1.3 [±1.5]	4.1 [±3.4]	8 [±7.0]	3.8 [±3.2]	1.8 [±2.3]

* p<.05; ** p<.01; *** p<.001

Subtask	Measure	2017 Karak G2	2019 Karak G2	2017 Karak G3	2019 Karak G3	2017 Ma'an G2	2019 Ma'an G2	2017 Ma'an G3	2019 Ma'an G3
Number Identification	fluency (correct items per min.)	28.8 [±4.6]	38.6 *** [±3.4]	39.7 [±3.9]	47.6 * [±5.7]	26.8 [±4.0]	30.8 [±3.7]	35.8 [±5.2]	42.7 * [±3.8]
	% correct of items attempted	82.5 [±6]	94.3 [±2.5]	93.5 [±2.6]	96.1 [±2.9]	84.8 [±5.4]	84.4 [±4.5]	90.9 [±3.2]	92.2 [±2.1]
	% of students with zero scores	.4 [±1.5]	0 [±.]	.6 [±1.9]	0 [±.]	.9 [±1.8]	1.3 [±2.5]	0 [±.]	0 [±.]
Quantity Comparison	% correct	68.1 [±6.7]	87.4 *** [±3.3]	83.4 [±4.1]	92.6 ** [±4.2]	64.2 [±7.3]	73.2 [±6.5]	77.6 [±5.6]	84.4 * [±3.2]
	% of students with zero scores	2.2 [±2.0]	.8 [±2.7]	2 [±2.6]	0 [±.]	6.9 [±7.0]	2.1 [±2.3]	1.2 [±1.7]	.5 [±1.5]
Addition L1	fluency (correct items per min.)	10.5 [±1.1]	12.8 ** [±1.0]	13.1 [±.0]	15.6 * [±1.6]	9.9 [±1.7]	10.5 [±1.2]	12.9 [±2.2]	12.7 [±1.3]
Subtraction L1	fluency (correct items per min.)	8.4 [±1.2]	10.8 ** [±.9]	10.7 [±.8]	12 [±1.4]	8.6 [±1.3]	8.4 [±.0]	10.1 [±1.1]	10.5 [±1.1]
Addition and Subtraction L1	% correct	47.2 [±5.8]	58.6 ** [±4.7]	58.4 [±3.6]	66.3 * [±6.6]	44.8 [±6.3]	46.8 [±4.9]	55.3 [±5.6]	55.7 [±4.6]
	% of students with zero scores	2.9 [±3.3]	1 [±1.4]	1.4 [±2.5]	3.1 [±6.2]	5.4 [±7.6]	6.1 [±4.3]	1.1 [±2.0]	1.3 [±2.0]
Addition and Subtraction L2	% correct	36.9 [±7.5]	59.8 *** [±7.1]	51.7 [±5.2]	67.3 ** [±9.2]	33.4 [±7.7]	40.5 [±6.1]	45.2 [±7.5]	50 [±5.1]
	% of students with zero scores	18 [±8.9]	5 * [±5.3]	6 [±3.0]	5.4 [±8.0]	24.6 [±15.1]	17.5 [±7.3]	7.6 [±5.6]	9.3 [±5.9]
Missing Number	% correct	53.1 [±7.0]	69.4 *** [±5.5]	66.1 [±5.3]	78.4 ** [±7.2]	50.1 [±7.0]	47.6 [±7.1]	60.5 [±7.2]	63.2 [±5.2]
	% of students with zero scores	5.1 [±3.2]	0 ** [±.]	2.3 [±2.5]	0 [±.]	7.6 [±6.8]	3.1 [±4.3]	2.4 [±2.0]	2 [±2.5]
Word Problems	% correct	50.6 [±5.2]	68.9 *** [±5.3]	63.1 [±5.4]	81.6 *** [±7.2]	41.2 [±6.5]	51.2 [±7.8]	55.4 [±7.1]	60.2 [±6.6]
	% of students with zero scores	10.6 [±4.4]	5.7 [±3.4]	6.3 [±3.5]	3.3 [±4.8]	15.1 [±6.0]	13.9 [±6.4]	8.5 [±4.8]	12.4 [±6.5]

* p<.05; ** p<.01; *** p<.001

Subtask	Measure	2017 Madaba G2	2019 Madaba G2	2017 Madaba G3	2019 Madaba G3	2017 Mafraq G2	2019 Mafraq G2	2017 Mafraq G3	2019 Mafraq G3
Number Identification	fluency (correct items per min.)	32.4 [±2.9]	37.7 * [±2.4]	40.9 [±2.7]	46.5 * [±4.1]	31.3 [±6.0]	31.3 [±3.3]	39 [±4.2]	41.7 [±2.9]
	% correct of items attempted	91.6 [±2.2]	91.9 [±1.8]	94 [±1.9]	94.7 [±2.1]	90.6 [±1.6]	86.9 [±2.4]	90.9 [±2.9]	91.4 [±2.3]
	% of students with zero scores	0 [±.]	.3 [±1.2]	.7 [±2.1]	0 [±.]	0 [±.]	.4 [±1.4]	0 [±.]	.5 [±1.8]
Quantity Comparison	% correct	82 [±3.1]	82.4 [±3.4]	86.6 [±2.7]	85.2 [±4.7]	68.6 [±9.5]	73.4 [±5.3]	79.9 [±6]	84.2 [±2.9]
	% of students with zero scores	0 [±.]	0 [±.]	2 [±3.0]	2.4 [±4.9]	.9 [±.8]	1.4 [±2.8]	1.4 [±4.1]	.8 [±1.4]
Addition L1	fluency (correct items per min.)	11.3 [±.8]	12.4 [±1]	13.1 [±1.1]	13.5 [±1.2]	9.8 [±1.3]	10.5 [±.7]	11.7 [±1.3]	12.3 [±.8]
Subtraction L1	fluency (correct items per min.)	9.3 [±.5]	10 [±.7]	10.4 [±.6]	10.6 [±1.0]	8.4 [±1.3]	8.5 [±.6]	9.8 [±.0]	9.6 [±.7]
Addition and Subtraction L1	% correct	51.4 [±2.9]	54.6 [±3.6]	57.4 [±3.3]	58.9 [±5.3]	45.4 [±6.3]	47.2 [±3.1]	52.6 [±4.8]	53.8 [±3.2]
	% of students with zero scores	1.2 [±2.2]	1.5 [±2.1]	2.3 [±2.2]	3.4 [±4.3]	1.2 [±1.8]	1.4 [±2.8]	3.1 [±4.3]	.8 [±1.6]
Addition and Subtraction L2	% correct	40.6 [±3.7]	49.5 ** [±5.0]	50.1 [±5.1]	52.8 [±7.6]	27.3 [±12.3]	36.4 [±4.3]	44.2 [±6.1]	45 [±5.8]
	% of students with zero scores	8.4 [±3.3]	6.7 [±3.6]	7.3 [±3.9]	8.4 [±5.5]	22 [±17.4]	11.9 [±3.3]	9.8 [±5.5]	9 [±4.7]
Missing Number	% correct	54.2 [±4.5]	55.2 [±4.0]	63.6 [±4.4]	66 [±6.2]	46.7 [±8.3]	49.4 [±5.0]	61.3 [±7.0]	59.4 [±3.8]
	% of students with zero scores	3.4 [±2.9]	3.7 [±2.5]	3.7 [±3.5]	3 [±4.6]	2.9 [±2.3]	4.8 [±4.4]	4.1 [±3.5]	2.3 [±2.2]
Word Problems	% correct	49.7 [±5.3]	57 [±5.7]	60.7 [±5.3]	60 [±7.6]	40.1 [±9]	48.1 [±5.3]	55.5 [±6.0]	60.4 [±3.7]
	% of students with zero scores	10 [±4.6]	10.4 [±4.5]	7.3 [±3.7]	7.6 [±5.6]	15.4 [±5.2]	10.9 [±4.9]	13.9 [±4.1]	4.2 *** [±3.0]

* p<.05; ** p<.01; *** p<.001

Subtask	Measure	2017 Tafilah G2	2019 Tafilah G2	2017 Tafilah G3	2019 Tafilah G3	2017 Zarqa G2	2019 Zarqa G2	2017 Zarqa G3	2019 Zarqa G3
Number Identification	fluency (correct items per min.)	35.3 [±2.8]	32.5 [±2.7]	41.8 [±3.2]	43.6 [±3.4]	31.9 [±3.0]	32.4 [±2.2]	38 [±4.5]	41.7 [±3.3]
	% correct of items attempted	92.7 [±1.8]	86.6 [±3.2]	94 [±2.2]	93.3 [±2.3]	92.1 [±2.7]	89.3 [±3.2]	94 [±1.9]	93.4 [±2.4]
	% of students with zero scores	.2 [±.7]	0 [±.]	0 [±.]	0 [±.]	0 [±.]	0 [±.]	0 [±.]	0 [±.]
Quantity Comparison	% correct	77.1 [±4.1]	77.7 [±4.5]	85 [±4.1]	85.7 [±3.3]	83.5 [±3.1]	78.1 * [±2.6]	89.1 [±2.6]	86.8 [±3.8]
	% of students with zero scores	0 [±.]	1 [±1.8]	1.1 [±2.1]	0 [±.]	0 [±.]	1 [±1.7]	0 [±.]	.8 [±1.5]
Addition L1	fluency (correct items per min.)	11.4 [±.8]	10.5 [±.7]	13.3 [±1.4]	13.1 [±.8]	12.3 [±.9]	11.7 [±.8]	12.7 [±.9]	13.6 [±1.3]
Subtraction L1	fluency (correct items per min.)	9.5 [±.8]	8.4 [±.6]	10.7 * [±.7]	10 [±.8]	10.2 [±.7]	9.1 * [±.7]	10.2 [±.8]	10.8 [±1.2]
Addition and Subtraction L1	% correct	51.8 [±3.5]	46.8 [±3.1]	58.2 * [±3.8]	56.7 [±3.3]	56.1 [±3.9]	51.5 [±2.9]	57 [±3.9]	58.7 [±4.7]
	% of students with zero scores	4.1 [±4.2]	5.4 [±3.0]	2.8 [±2.3]	.8 [±1.4]	.5 [±1.8]	1.3 [±1.7]	0 [±.]	.8 [±1.5]
Addition and Subtraction L2	% correct	38 [±5.1]	38.9 [±5.4]	47.8 [±4.8]	49.8 [±5.8]	47.2 [±6.9]	43.3 [±5.2]	46.7 [±8.0]	53.3 [±6.8]
	% of students with zero scores	12.7 [±5.9]	16.8 [±5.5]	8.6 [±4.5]	8.5 [±5.4]	4.7 [±4.3]	9 [±4.3]	8.9 [±8.7]	6 [±4.5]
Missing Number	% correct	60.5 [±4.9]	55.3 [±4.9]	68.2 [±4.3]	67.2 [±4.0]	63.5 [±5.1]	52 *** [±3.5]	68.6 [±7.9]	64.4 [±5.8]
	% of students with zero scores	3.9 [±4.3]	3.6 [±4.2]	1.5 [±1.8]	.6 [±2.0]	1.2 [±1.7]	3.5 [±2.4]	0 [±.]	2 * [±2.4]
Word Problems	% correct	47.4 [±5.9]	51.9 [±5.1]	62 [±6.0]	65.4 [±4.5]	65.1 [±6.4]	53.9 ** [±3.7]	68.2 [±7.0]	65.8 [±5.9]
	% of students with zero scores	14.8 [±7.8]	10.8 [±4.8]	8.1 [±4.0]	4.9 [±2.4]	3 [±3.5]	7.7 [±3.4]	.8 [±1.6]	4.7 * [±3.9]

* p<.05; ** p<.01; *** p<.001

Annex 15: EGMA Results by Cohort, Including Zero Scores

Subtask	Measure	Cohort 1 Grade 2 2017	Cohort 1 Grade 2 2019	Cohort 1 Grade 3 2017	Cohort 1 Grade 3 2019
Number Identification	fluency (correct items per min.)	31.9 [±2.0]	36.2 *** [±1.7]	39.3 [±3.2]	45.9 *** [±2.4]
	% correct of items attempted	90.4 [±2.1]	91.6 [±1.9]	94.3 [±1.4]	95 [±1.5]
	% of students with zero scores	.1 [±.3]	0 [±.]	.1 [±.4]	0 [±.]
Quantity Comparison	% correct	80.1 [±2.4]	81.8 [±1.8]	87.9 [±1.9]	89.6 [±2.4]
	% of students with zero scores	.5 [±.4]	.8 [±.0]	.4 [±.5]	.5 [±.8]
Addition L1	fluency (correct items per min.)	11.9 [±.6]	12.3 [±.5]	13 [±.7]	14.5 ** [±.9]
Subtraction L1	fluency (correct items per min.)	9.8 [±.5]	9.8 [±.4]	10.4 [±.6]	11.4 * [±.8]
Addition and Subtraction L1	% correct	54 [±2.7]	54.6 [±2]	58.1 [±2.8]	62.4 * [±3.1]
	% of students with zero scores	1 [±.9]	1.1 [±.0]	.3 [±.5]	1.2 [±1.4]
Addition and Subtraction L2	% correct	44.9 [±4.6]	49.1 [±3.5]	48.8 [±5.7]	58.5 ** [±4.4]
	% of students with zero scores	7.9 [±3.0]	7.3 [±2.7]	7.5 [±5.8]	5 [±2.9]
Missing Number	% correct	60.8 [±3.6]	58.3 [±2.5]	68.6 [±5.1]	70.6 [±3.7]
	% of students with zero scores	2 [±1.1]	2.4 [±1.4]	.5 [±.5]	1.3 [±1.3]
Word Problems	% correct	60.5 [±4.2]	59.3 [±2.9]	67.1 [±4.4]	71.8 [±3.7]
	% of students with zero scores	5 [±2.2]	7.1 [±2.2]	2.4 [±1.2]	3.7 [±2.3]

* p<.05; ** p<.01; *** p<.001

Subtask	Measure	Cohort 2 Grade 2 2017	Cohort 2 Grade 2 2019	Cohort 2 Grade 3 2017	Cohort 2 Grade 3 2019
Number Identification	fluency (correct items per min.)	34.4 [±2.4]	35.7 [±2.6]	43.2 [±2.3]	47.1 * [±3.2]
	% correct of items attempted	92.4 [±1.9]	89.4 [±2.4]	94.7 [±1.4]	95.3 [±1.6]
	% of students with zero scores	0 [±.0]	0 [±.]	0 [±.0]	0 [±.]
Quantity Comparison	% correct	83.4 [±2.0]	81.9 [±3.0]	89.1 [±1.7]	90.7 [±2.6]
	% of students with zero scores	.3 [±.9]	.9 [±1.0]	.1 [±.1]	0 * [±.]
Addition L1	fluency (correct items per min.)	12.2 [±.8]	12 [±.8]	13.8 [±.5]	14.4 [±.8]
Subtraction L1	fluency (correct items per min.)	9.7 [±.5]	9.7 [±.7]	10.8 [±.4]	11.4 [±.6]
Addition and Subtraction L1	% correct	54.5 [±3]	53.5 [±3.3]	60.6 [±1.9]	62.3 [±2.8]
	% of students with zero scores	.9 [±.9]	1.8 [±1.8]	.4 [±.6]	.4 [±1.1]
Addition and Subtraction L2	% correct	42.7 [±3.6]	48.7 [±5.7]	52.4 [±3.6]	59 * [±4.4]
	% of students with zero scores	7.2 [±2.8]	11.6 [±4]	4.5 [±2.5]	4.9 [±3]
Missing Number	% correct	64 [±3.9]	57.3 * [±4.7]	71.7 [±2.9]	74.6 [±3.9]
	% of students with zero scores	1.4 [±1.4]	4.8 * [±2.6]	1.2 [±.0]	.6 [±1.1]
Word Problems	% correct	52.8 [±4.2]	57.7 [±4.6]	68.9 [±3.6]	71.9 [±4.7]
	% of students with zero scores	8.2 [±3.5]	7.8 [±2.5]	5 [±1.8]	3.2 [±1.8]

* p<.05; ** p<.01; *** p<.001

Subtask	Measure	Cohort 3 Grade 2 2017	Cohort 3 Grade 2 2019	Cohort 3 Grade 3 2017	Cohort 3 Grade 3 2019
Number Identification	fluency (correct items per min.)	30.2 [±2.9]	34 * [±1.9]	38.1 [±2.3]	43 ** [±2.0]
	% correct of items attempted	89 [±1.6]	88.6 [±1.5]	92 [±1.6]	92.2 [±1.6]
	% of students with zero scores	.1 [±.3]	.6 [±.6]	.1 [±.3]	.4 [±.8]
Quantity Comparison	% correct	72.1 [±5.7]	76.8 [±2.8]	82 [±3.0]	84.6 [±2.2]
	% of students with zero scores	1.7 [±1.3]	1.3 [±1.1]	1.1 [±1.5]	1.4 [±1.4]
Addition L1	fluency (correct items per min.)	10.3 [±.8]	11.3 * [±.5]	12.3 [±.8]	12.9 [±.6]
Subtraction L1	fluency (correct items per min.)	8.7 [±.8]	9.2 [±.4]	10 [±.5]	10.1 [±.5]
Addition and Subtraction L1	% correct	47.2 [±3.8]	50.6 [±2.1]	54.6 [±2.7]	56.1 [±2.3]
	% of students with zero scores	2.4 [±1.7]	2 [±1.3]	2.5 [±1.9]	1.3 [±.9]
Addition and Subtraction L2	% correct	34.3 [±7.9]	41.7 [±3.0]	47.8 [±3.5]	48.9 [±3.5]
	% of students with zero scores	16.9 [±10.1]	11.1 [±2.4]	8.1 [±2.8]	8.2 [±3.1]
Missing Number	% correct	50.1 [±5.2]	53.1 [±3.0]	62.3 [±3.7]	63.4 [±3.0]
	% of students with zero scores	3.9 [±1.8]	3.6 [±2.0]	2.9 [±1.6]	2.9 [±2.1]
Word Problems	% correct	45.3 [±5.8]	53.9 ** [±3.1]	59.2 [±3.2]	61.6 [±2.9]
	% of students with zero scores	12.6 [±3.1]	9.2 [±2.6]	9.8 [±2.4]	6.8 [±2.6]

* p<.05; ** p<.01; *** p<.001

Annex 16: Instruments

Jordan 2019 EGRA Assessor Protocol

Form __

أداة تقييم مهارات القراءة في المرحلة الأساسية: نموذج تعليمات المقيّم 2019

تعليمات عامة

من المهم أن تضفي جوًّا من المرح على الطفل الذي سيخضع للتقييم كأن تبدأ معه بمحادثة بسيطة حول مواضيع تهمه (انظر المثال أدناه).أشعره بأن هذا التقييم هو تقريبًا بمثابة لعبة سيستمتع بها، وليس بالمهمة الصعبة.
من المهم جدًّا أن تقرأ محتوى المربعات **فقط**، بصوت عالٍ وبوضوح وتمهّل.

صباح الخير. اسمي_____ أسكن في_____. أريد أن أتكلم معك عن نفسي، لدي ..... من الأطفال، عمرهم .....؛ عندي في البيت.....، الرياضة التي أمارسها .....، إلخ.]
1. أخبرني عن نفسك وعن عائلتك؟ [انتظر الجواب؛ إذا كان التلميذ غير متحمس للكلام، وجّه إليه السؤال رقم 2. إذا تكلم بارتياح، انتقل لفقرة الموافقة الشفهية.]
2. ما اللعبة التي تحبها؟

الموافقة الشفهية

اسمح لي أن أقول لك لماذا أنا معك اليوم. أنا أعمل في وزارة التربية والتعليم، وأحاول أن أفهم كيف يتعلم الأطفال القراءة. لقد تم اختيارك للقيام بهذا الاختبار بشكل عشوائي.
أُحبّذ أن تتعاون معي في هذه العملية. ولكن إذا لم ترد المشاركة، فلك ذلك.
سنلعب لعبة القراءة؛ إذ سأطلب منك أن تقرأ بعض الحروف وبعض الكلمات وقصة قصيرة بصوت عالٍ.
سأستعمل ساعة أو مؤقِّتًا لأحسب الوقت الذي تحتاجه في القراءة.
هذا ليس امتحانًا، وليس له أي تأثير على علاماتك المدرسية.
سأسألك بعض الأسئلة الأخرى عن عائلتك.
لن أكتب اسمك على ورقة الاختبار. لن يرى أي أحد إجاباتك عليها.
مرة أخرى، أنت غير ملزم بالمشاركة إذا لم تكن ترغب في ذلك، وإذا بدأنا ولم ترد الإجابة عن أي سؤال، فلا مشكلة في ذلك.
هل لديك سؤال؟هل أنت مستعد؟

إذا حصلت على الموافقة الشفهية للطفل ضع علامة (×)في هذا المربع ☐ نعم
(إذا لم تحصل على الموافقة، اشكر الطفل وانتقل إلى الذي يليه واستعمل نفس الاستمارة)

1. تاريخ التقييم مثال: 15 يوليو 2019 = 2019-04-15	اليوم: __ الشهر: __ السنة: __
2. المحافظة:	
3. مديرية التربية والتعليم:	
4. اسم المدرسة:	
5. الرقم الوطني للمدرسة:	
6 فترة دوام الطفل	○ فترة واحدة ○ فترة صباحية ○ فترة مسائية
7. اسم المقيّم:	
8. رمز المقيّم: (ذاتي)	
9. الصف:	○ الثاني (2) ○ الثالث (3)
10. الشعبة:	
11. رقم الطفل:	
12. تاريخ ميلاد الطفل:	الشهر: __ السنة: __
13. جنس الطفل:	○ ذكر ○ أنثى
14. الطفل مصاب بإعاقة:	_____: ☐ نعم (اختر واحدة فيما يأتي) ☐ سمعية

60 ثانية	القسم 1: التعرف على أصوات الحروف	1

بعد مرور 60 ثانية، ستقول للطفل**'توقف'**.

إذا تردد الطفل في قراءة الحرف لمدة تزيد عن 3 ثوانٍ، أشر للحرف الذي يليه وقل : **"لنكمل من فضلك"**.

قاعدة التوقف المبكّر: إذا وضعت علامة (/) على جميع الإجابات في السطر الأول على أنها خطأ ولم يصحح الطفل أي خطأ من أخطائه، قل **"شكرا"** وأوقف التمرين. ضع علامة (×) في المربع الموجود في أسفل الصفحة وانتقل للتمرين الذي يليه.

🗣 **هذه ورقة تضم حروفًا، اقرأ قدر ما تستطيع منها (اقرأ صوت الحرف وليس اسمه).**
مثلا، صوت هذا الحرف [*أشر إلى الحرف "كـ"***] هو" كـ ".**

🗣 **والآن لنقم بهذا التمرين: قل لي صوت هذا الحرف [***وأشر إلى الحرف "لـ"***]:**
🗣✓**: جيد، صوت هذا الحرف هو"لـ"**
🗣×**: صوت هذا الحرف هو"لام"**

🗣 **لنجرب مثالاً آخر: قل لي صوت هذا الحرف [***أشر إلى الفتحة "*ص*"***]:**
🗣✓**: أحسنت، صوت هذا الحرف هو"**ص**"**
🗣×**: صوت هذا الحرف هو"**صاد**"**

🗣 **هل فهمت المطلوب منك؟**
عندما أقول لك "لنبدأ"، اقرأ صوت الحروف بدقة وبأكبر سرعة ممكنة. سنبدأ من هنا ونكمل بهذه الطريقة [*أشر إلى الحرف الأول في السطر الأول، وتتبع معه بإصبعك على الحروف الموجودة في السطر الأول بأكمله***]. هل أنت مستعد؟ لنبدأ**

✍ضع **بوضوح** علامة (/) على أي خطأ يرتكبه الطفل.
✍في حالة قيام الطفل بتصحيح نفسه، قم بوضع دائرة ⬭ حول علامة (/) التي وضعتها مسبقاً له.
✍ضع العلامة (**]**) على آخر حرف قرأه الطفل.

مثال: كـ لـ ص

	10	9	8	7	6	5	4	3	2	1
(10)	ك	ص	ظ	ز	ـهـ	خـ	د	ل	و	جـ
(20)	س	و	ة	د	ر	قـ	خ	فـ	ح	ا
(30)	هـ	ص	ظ	ش	ج	تـ	ف	حـ	بـ	نـ
(40)	ع	ـه	ـغـ	ـة	ز	و	ب	خ	ق	ضـ
(50)	ق	ثـ	ب	ضـ	يـ	ح	ـهـ	ذ	غـ	خـ
(60)	ظ	و	ص	ـهـ	ط	نـ	ي	سـ	ذ	ـعـ
(70)	ن	خـ	بـ	ق	غ	ي	شـ	د	ء	تـ
(80)	فـ	ن	ض	سـ	ح	ث	ط	ب	م	ذ
(90)	ج	ط	حـ	خ	ضـ	ـه	ـعـ	ش	ث	غ
(100)	ث	لـ	عـ	ف	مـ	خ	ذ	ء	س	ز

✍ الوقت المتبقي من وقت التمرين (عدد الثواني):

✍ ضع علامة (×) في هذا المربع □ في حال أوقفت هذا الجزء من التقييم؛ لأن الطفل لم يقرأ أيًّا من الكلمات في السطر الأول بشكل صحيح.

2 القسم 2: قراءة المقاطع	60 ثانية

هذه ورقة تضم مقاطعَ، اقرأ قدر ماتستطيع منها (اقرأ المقطع).
مثلا، نقرأ هذا المقطع [*أشر إلى المقطع***" را"]**

و الآن لنقم بهذا التمرين: اقرأ هذا المقطع [*وأشر إلى المقطع***" أُع"]:**
✓: جيد، نقرأ هذا المقطع هكذا " أُع "
×: نقرأ هذا المقطع" أُع "

لنجرب مثالاً آخر: اقرأ لي هذا المقطع [*أشر إلى المقطع*سى]:
✓: أحسنت، نقرأ هذا المقطع هو "سى"
×: نقرأ هذا المقطع هكذا "سى"

هل فهمت المطلوب منك؟
عندما أقول لك "لنبدأ"، اقرأ المقطع بدقة وبأكبر سرعة ممكنة. سنبدأ من هنا ونكمل بهذه الطريقة *[أشر إلى المقطع الأول في السطر الأول، وتتبع معه بإصبعك على المقاطع الموجودة في السطر الأول بأكمله]*. **هل أنت مستعد؟ لنبدأ.**

ضع **بوضوح** علامة (/) على أي خطأ يرتكبه الطفل.
في حالة قيام الطفل بتصحيح نفسه، قم بوضع دائرة ⬭ حول علامة (/) التي وضعتها مسبقا له
ضع العلامة (**]**) على آخر كلمة قرأها الطفل.

مثال: را أُع سى

	10	9	8	7	6	5	4	3	2	1
(10	مِح	د	ئـ	كث	مِن	ن	بِ	صو	نِ	دا
(20	عِن	لت	دا	ضِ	ة	را	ةِ	جا	ها	فِ
(30	حًا	لى	ر	ذا	رو	دًا	قِ	تِ	طو	مِ
(40	ظ	إ	ه	ذي	حو	جا	كو	دي	م	يح
(50	غُص	صو	مُ	أن	سُ	ـه	أل	حت	ضِ	رُ
(60	رس	ر	قب	قو	رِ	ق	كُ	ال	ءُ	ه
(70	مِي	فـو	تى	هِ	وق	زِ	بِ	جب	دي	ظ
(80	قة	يِن	ـهِ	خ	عُ	د	رى	ط	قا	خى
(90	أُ	غِي	عِن	كِن	مز	تُ	ز	مِن	بع	عا
(100	تن	رخ	با	أج	هق	فـي	حا	كو	عا	دز

الوقت المتبقي من وقت التمرين (عدد الثواني):

ضع علامة (×) في هذا المربع □ في حال أوقفت هذا الجزء من التقييم لأن الطفل لم يقرأ أيًّا من الكلمات في السطر الأول بشكل صحيح.

بعد مرور 60 ثانية، ستقول 'توقف'.

إذا تردد الطفل في قراءة كلمة لمدة تزيد عن 3 ثوانٍ. أشر للكلمة التالية وقل: **"لنكمل من فضلك"**.

قاعدة التوقف المبكّر: إذا وضعت علامة (/) على جميع الإجابات في السطر الأول لأنها خاطئة ولم يصحح الطفل أي خطأ من أخطائه، قل **"شكرا"** وأوقف التمرين. ضع علامة (×) في المربع الموجود في أسفل الصفحة وانتقل للتمرين الذي يليه.

3 القسم 3: قراءة كلمات مخترعة	60 ثانية

هذه بعض الكلمات المخترعة. اقرأ بشكل صحيح أكبر عدد ممكن منها. لا تقرأ حرفا بحرف بل اقرأ الكلمة بالكامل. مثلاً هذه الكلمة المخترعة هي "الفَلاَّطُ".

الآن اقرأ الكلمة التالية: [*أشر إلى كلمة شلاَمِيذُ***]:**

✓: أحسنت، " شلاَمِيذُ"

×: " شلاَمِيذُ " بشكل صحيح، قل: هذه الكلمة المخترعة هي "شَلاَمِيذُ"

لنجرب الآن كلمة أخرى: اقرأ هذه الكلمة [*أشر إلى كلمة "ناسِبَ"***]:**

✓: جيد جداً، "ناسِبَ"

×: هذه الكلمة المخترعة هي "ناسِبَ"

عندما أقول لك "ابدأ"، اقرأ الكلمات بدقة وبأكبر سرعة ممكنة. سنبدأ من هنا ونكمل بهذه الطريقة [*أشر إلى الكلمة الأولى في السطر الأول، وتتبع معه بإصبعك الكلمات في السطر الأول بأكمله.***]**

هل أنت مستعد؟ لنبدأ.

بعد مرور 60 ثانية، ستقول 'توقف'.

إذا تردد الطفل في قراءة كلمة لمدة تزيد عن 3 ثوانٍ. أشر للكلمة التالية وقل: **"لنكمل من فضلك"**.

قاعدة التوقف المبكّر: إذا وضعت علامة (/) على جميع الأجوبة في السطر الأول لأنها خاطئة ولم يصحح الطفل أي خطأ من أخطائه، قل **"شكرا"** وأوقف التمرين. ضع علامة (×) في المربع الموجود في أسفل الصفحة وانتقل للتمرين الذي يليه.

ضع **بوضوح** علامة (/) على أي خطأ يرتكبه الطفل.

في حالة قيام الطفل بتصحيح نفسه، قم بوضع دائرة حول علامة (/) التي وضعتها مسبقا له.

ضع العلامة (**]**) على آخر كلمة قرأها الطفل.

مثال: الفَلاَّطُ شَلاَمِيذُ ناسِبَ

1	*2*	*3*	*4*	*5*	
ضا	تاري	را	أمْنُ	تخمُ	*(5)*
ذف	داف	صلابُ	سا	محبُ	*(10)*
تشبرون	أظي	قير	تِماجي	بُجى	*(15)*
ولمُ	فِع	أغي	صلادُ	جيها	*(20)*
سى	ثُب	ذى	سعِمةُ	يمضُ	*(25)*
شمد	عصِلُ	يهِ	ثول	خبلةُ	*(30)*
فيص	سلعبُ	قملي	بلِخُ	فلا	*(35)*
خزاءً	قسِهُ	جُدءُ	خمب	سمهُ	*(40)*
قدحُن	فُتلانًا	ملصي	شاو	أُحّي	*(45)*
فلُّي	نبُر	غِمسُ	قاط	سحت	*(50)*

الوقت المتبقي من وقت التمرين (عدد الثواني):

ضع علامة (×) في هذا المربع □ في حال أوقفت هذا الجزء من التقييم لأن الطفل لم يقرأ أيًّا من الكلمات في السطر الأول بشكل صحيح.

📖القسم 4 – أ	**قراءة نص شفهياً**

🗣هذه قصة قصيرة، ركز جيدًا واقرأها بشكل صحيح **مع الحركات** وبصوتٍ عالٍ وبأقصى سرعة ممكنة. حين تنتهي، سأسألك بعض الأسئلة حول ما قرأته. هل فهمت المطلوب منك؟ حين أقول لك "لنبدأ"، ابدأ بالقراءة. مستعد؟ **ذكّر الطفل بضرورة القراءة مع الحركات بقولك: "ضع اصبعك على الكلمة الأولى وتذكّر أن تكون قراءتك مع الحركات"، لنبدأ.**

⏱ 60 ثانية

✋ **بعد مرور 60 ثانية**، ستقول **'توقف'**.

➲ حين يتردد الطفل لمدة تزيد على 3 ثوانٍ في قراءة الكلمة. أشر إلى الكلمة التالية وقل: "**لنكمل من فضلك**"

✋ قاعدة التوقف المبكّر: إذا وضعت علامة (/) على جميع الكلمات في السطر الأول على أنها خطأ ولم يصحح الطفل أي خطأ من أخطائه، قل "**شكرًا**" وأوقف التمرين. ضع علامة (×) في المربع الموجود في أسفل الصفحة وانتقل إلى التمرين الذي يليه.

📖القسم 4 – ب	**فهم المقروء**

⏱

اسحب نص القصة من أمام الطفل ووجه إليه الأسئلة أدناه.

➲ اترك للطفل 15 ثانية على الأكثر كي يجيب عن كل سؤال.

وجّه السؤال المقابل لكل سطر قرأه الطفل حتى تصل إلى السطر الذي يحتوي العلامة ([) والتي تشير إلى مكان توقف الطفل عن القراءة.

🗣**سأوجه إليك الآن بعض الأسئلة حول القصة التي قرأت. أجب عن الأسئلة بشكل صحيح.**

✍ضع علامة (/) على أي خطأ يرتكبه الطفل أثناء القراءة. ضع علامة ( ]) بعد آخر كلمة قرأها الطفل.		✍ضع علامة (×) في الخانة التي تتناسب مع إجابة الطفل، ومن ثم انتقل إلى السؤال الذي يليه.	لا إجابة	إجابة خاطئة	إجابة صحيحة
رمى زيد لعبته فسقطت خارج سور الحديقة.	7	**1.** أين سقطت لعبة زيد؟ خارج سور الحديقة.			
ذهب ليحضرها ووجد الأرض متسخة فشعر بالحزن.	41	**2.** لماذا شعر زيد بالحزن؟ لأن الأرض متسخة.			
عاد إلى بيته وأحضر المكنسة ثم بدأ بتنظيف أرض الحديقة من الأوراق.	62	**3.** ماذا فعل زيد بالمكنسة؟ نظّف أرض الحديقة من الأوراق.			
كان عامل النظافة ينظف الحديقة فسلم على زيد مبتسمًا.	53	**4.** من قابل زيد في الحديقة؟ عامل النظافة.			
شعر زيد بالسعادة وعاد إلى بيته فرحًا.	16	**5.** لماذا شعر زيد بالسعادة؟ إجابة مقترحة: أن شارك في تنظيف الشارع. أو أي إجابة صحيحة. يمكن اضافة اجابات اخرى مثل: ساعد عامل النظافة.			

✍ الوقت المتبقي من وقت التمرين (عددالثواني):	
✍ ضع علامة (×) في هذا المربع □ في حال أوقفت هذا الجزء من التقييم لأن الطفل لم يقرأ أي كلمة في السطر الأول بشكل صحيح.	

🕮 X	**القسم 5 - فهم المسموع**

X⏱

يقرأ المقيّم بصوت عال النص التالي ولمرة واحدة فقط وبتأنٍ (كلمة كل ثانية تقريبًا). قل للطفل:

سأقرأ عليك قصة قصيرة بصوت عال، مرة واحدة فقط. و بعد ذلك سأوجّه إليك بعض الأسئلة.
اسمع جيدًا من فضلك وأجب عنها بشكل صحيح. هل فهمت المطلوب منك؟

"استيقظ أبو سعيد في الصباح الباكر نشيطًا؛ ليذهب إلى مزرعته. تناول فطوره ثم لبس معطفه القطني الخفيف. وعندما فتح الباب تفاجأ فقال: سبحان الله! ما أجمل هذا المنظر! الأرض بساط أخضر. عاد أبو سعيد وأيقظ أبناءه مناديًا: نبتت الزروع، تعالوا وانظروا إلى الأعشاب الخضراء وهي تنمو. نهض الأبناء فرحين، وخرجوا إلى الحقل ثم جمعوا بعض الأزهار."

	لا إجابة	إجابة خاطئة	إجابة صحيحة
من الذي استيقظ في الصباح الباكر؟ أبو سعيد.			
أين أراد أبو سعيد الذهاب؟ إلى مزرعته.			
كيف نهض الأبناء؟ فرحين.			
ماذا جمع الأبناء من الحقل؟ بعض الأزهار.			
في أي فصل حدثت القصة؟ في فصل الربيع.			

Jordan 2019 EGRA Stimulus Sheet

ك ـل ـص

جـ	و	ل	د	خـ	ـهـ	ز	ظ	ص	ك
ا	ح	فـ	خ	قـ	ر	د	ة	و	س
نـ	بـ	حـ	ف	تـ	ج	ش	ظ	ص	هـ
ضـ	ق	خ	ب	و	ز	ـة	ـغـ	ـه	ع
خـ	غـ	ذ	ـهـ	ح	يـ	ضـ	ب	ثـ	ق
ـعـ	ذ	سـ	ي	نـ	ط	ـهـ	ص	و	ظ
تـ	ء	د	شـ	ي	غ	ق	بـ	خـ	ن
ذ	م	ب	ط	ث	ح	سـ	ض	ن	فـ
غ	ث	ش	ـعـ	ـه	ضـ	خ	حـ	ط	ج
ز	س	ء	ذ	خ	مـ	ف	عـ	لـ	ث

را أُع سى

دا	نِ	صو	بِ	ن"	مِن	كث	ئـ	د	مِح
فِ	ها	جا	ةِ	را	ة	ضِ	دا	لت	عِن
مِ	طو	تِ	قِ	دًا	رو	ذا	ر	لى	حًا
يح	م	دي	كو	جا	حو	ذي	ه	إِ	ظ
رُ	ضِ	حت	أل	ـه	سُ	أن	مُ	صو	غُص
ه"	ءُ	ال	كُ	ق	رِ	قو	قب	ر	رس
ظ	دي	حب	بِ	زِ	وق	هِ	تى	فـو	مي
خى	قا	ط	رى	د	عُ	خ	ـهِ	ين	قة"
عا	بع	من	ز	ثُ	مز	كِن	عِن	غي	أُ
دز	عا	كو	حا	فـي	هق	أج	با	رخ	تن

قل ط ش يذ نلسب

ضا	تاري	را	أشنُ	تخمُ
ذف	داف	صلِبُ	سا	محبُ
تشِرون	أظي	قير	تِماجي	بُجى
ريمُ	فِع	أغي	صلِدُ	جيها
سى	نُلُب	لى	سعِمةُ	يمضُ
شمد	عطِلُ	يهِ	ثول	خبةُ
فيص	سعلبُ	قِملي	بلِخُ	فلأا
خاءً	قسِهُ	جُدءُ	خمب	سمهُ
قدحُن	فنّلاً	ماصي	شاو	أحّي
فلّي	نرُ	غسمُ	قاط	سحت

رمى زيد الكُعكة فسقطت خارج الحديقة. ذهب صاحب الأرض ووجد الأرض مُتسخة فشعر بالحُزن. عاد إلى بيته وأحضر المِكنسة ثم بدأ بتنظيف أرض الحديقة من الأوراق. كان عامل النظافة يُنظّف الحديقة، فسلّم على زيد مُتسمًا. شعر زيد بالسعادة وعاد إلى بيته فرحًا.

Jordan 2019 EGRA Silent Reading Stimulus Sheet

جلستِ العائلةُ حولَ المائدةِ؛ لتناوُلِ طعامِ الإفطارِ. بعد انتهاءِ مِن
الإفطارِ قال الأبُ: هيّا يا راما، حان موعدُ الذهابِ إلى المدرسةِ.
وضعت راما يدها على خدِّها وصرخت: سنّي تؤلمُني. فاصطحبها
أبوها إلى عيادةِ الأسنانِ؛ لمعالجةِ سنّها. في صباحِ اليومِ التالي قالت
لأمِّها: لن أشتري الحلوى. أعِدّي لي يا أمي شطيرةَ الجُبنِ.

Jordan 2019 EGMA Assessor Protocol

الأردن | USAID
من الشعب الأمريكي

المهمة 1: التعرف إلى الأعداد	A	60 ثانية
في ما يلي بعض الأعداد، أريد منك أن تقرأ لي كل عدد. عندما أقول ابدأ، اقرأ الأعداد بصوت مرتفع وأنت تتبع مع إصبعك. ابدأ من هذا العدد وتابع من اليمين إلى اليسار سطراً بسطر. (أشر إلى أول عدد) ابدأ من هنا. هل أنت مستعد؟ ابدأ. ما هو هذا العدد ؟		• عند انتهاء الوقت المحدد (60 ثانية) (ضع منبه ساعة التوقيت). • إذا توقف الطالب عن التسلسل لمدة 5 ثوان.

(/) غير صحيح أو بدون إجابة
( ]) عند آخر بند إجابة الطالب

02	2	8	9	2
88	58	54	09	44
٦٢	89	85	52	99
989	٦٠٠	750	3٨٦	١٠1

الوقت المتبقي (بالثواني)

ال مه مة 2 :مق ارن ة ا عداد ـتم ري ن	B1	×
انظر إلى هذين العددين. أيهما أكبر؟		× ×
✓ صحيح ٦ هو ا لمتبادر		
✗ العدد ٦ هو ا [أشر إلى ٦] هذا هو العدد ٦ . [أشر إلى ٤] هذا هو العدد ٤. العدد ٦ أكبر من ٤. لمتبادر .		
انظر إلى هذين العددين. أيهما أكبر؟		
✓ صحيح ١٢ هو ا لمتبادر		
✗ العدد ١٢ هو ا [أشر إلى ١٠] هذا هو العدد ١٠. [أشر إلى ١٢] هذا هو العدد ١٢. العدد ٢ أكبر من ١٠. لمتبادر .		

ال مه مة 2 :مق ارن ة ا عداد	B2 & B3	×
انظر إلى كل عددين وأخبرني أيهما أكبر (كرر مع كل مبتدئ)		• في حال أعطى الطالب إجابات خاطئة تشكل متتالي
(/) غير صحيح أو بدون إجابة		

5	7	7	65	٧٣	٧٣
٦4	١3	٦4	١4٦	١2٦	١4٦
٦9	٣2	٣2	٥١٩	٢١٧	٥١٩
٤٧	٦3	٦3	4٦8	48٦	480
٨٦	٨٤	٨٦	68٦	٦٦٤	686

• إذا توقف الطالب عن المحاولة لمدة 5 ثوان.

✗ ⏱	C1 📖	**المهمة 3: العدد الناقص - تمرين**
✗ ✋ ✗ ⊂	🗣 انظر إلى الأعداد التالية ١، ٢، ٤، ما هو العدد المناسب؟ ١ · ٢ · (٣) · ٤ ✓ 🗣 صحيح ٣. لننتقل إلى مثال آخر. ✗ 🗣 العدد هو ٣. وانتبه وأنا أعدّ معي (مع الإشارة إلى كل عدد على حدة): ١، ٢، ٣، ٤، إذن العدد المناسب هو ٣. لننتقل إلى مثال آخر.	
	🗣 انظر إلى الأعداد التالية ٥، ١٠، ١٥، ما هو العدد المناسب (أشر إلى الفراغ) ٥ · ١٠ · ١٥ · (٢٠) ✓ 🗣 صحيح ٢٠. لنتابع. ✗ 🗣 العدد المناسب هو ٢٠ وانتبه وأنا أعد معي (مع الإشارة إلى كل عدد على حدة): ٥، ١٠، ١٥، ٢٠، إذن العدد المناسب هو ٢٠. لنتابع.	

✗ ⏱	C3 & C2 📖	**المهمة 3: العدد الناقص**
✋ ● في حال أعطى الطفل أربع إجابات خاطئة بشكل متتالي. ⊂ ● إتوقف الطالب عن ... لمدة 5 ثوان.	🗣 فيما يلي لدينا أسئلة أخرى من هذا النوع: ضع العدد المناسب داخل المستطيل الفارغ. (كرر هذه التعليمات لكل بند)	
	✍ (/) غير صحيح أو بدون إجابة	

1	3	4	5	**6**	6	6٢٨	6٢٩	**6٣٠**	6٣١
2	١3	١4	**١5**	١6	7	28	**٦6**	٦4	٦2
3	5٠	**6٠**	7٠	8٠	8	6٠	6٥	**7٠**	7٥
4	**1٦٦**	2٦٦	3٦٦	4٦٦	9	7٦٦	7٧٠	7٦٠	**7٥٠**
5	3	5	7	**9**	01	4	9	**١4**	١9

المهمة 4A: عملية الجمع – المستوى 1	D2 & D1	60 ثانية
إليك بعض أسئلة الجمع (مرر يدك على سطر من اليمين إلى اليسار). جد نتائج الجمع لكل ما يأتي. إذا لم تتمكن من معرفة النتائج. انتقل إلى السؤال التالي. هل أنت مستعد؟......ابدأ من هنا (أشر إلى السطر الأول)		• عند انتهاء الوقت المحدد (60 ثانية) ضمن ساعة التوقيت. • إن توقف الطالب عند سؤال لمدة 5 ثوان.
(/) غير صحيح أو بدون إجابة ([ ) عند آخر بند إجابه الطالب		

1 + 2 = (٣)	14 + ٦ = (16)
١ + ٤ = (٥)	١١ + ٦ = (١٧)
8 + ١ = (9)	16 + ٣ = (19)
٦ + ٥ = (٧)	٦ + ٦ = (١٤)
3 + 3 = (6)	9 + 8 = (١٧)
٣ + ١ = (٤)	٥ + ٧ = (١٦)
3 + ٤ = (7)	7 + 7 = (14)
٣ + ٦ = (٩)	٦ + ١٢ = (١٤)
4 + 4 = (8)	١٠ + 4 = (14)
٦ + ٨ = (١٠)	٥ + ١٠ = (١٥)

الوقت المتبقي بالثواني ( )

⏱ ✗	D3 📖	**المهمة 4B: عملية الجمع – المستوى 2**
✋ ● إذا أخطأ الطالب في بلة عن أول خمس و في الم ول. ● في حال أعطى الطفل أربع إجابات خاطئة بشكل متتالي. ⮌ ● إذا قام الطالب باستخدام رق غير فعالة (كاستخدام بع وا شارات)، اطلب من الطالب أن يستخدم طريقة أخرى لحل المسألة. ● إن توقف الطالب عن دلين د لمدة 5 ثوان.		❖ ✏ ورق وقلم 🗣 إليك بعض مسائل الجمع رى. يمكنك استخدام القلم والورقة إذا شئت. ابدأ من هنا ✍ (/) غير صحيح أو بدون إجابة

17 + 2 = (١٩)
14 + ٩ = (23)
24 + 15 = (39)
34 + 46 = (80)
27 + 36 = (63)

المهمة 5A: عملية الطرح المستوى 1	E1 & E2	60 ثانية
قل للطالب: سوف أريك بعض أسئلة الطرح. حاول أن تجيب عن كل سؤال بسرعة. إذا لم تعرف الجواب عن سؤال ما فانتقل إلى السؤال التالي. هل أنت مستعد؟.....ابدأ من هنا وأشر إلى السؤال الأول.		• عندما تبدأ الوقت حدد الوقت لمدة 60 ثانية (ضع علامة على الوقت).
(/) غير صحيح أو بدون إجابة [ ] عدد آخر بين إجابة الطالب		• إذا توقف الطالب عن المحاولة لمدة 5 ثوان.

2 - ١ = (1)	16 - ٢ = (14)
٥ - ٢ = (٣)	١٧ - ١١ = (٦)
8 - ١ = (7)	19 - ٣ = (16)
٧ - ٢ = (٥)	١٥ - ٨ = (٧)
6 - ٤ = (2)	14 - ٩ = (5)
٤ - ٣ = (١)	١٦ - ٩ = (٧)
١٠ - 6 = (4)	14 - 7 = (7)
٩ - ٣ = (٦)	١٤ - ٢ = (١٢)
6 - 3 = (3)	17 - 7 = (١٠)
١٠ - ٢ = (٨)	١٥ - ١٠ = (٥)

الوقت المتبقى (بالثواني)

✗ ⏱	E3 📖	**المهمة 5B: عملية الطرح المستوى 2**

❖ ✏ ورقة وقلم

🗣 إليك بعض مسائل الطرح رى يمكنك استخدام القلم والورقة إذا شئت.
ابدأ من هنا

✍ (/) غير صحيح أو بدون إجابة

)14(= ٣ - 17
)16(= ٩ - 25
)13(= ١٤ - 27
)26(= ٢٤ - 50
)18(= ٢٦ - 44

✋

- إذا أخطأ الطالب في بة عن أول خمس بنود في ال م ول.
- في حال أعطى الطفل أربع إجابات خاطئة بشكل متتالي.

➡

- إذا قام الطالب باستخدام طرق غير فعالة (كاستخدام ا بع شارات)، اطلب من الطالب أن يستخدم طريقة أخرى لحل المسألة.
- إذا توقف الطالب عن الحل لمدة 5 ثوان.

المهمة 6 : المسائل الكلامية	× 📖	× ⏱

❖ ✏ عدادات، ورقة، وقلم.

🗣 لديّ بعض المسائل الحسابية وسوف أطلب منك حلها. هذه بعض الأشياء التي يمكن أن تساعدك. تستطيع استعمالها إذا احتجت لها، ولكنك لست مجبراً على استعمالها. استمع جيداً لكل من هذه المسائل. سأكرر المسألة في حال احتجت لذلك. جيد لنبدأ.

🗣 الكرة تدريبية
ركب ثلاثة أطفال في حافلة. [توقف وتحقق من الطفل]
نزل طفل من الحافلة. [توقف وتحقق من الطفل]
كم عدد الأطفال الذين بقوا في الحافلة؟

✓ 🗣 هذا صحيح. بقي طفلان في الحافلة. لنقم بحل تمرين إضافي.

✗ 🗣 تخيل هذه المعدودات أطفالاً.
قم باختيار ثلاثة أطفال. ركب هؤلاء الأطفال في الحافلة.
نزل طفل من الحافلة. أشر إلى الطفل الذي سيقوم بالنزول من الحافلة باستعمال المعدودات.
كم عدد الأطفال الذين بقوا في الحافلة؟
هذا صحيح. بقي طفلان في الحافلة. لنقم بحل تمارين إضافية.

🗣	المسألة ١ عند ريم قصتان. [توقف وتحقق من الطفل] أهداها والدها ... قصص. [توقف وتحقق من الطفل] كم قصة أصبح معها؟	الإجابة الصحيحة: ٥ ☐✓ ☐× ☐--
🗣	المسألة ٢ هناك في ملعب ... ثمانية أطفال. [توقف وتحقق من الطفل] خمسة منهم ذكور والباقي إناث. [توقف وتحقق من الطفل] كم عدد الإناث؟	الإجابة الصحيحة: ٣ ☐✓ ☐× ☐--
🗣	المسألة ٣ أم عندها سبع قبعات ولديها ثلاث ... سحبات من ... [توقف وتحقق من الطفل] كم تحتاج إضافية حتى يأخذ كل منهم ... واحدة؟	الإجابة الصحيحة: 1 ☐✓ ☐× ☐--
🗣	المسألة ٤ عند فراس عدد من الخراف. [توقف وتحقق من الطفل] اشترى أربعة خراف أخرى. [توقف وتحقق من الطفل] فأصبح عنده عشرة خراف. [توقف وتحقق من الطفل] فكم خروفاً كان عنده؟	الإجابة الصحيحة: 6 ☐✓ ☐× ☐--
🗣	المسألة ٥ وزعت معلمة تسع عشرة قطعة من الحلوى على ثلاثة أطفال بالتساوي. [توقف وتحقق من الطفل] كم قطعة حلوى أخذ كل طفل؟	الإجابة الصحيحة: 1: ☐✓ ☐× ☐--
🗣	المسألة ٦ زرع عمر صفين من ... [توقف وتحقق من الطفل] في كل صف أربع أشجار. [توقف وتحقق من الطفل] كم شجرة زرع؟	الإجابة الصحيحة: ٦ ☐✓ ☐× ☐--

✋
● في حال أعطى الطفل أربع إجابات خاطئة بشكل متتالي.

⮜
● في حال توقف الطفل عند السؤال لمدة ٥ ثواني (ولم يحاول استعمال العدادات، أو الأصابع، أو الورقة والقلم)
أو
● في حال لم يجب الطفل عن السؤال بعد مرور ٣٠ ثانية على توجيه السؤال له.

ملاحظة: تشير عبارات "[توقف وتحقق من الطفل]" في كل مسألة إلى أنك يجب أن تتأكد من فهم الطفل لكل جزء قبل أن تكمل. قد تحتاج لسؤال الطفل، "هل فهمت؟"

Jordan 2019 EGMA Stimulus Sheets

A

20	0	8	9	2
85	45	47	29	77
٦٢	89	54	40	91
989	٦٠٠	750	3٨٦	١٠1

B1

٤	٨
١٢	١٠

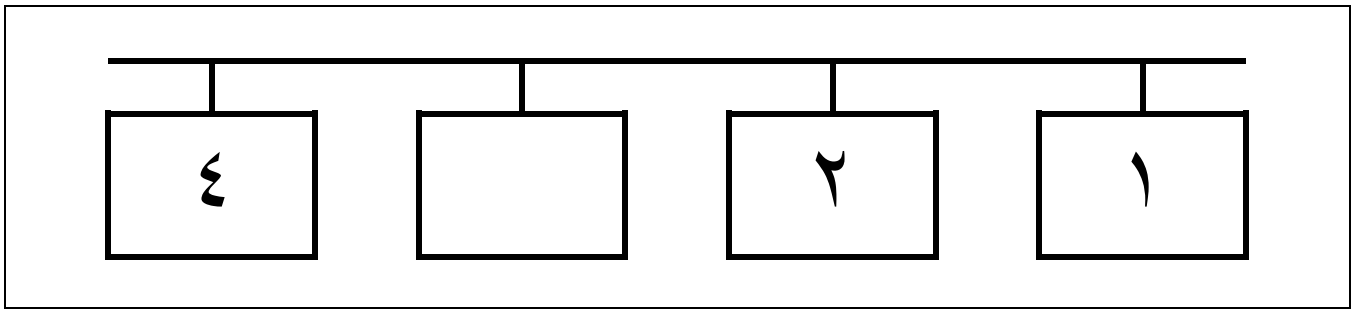
B2

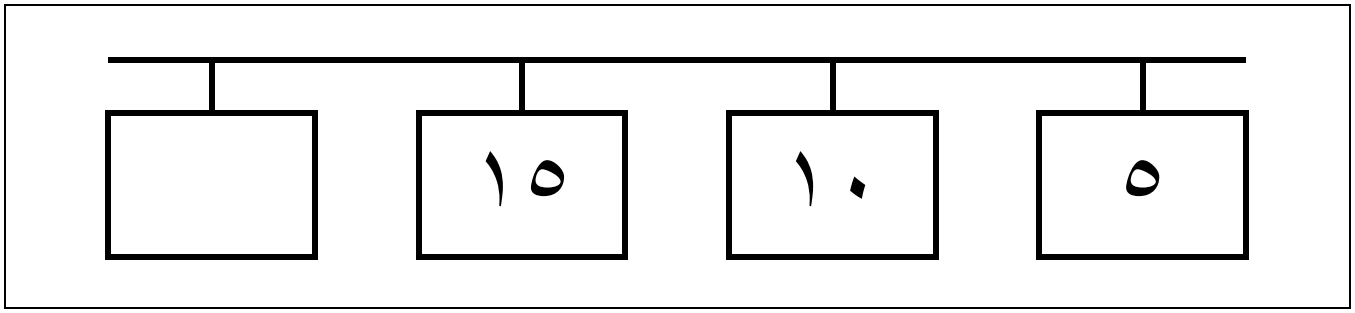
7	5
١3	٢4
٣2	٢9
٦3	4٧
8٤	8٦

B3

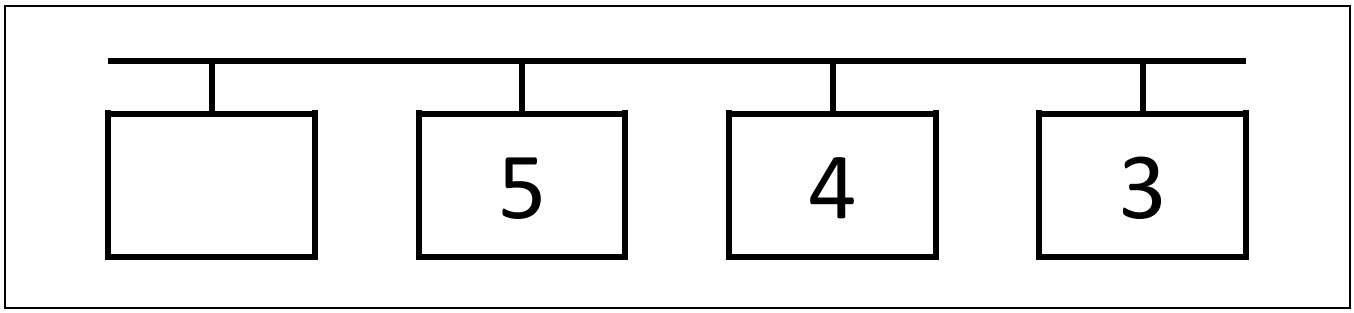
٧٣	65
١2٨	١4٢
٢١٧	٥١٩
48٠	4٠8
6٨٤	68٦

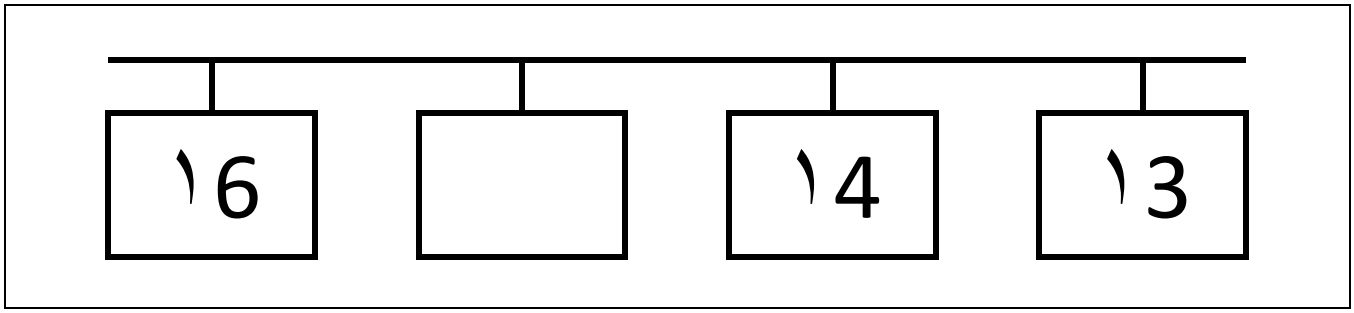
C1

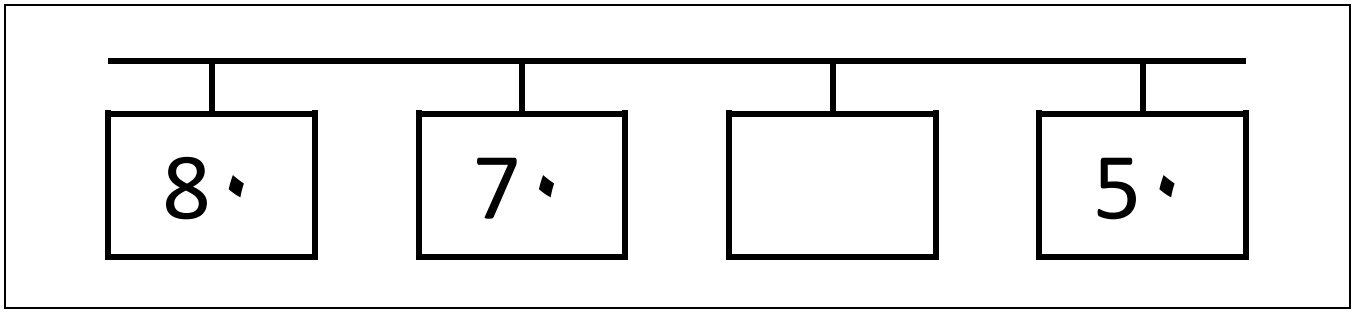


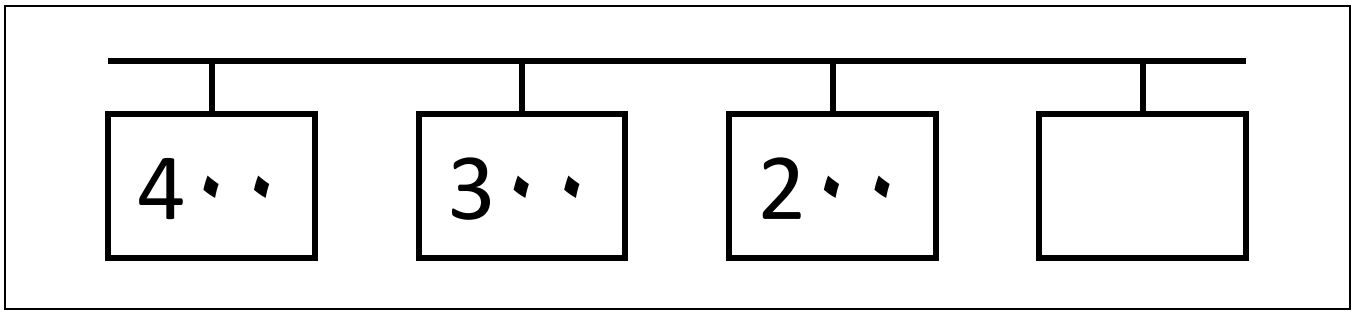


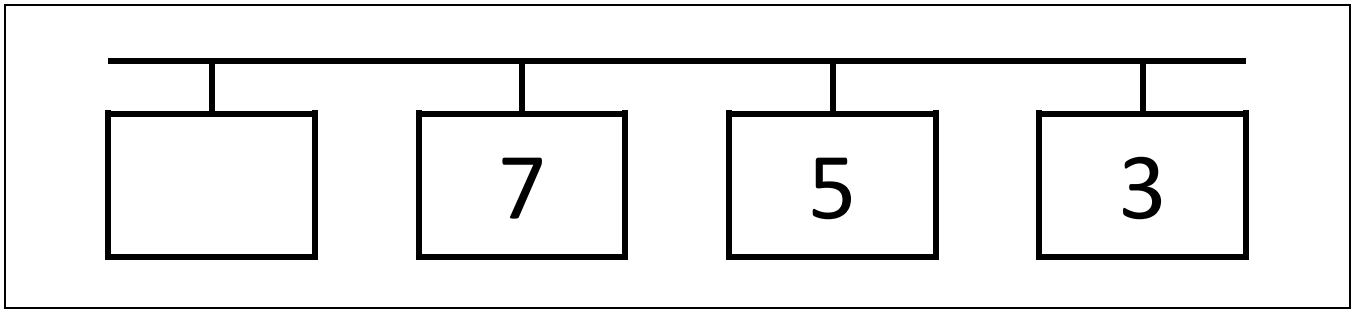
C2



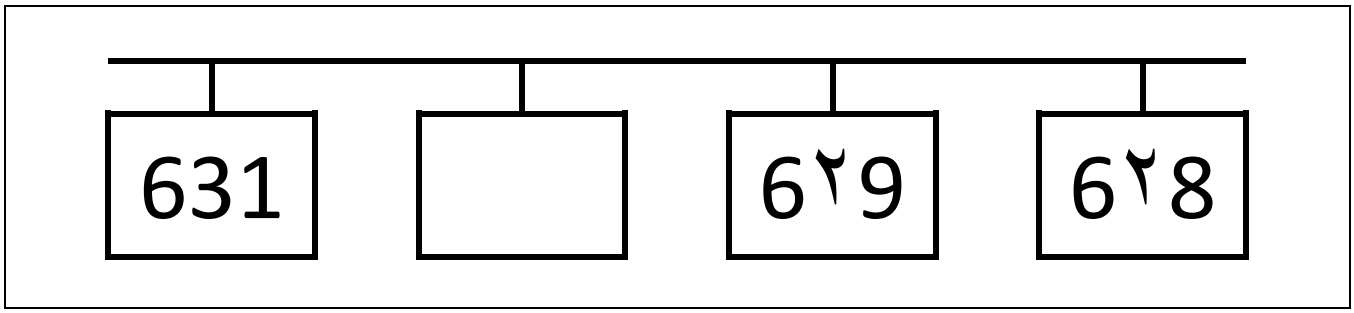


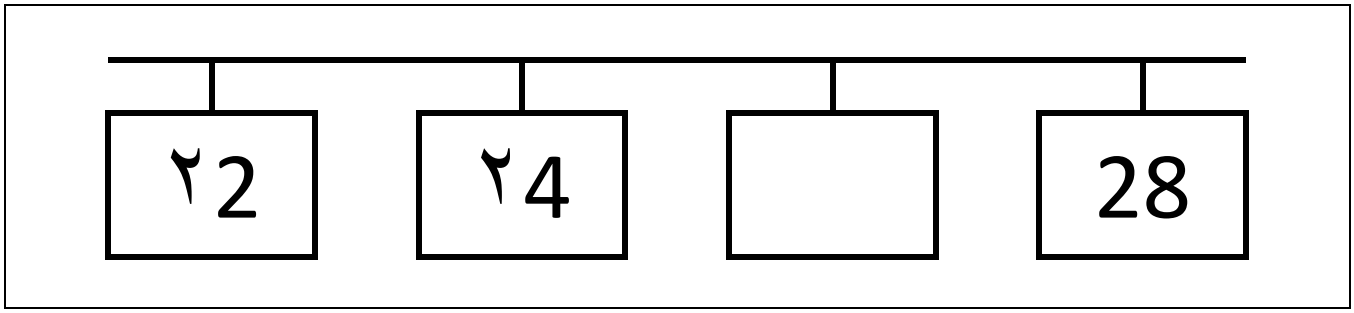


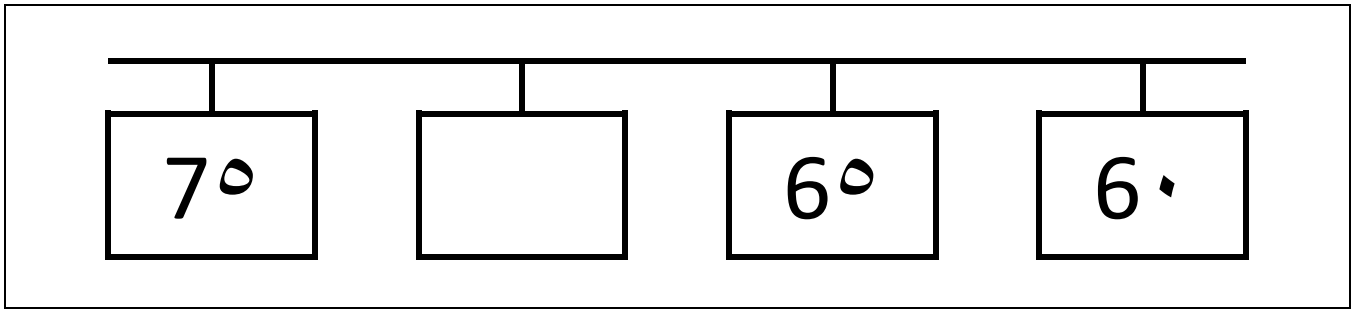


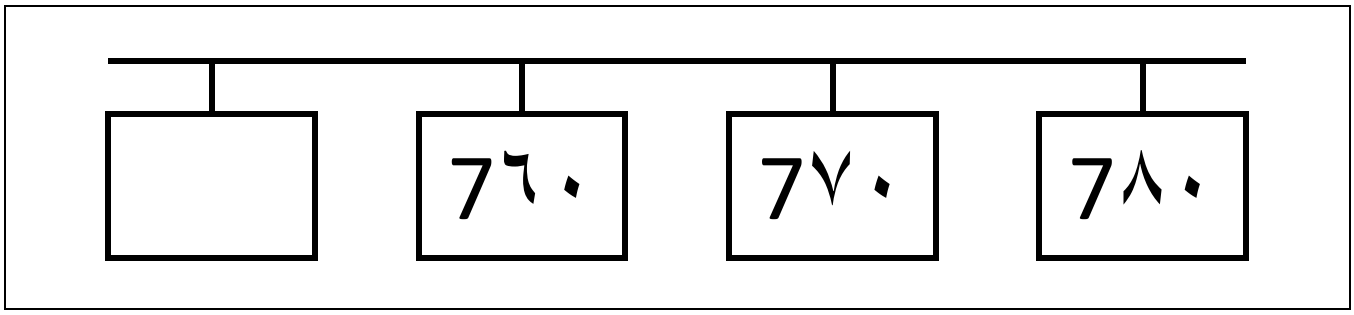


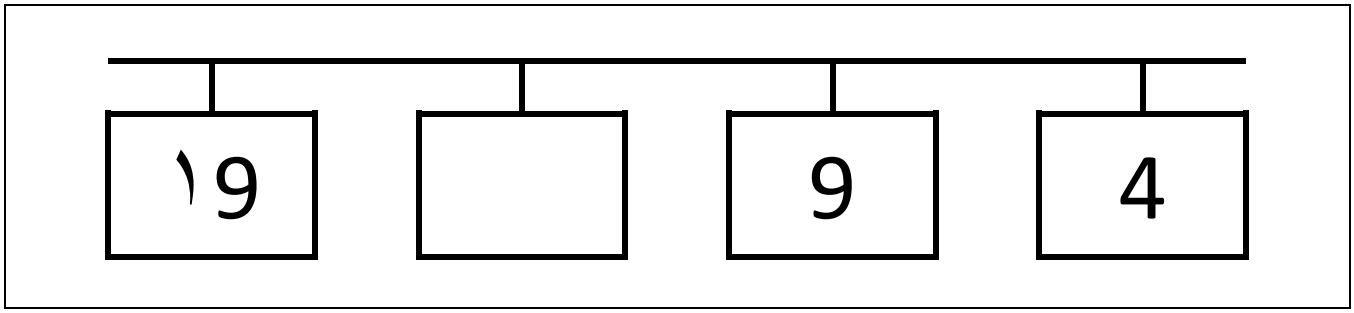
C3



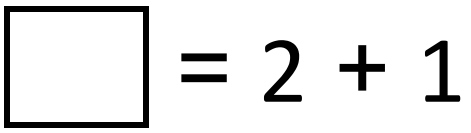
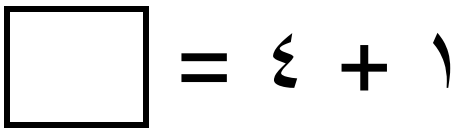
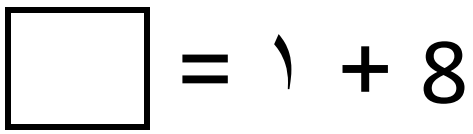
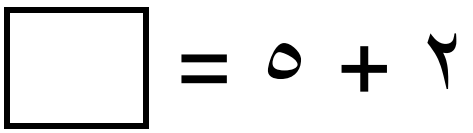
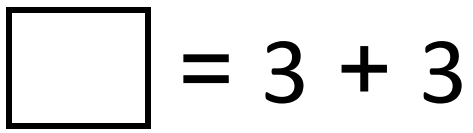
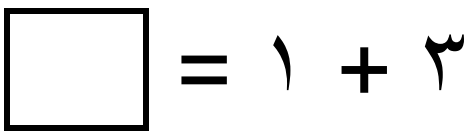
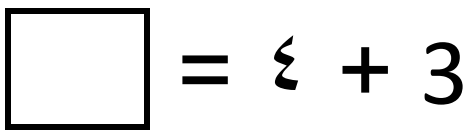
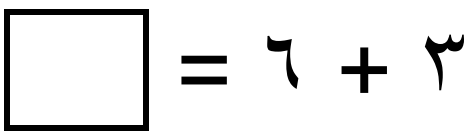
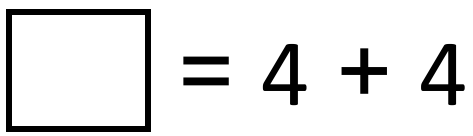








D1


D2

☐ = ٢ + 14
☐ = ٦ + ١١
☐ = ٣ + 16
☐ = ٦ + ٨
☐ = 8 + 9
☐ = ٧ + ٩
☐ = 7 + 7
☐ = ١٢ + ٢
☐ = 4 + ١٠
☐ = ١٠ + ٥

D3

□ = 2 + 17
□ = ٩ + 14
□ = 15 + 24
□ = 46 + 34
□ = 36 + 27

E1

☐ = ١ - 2
☐ = ٢ - ٥
☐ = ١ - 8
☐ = ٢ - ٧
☐ = ٤ - 6
☐ = ٣ - ٤
☐ = 6 - ١٠
☐ = ٣ - ٩
☐ = 3 - 6
☐ = ٢ - ١٠

E2

☐ = ٢ - 16
☐ = ١١ - ١٧
☐ = ٣ - 19
☐ = ٨ - ١٥
☐ = ٩ - 14
☐ = ٩ - ١٦
☐ = 7 - 14
☐ = ٢ - ١٤
☐ = 7 - 17
☐ = ١٠ - ١٥

E3

☐ = ٣ - 17
☐ = ٩ - 25
☐ = ١٤ - 27
☐ = ٢٤ - 50
☐ = ٢٦ - 44

Principal and School Questionnaire

RAMP Initiative, midline survey: Principal and school questionnaire

1.	Starting time **[Use 24-hour time HH:MM]**	* * : * *	* * : * *	وقت البدء **[استخدم توقيت 24 ساعة – س س:دد]**	
2.	Interview date **[DD/MM/YY]**	* * / * * / * *	* * / * * / * *	تاريخ المقابلة **[يوم شهر سنة]**	
3.	Interview status	Refused → **Thank teacher and end interview** .....1 Partially completed .....2 Completed 3	رفض المقابلة ← **اشكر المعلم وقم بإنهاء المقابلة**.....4 تمت المقابلة جزئيا.....6 تمت المقابلة كاملة.....5	حالة المقابلة	
4.	What is your position at the school?	Principal.....1 Deputy Principal.....2 Other.....3 Don't know / Refuse.....888	مدير.....1 مساعد مدير.....2 غير ذلك.....3 عرف/رفض الإجابة.....888	ما هو وصفك الوظيفي في المدرسة؟	
5.	**Gender**	Male.....1 Female.....2	ذكر.....4 أنثى.....6	**جنس المدير**	
	I would like to begin by asking you a few questions about the teachers in the school.			**أود البدء بسؤالك بضعة أسئلة حول المعلمين في المدرسة.**	
6.	How many total teachers are there teaching KG2 to Grade 3 in this school? [Note (a) is the total of all the teachers in question 7]	Teachers.....***** Don't know/Refuse.....888	*****.....العدد الإجمالي للمعلمين عرف/رفض الإجابة.....888	كم العدد الإجمالي للمعلمين الذين يُدرّسون حاليا في الصفوف من الروضة إلى الصف الثالث في هذه المدرسة؟ (أدخل الرمز 888 للإجابة "لا أعرف" أو 999 للإجابة "لا توجد معلومات") العدد الإجمالي للمعلمين في المدرسة.	

7.	How many of the *(a)* **KG2 to Grade 3** teachers teaching in the school now were also teaching in the school at the **start of the school year**? *[Note (a) is the total of all the teachers in question 6]*	Teachers .... ***** Don't know/Refuse 888	***** .... المعلمين 888 .... عرف/أفض ا بة	كم عدد (أ) معلمي **الروضة الثانية حتى الصف الثالث** الذين يُدرّسون حالياً في المدرسة وكانوا أيضاً يُدرّسون في المدرسة **في بداية العام الدراسي؟** م ة: يمثل الرمز (أ) في السؤال السادس العدد ا لمي للمعلمين في المدرسة.	
	Now I would like to ask you a few questions about the students in the school.			**أود طرح بعض ا لة عليك عن الطلبة في المدرسة.**	
8.	How many total students are there in KG2 to Grade 3 in this school? [Note (b) is the total of all the teachers in question 9]	Students .... ***** Don't know/Refuse 888	***** .... ال ب 888 .... عرف/أفض ا بة	كم العدد ا لمي للطلبة من الروضة الثانية حتى الصف الثالث في هذه المدرسة؟ م ة يمثل الرمز (ب) في السؤال التاسع العدد ا لمي للطلبة في المدرسة.	
9.	How many of the *(b)* **KG2 to Grade 3** students in the school are **Syrian refugees**? *[Note (b) is the total of all the students in question*8	Students............................ ***** Don't know/Refuse 888	***** .... ال ب 888 .... عرف/أفض ا بة	كم عدد (ب) ب **الروضة الثانية حتى الصف الثالث** في المدرسة **من اجئين السوريين؟** م ة يمثل الرمز (ب) في السؤال الثامن العدد ا لمي للطلبة في المدرسة.	
10.	How many of the **KG2 to Grade 3** students in the school have **visible disabilities**? (e.g. children using wheelchairs, crutches, hearing aids, visual aids [such as domes, handheld magnifiers etc.]. NOTE: not every child with spectacles is a child with a visual disability)	Students............................ ***** Don't know/Refuse 888	***** .... ال ب 888 .... عرف/أفض ا بة	كم عدد ب **الروضة الثانية حتى الصف الثالث** في المدرسة **من ذوي ا قات المرئية؟** (على سبيل المثال: اطفال الذين يستخدمون الكراسي المتحركة أو العكازات أو المعينات السمعية أو المعينات البصرية كالعدسات المكبرة أو ما شابه. م ة: ليس كل من يرتدي نظارات هو طفل من ذوي ا عاقات البصرية.)	
	Now I would like to ask you some GENERAL questions about the RAMP initiative that you and your school have been implementing. Let us begin with the training			**أود أنْ أطرح على حضرتك بعض ا لة العامة حول مبادرة الـRAMP التي اشتركت في تطبيقها أنت ومدرستك. دعنا نبدأ بالحديث عن التدريب.**	

11.	How many **KG2 to Grade 3** teachers teaching in the school have attended all the RAMP training that they should have?	Teachers ..... ***** Don't know/Refuse 888	***** ..... عدد المعلمين 888 ..... لا أعرف/لا أفضل الإجابة	كم عدد معلمي **الروضة الثانية حتى الصف الثالث** الذين يُدرّسون في المدرسة ولديهم حضروا جميع جلسات التدريب التي كان يجب أن يكونوا قد حضروها؟	
12.	What are the main reasons for some teachers not attending the training? **[Do not read the options; just circle all that apply.]**	None of the teachers missed the training (if selected, the other options do not apply) ..... 1 They were not teaching at the school ..... 1 They were not notified on time. 1 They were ill ..... 1 They did not want to attend ..... 1 Other (specify): ..... 1 _____ Don't know/Refuse 888	لم يكن المعلمون يُدرّسون في المدرسة ..... 4 لم يتم إخطارهم في الوقت المطلوب ..... 4 كانوا مرضى ..... 4 لديهم أسباب أخرى للغياب عن جلسات التدريب ..... 4 أخرى (حدد): ..... 4 _____ لا أعرف/لا أفضل الإجابة 888	ما هي الأسباب الرئيسية التي جعلت بعض المعلمين يتغيبون عن حضور جلسات التدريب؟ **[لا تقرأ الخيارات. ضع دائرة حول جميع ما ينطبق من الخيارات.]**	
13.	Did you attend the3 to 4-day teacher training workshop for **Principals?** If yes, did you attend all the days, more than half of the days, or less than half of the days? **[Read all the options to the teacher; only circle one response]** **If yes, skip to 15**	No ..... 0 Yes, all days ..... 1 Yes, more than half of the days 2 Yes, less than half of the days ..3 Don't know/Refuse 888	0 ..... نعم، كل الأيام ..... 4 نعم، أكثر من نصف الأيام ..... 6 نعم، أقل من نصف الأيام ..... 5 لا أعرف/لا أفضل الإجابة 888	هل حضرت ورشة العمل التدريبية الخاصة بالمديرين/المديرات التي امتدت من ثلاثة إلى أربعة أيام في أي من الأماكن التي تلقيت التدريب فيها، أم أكثر من ذلك؟ **[إذا لزم الأمر، أسأل: هل حضرت أكثر من نصف الأيام أم أقل من نصف الأيام؟ ثم ضع دائرة حول الخيار الواحد فقط الذي يمثل الإجابة.]** **إذا كانت الإجابة بنعم لجميع الأيام، انتقل إلى السؤال 15**	

14.	Why did you not attend the training? **[Do not read the options; just circle all that apply.]**	I was not teaching at the school 1 I was not notified on time ..... 1 I was ill ..... 1 I did not want to attend ..... 1 Other (specify): ..... 1 _____ Don't know/Refuse 888	لم أكن أُدرّس في هذه المدرسة.....4 لم يتم إعلامي في الوقت المطلوب.....4 كنت مريضاً .....4 لم أكن أرغب في الحضور .....4 أخرى (حدد).....4 _____ عرف/فضل ال إجابة. 888	لماذا لم تحضر التدريب؟ [**اقرأ الخيارات؛ اضغط على جميع الردود التي تنطبق من الخيارات.**]	
	Next I want to get a sense of your overall impression of the RAMP initiative. **Please respond to each of the following statements by indicating whether you strongly agree, agree, are neutral about the statement, disagree or strongly disagree.**			**أود أن أعرف ما انطباعك العام حول برنامج RAMP.** **أرجو الرد على كلّ من العبارات التالية بما يوافق بشدة، أو يوافق، أو محايد، أو لا يوافق، أو لا يوافق بشدة.**	
15.	I understand the goals of the RAMP initiative.	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse 888	أوافق بشدة .....4 أوافق.....6 محايد.....5 لا أوافق.....1 لا أوافق بشدة .....3 عرف/فضل ال إجابة. 888	أنا أفهم أهداف برنامج RAMP.	

16.	The training my teachers received for the RAMP initiative was adequate.	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse..... 888	أوافق بشدة ..... 4 أوافق..... 6 محايد..... 5 لا أوافق..... 1 لا أوافق بشدة..... 3 لا أعرف/رفض الإجابة..... 888	كان التدريب الذي حصل عليه معلموي مناسبًا لقدرتهم على تنفيذ برنامج RAMP.	
17.	I feel confident that the teachers in my school are implementing the RAMP initiative in their classes.	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse..... 888	أوافق بشدة ..... 4 أوافق..... 6 محايد..... 5 لا أوافق..... 1 لا أوافق بشدة..... 3 لا أعرف/رفض الإجابة..... 888	أثق بأن المعلمين في مدرستي يطبقون قدرتهم بشكل مناسب في برنامج RAMP.	
18.	Sufficient guidance and support provided by the coaches and/or supervisors have helped the teachers implement the activities of the RAMP initiative.	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse..... 888	أوافق بشدة ..... 4 أوافق..... 6 محايد..... 5 لا أوافق..... 1 لا أوافق بشدة..... 3 لا أعرف/رفض الإجابة..... 888	لم يتلق المعلمون الدعم الكافي من المدربين و/أو المشرفين على تطبيق أنشطة قدرتهم في برنامج RAMP.	

19.	The RAMP initiative has positively impacted student achievement in the school.	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse..... 888	أوافق بشدة ..... 4 أوافق..... 6 محايد..... 5 أختلف..... 1 أختلف بشدة..... 3 لا أعرف/أرفض الإجابة. 888	إلى أي مدى تعتقد أن برنامج RAMP قد أثر إيجابياً على مستوى تحصيل الطلاب في المدرسة؟	
20.	Students in this school are more enthusiastic about learning because of the RAMP initiative.	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse..... 888	أوافق بشدة ..... 4 أوافق..... 6 محايد..... 5 أختلف..... 1 أختلف بشدة..... 3 لا أعرف/أرفض الإجابة. 888	هل يبدي الطلاب في هذه المدرسة حماساً أكثر تجاه التعلم بسبب برنامج RAMP؟	
21.	The materials provided to implement the RAMP initiative were sufficient.	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse..... 888	أوافق بشدة ..... 4 أوافق..... 6 محايد..... 5 أختلف..... 1 أختلف بشدة..... 3 لا أعرف/أرفض الإجابة. 888	المواد التعليمية المُزودة لتطبيق برنامج RAMP كانت كافية.	
	Here are the last questions about RAMP			**الأسئلة الأخيرة عن برنامج RAMP**	

22.	What are the **overall** aspects that you think are positive? **[Do not read the options; just circle all that apply.]**	The initiative had positive impact on learning ..... 1 Activities support learning..... 1 Development of thinking skills. 1 Improvement of student skills .. 1 Improvement of teaching skills 1 Training..... 1 Supervisor support (visits, feedback and monthly meetings) ..... 1 Encouragement of school and/or district ..... 1 Parents enjoyed the project..... 1 Other ..... 1 Don't know/Refuse 888	لقد كان للمبادرة تأثير إيجابي على عملية التعلم..... 1 تدعم الأنشطة عملية التعلم.....1 تنمية مهارات التفكير.....1 تحسين مهارات الطلبة.....1 تحسين مهارات التدريس.....1 التدريب.....1 دعم المشرف (الزيارات، التغذية الراجعة الأسبوعية، الاتصالات الشهرية).....1 تشجيع المدرسة و/أو المجتمع.....1 لقد استمتع أولياء الأمور بهذا المشروع.....1 غير ذلك .....1 عرف/فضل عدم الإجابة.....888	ما هي الجوانب **الإجمالية** التي تعتقد أنها إيجابية؟ **[لا تقرأ الخيارات؛ بل فقط ضع دائرة حول كل الخيارات التي تنطبق من الخيارات.]**	

23.	What are the aspects with respect to **reading** that you think are positive? **[Do not read the options; just circle all that apply.]**	Students enjoyed the activities.. 1 Activities and materials support the curriculum ..... 1 Materials (teacher notes, lesson notes, workbooks)..... 1 Other ..... 1 Don't know/Refuse 888	لقد استمتع التلاميذ بالأنشطة .....1 لقد دعمت الأنشطة والمواد المنهج .....1 المواد (ملاحظات المعلم، كراسة التلميذ، الدوس) .....1 غير ذلك .....1 عرف/فضل عدم الإجابة 888	ما الجوانب التي تعتقد بأنها إيجابية فيما يخص **القراءة؟** **[قرأ الخيارات وضع دائرة حول جميع الخيارات المطبقة من الخيارات.]**	
24.	What are the aspects with respect to **mathematics** that you think are positive? **[Do not read the options; just circle all that apply.]**	Students enjoyed the activities.. 1 Activities and materials support the curriculum ..... 1 Materials (teacher notes, lesson notes, workbooks)..... 1 Other ..... 1 Don't know/Refuse 888	لقد استمتع التلاميذ بالأنشطة .....1 لقد دعمت الأنشطة والمواد المنهج .....1 المواد) المعلم، م ات الدوس، كراسة الطالب(.....1 غير ذلك .....1 عرف/فضل عدم الإجابة 888	ما الجوانب التي تعتقد بأنها إيجابية فيما يخص **الحساب؟** **[قرأ الخيارات وضع بعض دائرة حول جميع الخيارات المطبقة من الخيارات.]**	

25.	What are the **overall** aspects that you think are frustrating or negative? **[Do not read the options; just circle all that apply.]**	Objectives of project and materials not clear ..... 1 Increased workload for teachers 1 Training (dates/timing; duration; arrangements) ..... 1 Training (content, presentation) 1 Supervisors not providing support (visits and feedback) .... 1 Supervisors creating confusion (feedback) ..... 1 Initiative did not have a positive impact on learning ..... 1 Insufficient encouragement and support, no reward ..... 1 Too much time/effort required to mark the student workbooks ..... 1 Teachers transferring in and out of school..... 1 Other ..... 1 Don't know/Refuse 888	لم تكن أهداف ومواد المشروع واضحة 1..... زيادة عبء العمل على المعلمين 1..... التدريب (الموعد/المدة الزمنية؛ التوقيت) ..... 1 التدريب (المحتوى، العرض) 1..... عدم قيام المشرفين بتقديم الدعم (الزيارات والتغذية الراجعة) ..... 1 يُحدث المشرفون حالة من الارتباك (التغذية الراجعة) ..... 1 لم يكن للمبادرة تأثير إيجابي على التعلم 1..... عدم وجود الدعم والتشجيع الكافيين، وانعدام الحوافز 1..... الحاجة لجهد ووقت كبيرين لتصحيح كراسة الطلبة ..... 1 تنقل المعلمين داخل وخارج المدرسة 1..... غير ذلك ..... 1 عرف/رفض الإجابة 888	ما هي الجوانب الجوهرية التي تعتقد أنها تكون مُحبطة أو سلبية؟ [**لا تقرأ الخيارات؛ فقط اطلب وضع دائرة حول جميع ما ينطبق من الخيارات.**]	

26.	What are the aspects with respect to **reading** that you think are frustrating or negative? **[Do not read the options; just circle all that apply.]**	Activities did not support the curriculum ..... 1 Activities too time consuming/too many for each day ..... 1 Activities too difficult ..... 1 Activities too easy ..... 1 Students did not enjoy the activities ..... 1 Other ..... 1 Don't know/Refuse 888	تعمل الأنشطة على تعزيز المهارات الجديدة ..... 1 تستغرق الأنشطة الكثير من الوقت/عدد الأنشطة كثير كل يوم ..... 1 الأنشطة قصيرة جداً ..... 1 الأنشطة سهلة جداً ..... 1 لا يستمتع الطلاب بالأنشطة ..... 1 غير ذلك ..... 1 عرف/فضل ألا يجيب 888	ما الجوانب التي تعتقد أنها مُحبطة أو سلبية فيما يخص **القراءة؟** **[قرأ الخيارات فقط وضع دائرة حول جميع الخيارات التي تنطبق من الخيارات.]**	
27.	What are the aspects with respect to **mathematics** that you think are frustrating or negative? **[Do not read the options; just circle all that apply.]**	Activities did not support the curriculum ..... 1 Activities too time consuming/too many for each day ..... 1 Activities too difficult ..... 1 Activities too easy ..... 1 Students did not enjoy the activities ..... 1 Other ..... 1 Don't know/Refuse 888	تعمل الأنشطة على تعزيز المهارات الجديدة ..... 1 تستغرق الأنشطة الكثير من الوقت/عدد الأنشطة كثير كل يوم ..... 1 الأنشطة قصيرة جداً ..... 1 الأنشطة سهلة جداً ..... 1 لا يستمتع الطلاب بالأنشطة ..... 1 غير ذلك ..... 1 عرف/فضل ألا يجيب 888	ما الجوانب التي تعتقد أنها مُحبطة أو سلبية فيما يخص **الحساب؟** **[قرأ الخيارات فقط وضع دائرة حول جميع الخيارات التي تنطبق من الخيارات.]**	

	Finally, I would like to ask you about parental involvement in the school.			وأخيراً، أودّ أن أطرح عليك أسئلة حول مشاركة أولياء الأمور في المدرسة.	
28.	Which of the **scheduled PTC meetings** took place this year? **[Circle all that apply]**	No PTC meetings took place this year (if selected, the other options do not apply) ..... 1 The first PTC meeting took place as scheduled ..... 1 The second PTC meeting took place as scheduled..... 1 The third PTC meeting took place as scheduled..... 1 The fourth PTC meeting took place as scheduled..... 1 Don't know/Refuse 888	لم تُعقد أي اجتماعات لمجالس أولياء الأمور هذا العام (إذا تم اختيار هذه الخيارات الأخرى تنطبق)1 عقد اجتماع مجلس أولياء الأمور والمعلمين كما هو مقرر ..... 1 عقد اجتماع مجلس أولياء الأمور والمعلمين الثاني كما هو مقرر ..... 1 عقد اجتماع مجلس أولياء الأمور والمعلمين الثالث كما هو مقرر ..... 1 عقد اجتماع مجلس أولياء الأمور والمعلمين الرابع كما هو مقرر ..... 1 أعرف/فضل ألا أجيب 888	أي من اجتماعات مجلس أولياء الأمور والمعلمين تم عقدها هذا العام؟ [اعرض دائرة حول جميع الإجابات التي تنطبق.]	
29.	Was a **KG2-G3 parent open day conducted** this year? **[Circle all that apply]**	No ..... 0 Yes ..... 1 Don't know/Refuse 888	0..... نعم.....4 عرف/فضل ألا أجيب 888	هل تم عمل يوم مفتوح لأولياء الطلبة في الصف الثالث من المرحلة الابتدائية حتى الآن هذا العام؟ [اعرض دائرة حول جميع الإجابات التي تنطبق.]	
30.	Was a **KG2-G3 parent awareness day** conducted this year? **[Circle all that apply]**	No ..... 0 Yes ..... 1 Don't know/Refuse 888	0..... نعم.....4 عرف/فضل ألا أجيب 888	هل تم عمل يوم توعوي لأولياء الطلبة في الصف الثالث من المرحلة الابتدائية حتى الآن هذا العام؟ [اعرض دائرة حول جميع الإجابات التي تنطبق.]	
31.	Ending time **[Use 24-hour time HH:MM]**	*__ : *__	*__ : *__	وقت الانتهاء [إذا تعمل تقويت 42 ساعة، سجل: 00]	
	Thank you very much.			شكراً جزيلاً لك.	

Teacher Questionnaire

RAMP Initiative, midline survey: Teacher questionnaire

1.	Starting time **[Use 24-hour time HH:MM]**	*__/*__	*__/*__	وقت البدء **[سجل الوقت بتنسيق 24 ساعة – س س : د د]**	
2.	Interview date **[DD/MM/YY]**	*__/*__/*__	*__/*__/*__	تاريخ المقابلة **[يوم/شهر/سنة]**	
3.	Interview status	Refused → **Thank teacher and end interview** .....1 Partially completed .....2 Completed 3	رفض المعلم المشاركة ← **شكر المعلم وقم بإنهاء المقابلة** .....4 تمت المقابلة جزئيا .....6 تمت المقابلة كاملة 5	حالة المقابلة	
4.	**Gender**	Male .....0 Female.....1	ذكر .....1 أنثى 6	جنس المعلم	
5.	What grade are you currently teaching?	Grade 2.....1 Grade 3.....2	الصف الثاني.....1 الصف الثالث.....2	ما الصف الذي تدرسه حاليا؟	
6.	How many girls are enrolled in your classroom? Note: this number can be collected from the teacher.	The number is…… ***** Don't know/Refuse.....888	العدد ..... ***** رفض/عرفة 888	كم عدد الطالبات الإناث المسجلات في صفك؟ ملاحظة: هذا الرقم يمكن الحصول عليه من المعلم.	
7.	How many boys are enrolled in your classroom? Note: this number can be collected from the teacher.	The number is…… ***** Don't know/Refuse.....888	العدد ..... ***** رفض/عرفة 888	كم عدد الطلاب الذكور المسجلين في صفك؟ ملاحظة: هذا الرقم يمكن الحصول عليه من المعلم.	

8.	How many girls are present today in the sampled classroom for this grade? Note: this number should be recorded during the student sampling process to reflect the number of students present.	The number is…… ***** Don't know/Refuse 888	***** ..... العدد 888 ..... عرف/رفض إجابة	كم عدد الطالبات الحاضرات اليوم في الشعبة التي ترتفي العينة هذا الصف؟ مهمة يجب تسجيل هذا الرقم لعملية أخذ العينة التي ستحدد الطالبات الحاضرات	
9.	How many boys are present today in the sampled classroom for this grade? Note: this number should be recorded during the student sampling process to reflect the number of students present.	The number is…… ***** Don't know/Refuse 888	***** ..... العدد 888 ..... عرف/رفض إجابة	كم عدد الطلاب الذكور الحاضرين اليوم في الشعبة التي ترتفي العينة هذا الصف؟ مهمة يجب تسجيل هذا الرقم لعملية أخذ العينة التي ستحدد عدد الطلاب الحاضرين.	
	I would like to begin by asking you a few questions about your class.			أود البدء بطرح بعض الأسئلة حول صفك.	
10.	Have you been the only teacher to teach reading and mathematics to this class since the beginning of the school year? If no, how any other teachers, as far as you know, have taught reading and mathematics to this class? **[Read all the options to the teacher; only circle one response]**	Yes ..... 1 No, one other teacher ..... 2 No, two other teachers ..... 3 No, three or more other teachers 4 Don't know/Refuse 888	4 ..... نعم 6 ..... معلم واحد آخر 5 ..... معلمان آخران 1 ..... ثلاثة معلمين آخرين أو أكثر 888 ..... عرف/رفض إجابة	هل أنت المعلم الوحيد الذي يقوم بتدريس القراءة والحساب لهذا الصف منذ بدء العام الدراسي الحالي؟ إذا كانت الإجابة كم عدد المعلمين الآخرين الذين قاموا بتدريس الحساب والقراءة لهذا الصف بحسب علمك؟ **[قرأ الخيارات المتاحة على المعلم؛ ثم ضع دائرة حول رقم واحد فقط يمثل الإجابة.]**	
	Now I would like to ask you a few questions about yourself			**أود أن أسألك بضع أسئلة عن حضرتك**	

11.	What is your highest level of academic education?	Diploma ..... 1 Bachelor's degree ..... 2 Higher diploma ..... 3 Master's degree ..... 4 PhD ..... 5 Other (specify): ..... 6 _____ Don't know/Refuse 888	4 ..... دبلوم 6 ..... بكالوريوس 5 ..... دبلوم عالي 1 ..... ماجستير 3 ..... دكتوراه 2 ..... أخرى (حدد): ..... _____ 888 ..... عرف/فضل عدم الإجابة	ما أعلى مستوى دراسي أو شهادة أكاديمية حصلت عليها؟	
12.	For how many years have you been a teacher?	Years ..... ***** Don't know/Refuse 888	***** ..... سنوات 888 ..... عرف/فضل عدم الإجابة	منذ كم سنة تعمل كمعلم؟	
13.	Are you a substitute or a permanent teacher at this school?	Permanent teacher ..... 1 Substitute teacher ..... 2 Don't know/Refuse 888	4 ..... معلم أصيل 6 ..... معلم بديل 888 ..... عرف/فضل عدم الإجابة	هل أنت معلم أصيل في هذه المدرسة أم معلم بديل؟	
14.	During your **pre-service** training, did you receive any specific training on how to teach early grade (grade 1 to 3) **reading**?	No ..... 0 Yes ..... 1 Don't know/Refuse 888	0 ..... 4 ..... نعم 888 ..... عرف/فضل عدم الإجابة	هل تلقيت أي تدريب خاص حول كيفية تدريس **القراءة** في الصفوف (من الأول حتى الثالث) الابتدائي؟	
15.	**Not including the training for RAMP initiative**: have you attended any **in-service** training on how to teach early grade (grade 1 to 3) **reading**?	No ..... 0 Yes ..... 1 Don't know/Refuse 888	0 ..... 4 ..... نعم 888 ..... عرف/فضل عدم الإجابة	**عدا عن التدريبات المبادرة لـ RAMP**، هل سبق وتلقيت أي من التدريبات **أثناء الخدمة** حول كيفية تدريس **القراءة** في الصفوف (من الأول حتى الثالث) الابتدائي؟	

16.	During your **pre-service** training, did you receive any specific training on how to teach early grade (grade 1 to 3) **mathematics**?	No .... 0 Yes .... 1 Don't know/Refuse 888	0... 4...نعم 888................ لا أعرف/رفض الإجابة	أثناء تدريب ما قبل الخدمة، هل حصلت على أي تدريب خاص حول كيفية تدريس **الحساب** للصفوف المبكرة (من الأول حتى الثالث ابتدائي)؟	
17.	**Not including the training for RAMP initiative**: have you attended any **in-service** training on how to teach early grade (grade 1 to 3) **mathematics**?	No .... 0 Yes .... 1 Don't know/Refuse 888	0... 4...نعم 888................ لا أعرف/رفض الإجابة	**عدا عن تدريبات مبادرة الـRAMP**، هل سبق والتحقت بأي من التدريبات **أثناء الخدمة** حول كيفية تدريس **الحساب** للصفوف المبكرة (من الصف الأول حتى الثالث ابتدائي)؟	
	Now I would like to ask you some GENERAL questions about the Early Grade Reading and Mathematics Initiative (RAMP) that you and your school have been involved in. Let us begin with the training.			**أود الآن أنْ أوجه إليك بعض الأسئلة العامة عن مبادرة القراءة والحساب للصفوف المبكرة (RAMP) التي تشترك فيها أنت ومدرستك. دعنا نبدأ بالتدريب.**	
18.	Did you attend the 5-day teacher training workshop for **module 1?** If yes, did you attend all the days, more than half of the days, or less than half of the days? **[Read all the options to the teacher; only circle one response.]** ➔ **If yes, skip to 20**	No .... 0 Yes, all days .... 1 Yes, more than half of the days 2 Yes, less than half of the days .. 3 Don't know/Refuse 888	0... 4................................ نعم، كل الأيام 6.................. نعم، أكثر من نصف الأيام 5.................. نعم، أقل من نصف الأيام لا أعرف/رفض الإجابة 888. 888.................	هل حضرت ورشة العمل التدريبية التي امتدت لخمسة أيام **للمساق 1 الأول للتدريب الخاص بالقراءة**؟ في حال كانت الإجابة نعم، هل كنت حاضرًا في كل الأيام، أم لأكثر من نصفها، أم أقل من نصفها؟ **[اقرأ جميع الخيارات على المعلم؛ ثم ضع دائرة حول خيار واحد فقط يمثل الإجابة.]** ➔ **إذا كانت الإجابة نعم، تخطى إلى البند 20**	

19.	Why did you not attend the training? **[Do not read the options; just circle all that apply.]**	I was not teaching at the school 1 I was not notified on time ..... 1 I was ill ..... 1 I did not want to attend ..... 1 Other (specify): ..... 1 _____ _ Don't know/Refuse 888	لم أكن أدرس في هذه المدرسة.....4 لم يتم إعلامي في الوقت المطلوب.....4 كنت مريضاً .....4 لم أكن أرغب في الحضور .....4 أخرى (حدد): .....4 _____ _ عرف/فضل ألا أجيب. 888	لماذا لم تحضر التدريب؟] **قرأ الخيارات؛ فقط اضغط دائرة حول جميع الخيارات التالية تمثل الإجابة.]**	
20.	Did you attend the 5-day teacher training workshop for **module 2?** If yes, did you attend all the days, more than half of the days, or less than half of the days? **[Read all the options to the teacher; only circle one response.]** **If yes, skip to 22.**	No ..... 0 Yes, all days ..... 1 Yes, more than half of the days 2 Yes, less than half of the days ..3 Don't know/Refuse 888	.....0 نعم، كل الأيام .....4 نعم، أكثر من نصف الأيام.....6 نعم، أقل من نصف الأيام .....5 عرف/فضل ألا أجيب 888	هل حضرت ورشة عمل تدريبية على المعلمين لمدة خمسة أيام متتالية (من الاثنين إلى الجمعة) في حال كنت قد حضرت ورشة عمل، هل كنت قد حضرت كل الأيام، أم أكثر من نصفها، أم أقل من نصفها؟ **[قرأ جميع الخيارات على المعلم؛ ثم اضغط دائرة حول خيار واحد فقط يمثل الإجابة.]** **إذا كانت الإجابة نعم، تخطى إلى السؤال 22.**	
21.	Why did you not attend the training? **[Do not read the options; just circle all that apply.]**	I was not teaching at the school 1 I was not notified on time ..... 1 I was ill ..... 1 I did not want to attend ..... 1 Other (specify): ..... 1 _____ _ Don't know/Refuse 888	لم أكن أدرس في هذه المدرسة.....4 لم يتم إعلامي في الوقت المطلوب.....4 كنت مريضاً .....4 لم أكن أرغب في الحضور .....4 أخرى (حدد): .....4 _____ _ عرف/فضل ألا أجيب. 888	لماذا لم تحضر التدريب؟] **قرأ الخيارات؛ فقط اضغط دائرة حول جميع الخيارات التالية تمثل الإجابة.]**	

22.	Did you attend the 5-day teacher training workshop for **module? 3** If yes, did you attend all the days, more than half of the days, or less than half of the days? **[Read all the options to the teacher; only circle one response.]** **If yes, skip to 24**	No ....0 Yes, all days..............................1 Yes, more than half of the days 2 Yes, less than half of the days ..3 Don't know/Refuse...............888	0... نعم، كل الأيام ....4 نعم، أكثر من نصف الأيام..................6 نعم، أقل من نصف الأيام ....5 لا أعرف/رفض الإجابة 888.................	هل حضرت ورشة تدريب المعلمين التي امتدت لخمسة أيام **للمساق 3 (الدليل التدريبي الثالث الخاص بالنوع الاجتماعي والتقييم..)** في حال كانت الإجابة نعم، هل كنت حاضراً في كل الأيام، أم أكثر من نصفها، أم أقل من نصفها؟ **[اقرأ جميع الخيارات على المعلم؛ ثم ضع دائرة حول خيار واحد فقط يمثل الإجابة.]** **إذا كانت الإجابة نعم، تخطى إلى البند 24.**	
23.	Why did you not attend the training? **[Do not read the options; just circle all that apply.]**	I was not teaching at the school 1 I was not notified on time ....1 I was ill ....1 I did not want to attend ....1 Other (specify):.........................1 ______________________________ _ Don't know/Refuse...............888	لم أكن أدرس في هذه المدرسة..............4 لم يتم إبلاغي في الوقت المطلوب..........4 كنت مريضاً ....4 لم أكن أرغب في الحضور ....4 أخرى (حدد): ....4 ______________________________ _ لا أعرف/رفض الإجابة. 888	لماذا لم تحضر التدريب؟ [لا تقرأ الخيارات؛ **فقط ضع دائرة حول جميع الخيارات التي تمثل الإجابة.]**	
	Next I want to get a sense of your overall impression of the RAMP initiative. **Please respond to each of the following statements by indicating whether you strongly agree, agree, are neutral about the statement, disagree or strongly disagree.**			**أودّ الآن التعرف على انطباعك العام حول مبادرة الـRAMP.** **أرجو الإجابة عما يلي بموافق بشدة، أو موافق، أو محايد، أو غير موافق، أو غير موافق بشدة .**	

24.	I understand the goals of the RAMP initiative. **[Read all the options to the teacher; only circle one response.]**	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse..... 888	أوافق بشدة ..... 4 أوافق..... 6 محايد..... 5 أعارض..... 1 أعارض بشدة..... 3 لا أعرف/رفض الإجابة..... 888	أنا أفهم أهداف مبادرة القراءة والحساب (RAMP). **[اقرأ جميع الخيارات على المستجيب؛ شجع على الإجابة الواحدة فقط التي تمثل أفضل إجابة.]**	
25.	The training I received for the RAMP initiative was adequate. **[Read all the options to the teacher; only circle one response.]**	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse..... 888	أوافق بشدة ..... 4 أوافق..... 6 محايد..... 5 أعارض..... 1 أعارض بشدة..... 3 لا أعرف/رفض الإجابة..... 888	كان التدريب الذي تلقيته من أجل مبادرة الـRAMP كافيًا. **[اقرأ جميع الخيارات على المستجيب؛ شجع على الإجابة الواحدة فقط التي تمثل أفضل إجابة.]**	
26.	I feel confident about implementing the routines of the RAMP initiative in my class. **[Read all the options to the teacher; only circle one response.]**	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse..... 888	أوافق بشدة ..... 4 أوافق..... 6 محايد..... 5 أعارض..... 1 أعارض بشدة..... 3 لا أعرف/رفض الإجابة..... 888	أشعر بالثقة والتمكن حول تطبيق أنشطة مبادرة الـRAMP في صفي. **[اقرأ جميع الخيارات على المستجيب؛ شجع على الإجابة الواحدة فقط التي تمثل أفضل إجابة.]**	
27.	Did you get RAMP certificate?	Yes ..... 1 No ..... 2 Didn't attend the exam..... 3	نعم..... 1 2 لم أختبر..... 3	هل حصلت على شهادة مبادرة القراءة والحساب؟	

28.	Sufficient guidance and support was provided by my coach and/or supervisor to help me implement the activities of the RAMP initiative. **[Read all the options to the teacher; only circle one response.]**	Strongly agree .... 1 Agree.. 2 Neutral .... 3 Disagree .... 4 Strongly disagree .... 5 Don't know/Refuse 888	4.................................... أوافق بشدة 6... أوافق 5... محايد 1.. لا أوافق 3................................. لا أوافق بشدة 888................ لا أعرف/أرفض الإجابة.	حصلت على دعم وتوجيه كافيين من قبل مشرفي التربوي و/أو مُدربي لمساعدتي في تطبيق أنشطة المبادرة. **[اقرأ جميع الخيارات على المعلم؛ ثم ضع دائرة حول خيار واحد فقط يمثل الإجابة.]**	
29.	The RAMP initiative has positively impacted student achievement. **[Read all the options to the teacher; only circle one response.]**	Strongly agree .... 1 Agree.. 2 Neutral .... 3 Disagree .... 4 Strongly disagree .... 5 Don't know/Refuse 888	4.................................... أوافق بشدة 6... أوافق 5... محايد 1.. لا أوافق 3................................. لا أوافق بشدة 888................ لا أعرف/أرفض الإجابة.	لقد كان للمبادرة RAMP- تأثير إيجابي على مستوى التحصيل لدى الطلاب. **[اقرأ جميع الخيارات على المعلم؛ ثم ضع دائرة حول خيار واحد فقط يمثل الإجابة.]**	
30.	Students in this school are more enthusiastic about learning because of the RAMP initiative. **[Read all the options to the teacher; only circle one response.]**	Strongly agree .... 1 Agree.. 2 Neutral .... 3 Disagree .... 4 Strongly disagree .... 5 Don't know/Refuse 888	4.................................... أوافق بشدة 6... أوافق 5... محايد 1.. لا أوافق 3................................. لا أوافق بشدة 888................ لا أعرف/أرفض الإجابة.	يبدي الطلاب في هذه المدرسة حماساً أكثر للتعلم بسبب مبادرة RAMP-. **[اقرأ جميع الخيارات على المعلم؛ ثم ضع دائرة حول خيار واحد فقط يمثل الإجابة.]**	

31.	The materials provided to implement the RAMP initiative were sufficient. **[Read all the options to the teacher; only circle one response.]**	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse..... 888	أوافق بشدة ..... 4 أوافق..... 6 محايد..... 5 لا أوافق..... 1 لا أوافق بشدة..... 3 لا يعرف أو رفض الإجابة. 888	المواد المزودة لتطبيق مبادرة الـ RAMP كانت كافية . [اقرأ جميع الخيارات على المستجيب؛ ضع دائرة حول خيار واحد فقط يمثل الإجابة.]	
32.	Did RAMP materials reach school on-time (e.g. workbooks, teacher guide…)	Yes ..... 1 No ..... 2 I don't know..... 3	نعم..... 1 2 لا يعرف..... 3	هل وصلت المواد في وقتها؟ (على سبيل المثال كراسة الأنشطة للطلاب، دليل المعلم للمعلم، الخ....)	
	Next I would like to ask you some more specific questions about the RAMP initiative. Let us begin with the reading			**أودّ الآن أن أطرح عليك بعض الأسئلة المحددة حول مبادرة الـ RAMP. لنبدأ بالقراءة.**	
33.	The RAMP initiative has supported me in the teaching of the reading curriculum. **[Read all the options to the teacher; only circle one response.]**	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse..... 888	أوافق بشدة ..... 4 أوافق..... 6 محايد..... 5 لا أوافق..... 1 لا أوافق بشدة..... 3 لا يعرف/رفض الإجابة. 888	دعمني برنامج الـ RAMP في تدريس مهارات القراءة. [اقرأ جميع الخيارات على المستجيب؛ ضع دائرة حول خيار واحد فقط يمثل الإجابة.]	

34.	The RAMP initiative and training has improved my teaching of reading. **[Read all the options to the teacher; only circle one response.]**	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse 888	أوافق بشدة ..... 4 أوافق ..... 6 محايد ..... 5 لا أوافق ..... 1 لا أوافق بشدة ..... 3 لا أعرف أو رفض الإجابة 888	تحسّنت مهارتي في تدريس القراءة بعد تدريبي على برنامج RAMP. [اقرأ جميع الخيارات على المعلم؛ ثم ضع دائرة حول الخيار واحد فقط يمثل الإجابة.]	
35.	The RAMP initiative has improved the reading performance of students in my class. **[Read all the options to the teacher; only circle one response.]**	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse 888	أوافق بشدة ..... 4 أوافق ..... 6 محايد ..... 5 لا أوافق ..... 1 لا أوافق بشدة ..... 3 لا أعرف/رفض الإجابة 888	تحسّنت مهارتي في تدريس القراءة بعد تدريبي على برنامج RAMP. [اقرأ جميع الخيارات على المعلم؛ ثم ضع دائرة حول الخيار واحد فقط يمثل الإجابة.]	
	Let us turn our attention to mathematics			**دعنا الآن نتحدث عن مادة الحساب**	
36.	The RAMP initiative has supported me in the teaching of the mathematics curriculum. **[Read all the options to the teacher; only circle one response.]**	Strongly agree ..... 1 Agree..... 2 Neutral ..... 3 Disagree ..... 4 Strongly disagree ..... 5 Don't know/Refuse 888	أوافق بشدة ..... 4 أوافق ..... 6 محايد ..... 5 لا أوافق ..... 1 لا أوافق بشدة ..... 3 لا أعرف أو رفض الإجابة 888	لقد ساعدني برنامج RAMP في تدريس مادة الحساب في الفصل. [اقرأ جميع الخيارات على المعلم؛ ثم ضع دائرة حول الخيار واحد فقط يمثل الإجابة.]	

37.	The RAMP initiative and training has improved my teaching of mathematics. **[Read all the options to the teacher; only circle one response.]**	Strongly agree .... 1 Agree.. 2 Neutral .... 3 Disagree .... 4 Strongly disagree .... 5 Don't know/Refuse............... 888	4................................... أوافق بشدة 6.. أوافق 5.. محايد 1....................................... لا أوافق 3................................ لا أوافق بشدة 888............... لا أعرف/رفض الإجابة.	حسنت مبادرة الـ-RAMP وتدريبها من طريقتي في تدريس الحساب. **[اقرأ جميع الخيارات على المعلم؛ ثم ضع دائرة حول خيار واحد فقط يمثل الإجابة.]**	
38.	The RAMP initiative has improved the mathematics performance of students in my class. **[Read all the options to the teacher; only circle one response.]**	Strongly agree .... 1 Agree.. 2 Neutral .... 3 Disagree .... 4 Strongly disagree .... 5 Don't know/Refuse............... 888	4................................... أوافق بشدة 6.. أوافق 5.. محايد 1....................................... لا أوافق 3................................ لا أوافق بشدة 888............... لا أعرف/رفض الإجابة.	تمكنت مبادرة الـ-RAMP من تحسين مستوى أداء الطلبة في الحساب في صفي.	
	We are almost finished, I want to ask you about the support that you received from the supervisor.			**وشكنا على الانتهاء؛ أريد الآن أن أسألك عن الدعم الذي حصلت عليه من قبل المشرف.**	
39.	How often did the coach and/or supervisor visit your class after you received training? Once a week, once every two weeks, once a month, only once or twice in the semester, or not at all. **[Read all the options to the teacher; only circle one response.]**	Once a week.............................. 1 Once every two weeks .... 2 Once a month .... 3 Once or twice a semester .... 4 Not at all .... 5 Don't know/Refuse............... 888	4............................. مرة في الأسبوع 6............................ مرة كل أسبوعين 5................................ مرة في الشهر 1......... مرة أو مرتان في الفصل الدراسي 3.................................. لم يزر أبداً 888............... لا أعرف/رفض الإجابة.	كم عدد المرات التي قام فيها المشرف و/أو المدرب بزيارة صفك بعد تلقي التدريب؟ مرة في الأسبوع، مرة كل أسبوعين، مرة في الشهر، مرة أو مرتين في الفصل الدراسي أو لم يزرك أبداً. **[اقرأ جميع الخيارات على المعلم؛ ثم ضع دائرة حول خيار واحد فقط يمثل الإجابة.]**	

40.	In terms of the number of support visits that you received from the supervisor, did the coach and/or supervisor visit you: too often, just the right number of times, too seldom? **[Read all the options to the teacher; only circle one response.]**	Too often.................................1 Just right..................................2 Too seldom ....3 Don't know/Refuse...............888	4................................... متكررة جداً 6... كافية 5..................................... نادرة جداً 888................ لا أعرف/أرفض الإجابة.	بالنسبة لعدد زيارات الدعم من قبل المشرف لحصتك، هل كانت زيارات المدرب و/أو المشرف: متكررة جداً أم كافية أم نادرة جداً؟ **[اقرأ جميع الخيارات على المعلم؛ ثم ضع دائرة حول خيار واحد فقط يمثل الإجابة.]**	
41.	When the coach and/or supervisor visited you, did he/she give you feedback about your implementation of the reading and mathematics routines? If yes, was the feedback very helpful, helpful, neutral, not helpful or very unhelpful? **[Read all the options to the teacher; only circle one response.]**	No ....1 Yes, very helpful......................2 Yes, helpful..............................3 Yes, neutral ....4 Yes, not helpful.........................5 Yes, very unhelpful...................6 Don't know/Refuse...............888	4.. 6................................ نعم، مفيدة جداً 5.................................... نعم مفيدة 1................................... نعم، محايد 3............................... نعم، غير مفيدة 2.............. نعم، غير مفيدة على الإطلاق 888.............. لا أعرف/أرفض الإجابة.	هل قام المشرف و/أو المدرب عند زيارات لك في الصف بإعطائك التغذية الراجعة حول تطبيقك أنشطة القراءة والحساب؟ إذا كانت الإجابة نعم، هل كانت التغذية الراجعة مفيدة جداً أم مفيدة أم محايد أم غير مفيدة أم غير مفيدة على الإطلاق؟ **[اقرأ جميع الخيارات على المعلم؛ ثم ضع دائرة حول خيار واحد فقط يمثل الإجابة.]**	
42.	How do you respond to the following statement: I felt encouraged to communicate concerns, questions, and constructive ideas regarding the Early Grade Mathematics Initiative (RAMP) to the coach and/or supervisor and my head teacher? **[Read all the options to the teacher; only circle one response.]**	Strongly agree...........................1 Agree...2 Neutral ....3 Disagree ....4 Strongly disagree ....5 Don't know/Refuse...............888	4.................................... أوافق بشدة 6.. أوافق 5.. محايد 1...................................... لا أوافق 3................................ لا أوافق بشدة 888............ لا أعرف أو أرفض الإجابة.	كيف ترد على الجملة التالية: "كنت متشجعاً في توصيل مخاوفي وأسئلتي وأفكاري البنّاءة حول مبادرة القراءة والحساب للمشرف و/أو مدير المدرسة."؟ **[اقرأ جميع الخيارات على المعلم؛ ثم ضع دائرة حول خيار واحد فقط يمثل الإجابة.]**	
	Here are the last questions about RAMP			**فيما يلي الأسئلة الأخيرة حول مبادرة الـRAMP**	

43.	What are the **overall** aspects that you think are positive? **[Do not read the options; just circle all that apply.]**	The initiative had positive impact on learning ..... 1 Activities support learning..... 1 Development of thinking skills. 1 Improvement of student skills .. 1 Improvement of teaching skills 1 Training..... 1 Supervisor support (visits, feedback and monthly meetings) ..... 1 Encouragement of school and/or district ..... 1 Parents enjoyed the project..... 1 Other ..... 1 Don't know/Refuse..... 888	كان للمبادرة تأثير إيجابي على عملية التعلم..... 1 تدعم الأنشطة عملية التعلم.....1 تنمية مهارات التفكير.....1 تحسن مهارات الطلبة.....1 تحسن مهارات التدريس.....1 التدريب.....1 دعم المشرف (الزيارات، التغذية الراجعة للوالقاءات الشهرية).....1 تشجيع المدرسة و/أو المديرية.. .....1 لقد استمتع أولياء الأمور بهذا المشروع .....1 الأنشطة المعلمين.....1 كراسة القراءة.....1 كراسة الحساب.....1 غير ذلك .....1 عرف/أفضل عدم الإجابة.....888	ما الجوانب الإيجابية بشكل عام من وجهة نظرك؟ **[قرأ الخيارات؛ فقط بعض الاستجابات حول الخيارات المسجلة.]**	

44.	What are the aspects with respect to **reading** that you think are positive? **[Do not read the options; just circle all that apply.]**	Students enjoyed the activities.. 1 Activities and materials support the curriculum ..... 1 Materials (teacher notes, lesson notes, workbooks)..... 1 Other ..... 1 Don't know/Refuse 888	لقد استمتع التلاميذ بالأنشطة .....1 لقد دعمت الأنشطة والمواد المنهاج .....1 المواد (مذكرات المعلم، كراسات الطلبة، الدوس) .....1 غير ذلك .....1 عرف/رفض الإجابة 888	ما الجوانب التي تعتقد بأنها إيجابية فيما يخص **القراءة؟** **[قرأ الخيارات فقط وضع دائرة حول جميع الخيارات المناسبة.]**	
45.	What are the aspects with respect to **mathematics** that you think are positive? **[Do not read the options; just circle all that apply.]**	Students enjoyed the activities.. 1 Activities and materials support the curriculum ..... 1 Materials (teacher notes, lesson notes, workbooks)..... 1 Other ..... 1 Don't know/Refuse 888	لقد استمتع التلاميذ بالأنشطة .....1 لقد دعمت الأنشطة والمواد المنهاج .....1 المواد (مذكرات المعلم، كراسات الطلبة، الدوس) .....1 غير ذلك .....1 عرف/رفض الإجابة 888	ما الجوانب التي تعتقد بأنها إيجابية فيما يخص الحساب؟ **[قرأ الخيارات فقط وضع دائرة حول جميع الخيارات المناسبة.]**	

46.	What are the **overall** aspects that you think are frustrating or negative? **[Do not read the options; just circle all that apply.]**	Objectives of project and materials not clear.................... 1 Increased workload for teachers1 Training (dates/timing; duration; arrangements) .... 1 Training (content, presentation) 1 Supervisors not providing support (visits and feedback) .... 1 Supervisors creating confusion (feedback) .... 1 Initiative did not have a positive impact on learning .... 1 Insufficient encouragement and support, no reward .... 1 Too much time/effort required to mark the student workbooks..... 1 Teachers transferring in and out of school................................... 1 Other .... 1 Don't know/Refuse 888	لم تكن أهداف ومواد المشروع واضحة 1...... زيادة عبء العمل على المعلمين 1............. التدريب (المواعيد/المدة الزمنية، الترتيبات) .... 1 التدريب (المحتوى، العرض) 1............... يقوم المشرفون بتقديم الدعم (الزيارات والتغذية الراجعة) .... 1 يُحدِث المشرفون حالة من الارتباك (التغذية الراجعة) .... 1 لم يكن للمبادرة تأثير إيجابي على عملية التعلم.. 1 عدم وجود الدعم والتشجيع الكافي، وانعدام الحوافز .... 1 الحاجة لجهد ووقت كبيرين لمتابعة كراسة الطلبة وتصحيحها .... 1 تنقل المعلمين داخل وخارج المدرسة 1...... غير ذلك .... 1 لا أعرف/رفض الإجابة 888	ما الجوانب بالمجمل التي تعتقد بأنها كانت مُحبطة أو سلبية؟ **[لا تقرأ الخيارات فقط ضع دائرة حول جميع الخيارات المناسبة.]**	

47.	What are the aspects with respect to **reading** that you think are frustrating or negative? **[Do not read the options; just circle all that apply.]**	Activities did not support the curriculum ..... 1 Activities too time consuming/too many for each day ..... 1 Activities too difficult ..... 1 Activities too easy ..... 1 Students did not enjoy the activities ..... 1 Other ..... 1 Don't know/Refuse 888	تعلم الأنشطة لا تدعم المناهج ..... 1 تستغرق الأنشطة وقتاً طويلاً/أو أكثر من اللازم كل يوم ..... 1 الأنشطة صعبة جداً ..... 1 الأنشطة سهلة جداً ..... 1 لا يستمتع الطلاب بالأنشطة ..... 1 غير ذلك ..... 1 عرف/فضل بدون إجابة 888	ما الجوانب التي تعتقد بأنها مُحبطة أو سلبية فيما يخص **القراءة**؟ **[قرأ الخيارات، فقط ضع دائرة حول جميع الخيارات المناسبة.]**	
48.	What are the aspects with respect to **mathematics** that you think are frustrating or negative? **[Do not read the options; just circle all that apply.]**	Activities did not support the curriculum ..... 1 Activities too time consuming/too many for each day ..... 1 Activities too difficult ..... 1 Activities too easy ..... 1 Students did not enjoy the activities ..... 1 Other ..... 1 Don't know/Refuse 888	تعلم الأنشطة لا تدعم المناهج ..... 1 تستغرق الأنشطة إلى الكثير من الوقت/أي الكثير من اللازم كل يوم ..... 1 الأنشطة صعبة جداً ..... 1 الأنشطة سهلة جداً ..... 1 لا يستمتع الطلاب بالأنشطة ..... 1 غير ذلك ..... 1 عرف/فضل بدون إجابة 888	ما الجوانب التي تعتقد بأنها مُحبطة أو سلبية فيما يخص **الحساب**؟ **[قرأ الخيارات فقط ضع دائرة حول جميع الخيارات المناسبة.]**	

	Finally, I would like to ask you about how you assess and monitor student progress.			أخيراً، أود أن أسألك عن كيفية تقييمك وتتبعك لتقدم أداء الطلبة.	
49.	How do you check for student understanding during the lesson? **Do NOT READ the options. Tick ALL that apply.**	Ask comprehension questions to individual students ..... 1 Ask comprehension questions to whole class ..... 1 Ask comprehension questions to students in groups ..... 1 Give students a task and correct the responses before the end of the lesson ..... 1 Give students a task and correct the responses after the end of the lesson ..... 1 Don't know/Refuse 888	أسئلة فهم للطلبة بشكل فردي ..... 1 أسئلة فهم للصف كامل ..... 1 أسئلة فهم للمجموعات في الطابور ..... 1 مهام وفحص الإجابات قبل نهاية الحصة ..... 1 مهام وفحص الإجابات بعد نهاية الحصة ..... 1 لا ينطبق/رفض الإجابة 888	كيف تتحقق من مدى فهم الطلبة أثناء الدرس؟ **قرأ الخيارات.** **تحقق بما ينطبق.**	
50.	How do you use the results of students' oral and written assessments in your teaching? **Do NOT READ the options. Tick ALL that apply.**	Grade students ..... 1 Evaluate students' understanding of subject matter ..... 1 Plan teaching and learning activities ..... 1 Adapt teaching to better suit students' needs ..... 1 Arrange students in ability groups ..... 1 Other ..... 1 Don't know/Refuse 888	أعمال الطلبة ..... 1 تقييم فهم المادة للطلبة ..... 1 أخطط الدروس التالية بناءً على نتائج التقييمات ..... 1 أعمل على تغيير طريقة تدريسي بما يتناسب مع حاجات الطلبة ..... 1 أجمع الطلبة في مجموعات حسب مستوى قدرتهم ..... 1 أخرى ..... 1 لا ينطبق/رفض الإجابة 888	كيف تستخدم نتائج التقييمات الشفوية والكتابية في تدريسك؟ **قرأ الخيارات.** **تحقق بما ينطبق.**	

51.	In your class, how many parents / guardians review students' homework? None, some, most or all? **Read the responses.** **Tick only ONE response.**	None.....0 Some .....1 Most .....2 All .....3 Don't know/refuse 888	0.....حد 1.....بعض 2.....معظمهم 3.....جميعهم 888.....عرف/فضل بعة	في صفك، كم عدد أولياء امور الذين يراجعون الواجبات المنزلية للطلبة؟ أحد، أبعض هم، أم معظمهم، أم جميعهم؟ **قرأ الخيارات.** **وضح بجانب الملك بما ينطبق.**	
52.	Are you generally satisfied with parents' involvement in their children's schoolwork?	No .....0 Yes .....1 Don't know/Refuse.....888	0..... 1.....نعم 888.....عرف/فضل بعة	هل أنت بشكل عام راض عن مشاركة اهل في اعمال المدرسية التي تخص طفلهم؟	
53.	Ending time **[Use 24-hour time HH:MM]**	____/____	____/____	وقت النهاية **[استخدم نظام الـ 24 ساعة – س س : دد]**	
	Thank you very much.			**شكراً جزيلاً .**	

Student Questionnaire

RAMP Initiative, midline survey: Pupil questionnaire

1.	Starting time **[Use 24-hour time HH:MM**	__*___:___*__	__*___:___*__	وقت البدء **[استخدم توقيت 24 ساعة – س س:دد]**	1.
2.	Interview date **[DD/MM/YY]**	__*___/__*___/__*___	__*___/__*___/__*___	تاريخ المقابلة [اليوم/الشهر/السنة]	2.
3.	Interview status	Refused ..... 1 **→ Thank student and end interview** Partially completed ..... 2 Completed 3	رفض المقابلة ..... 1 **← شكر الطالب وأنهِ المقابلة** تم استكمال جزئي ..... 2 تم استكمال كامل 3	حالة المقابلة	3.
4.	**Gender**	Boy ..... 1 Girl 2	ذكر ..... 4 أنثى 6	**جنس الطالب**	4.
5.	How old are you?	Range: 5-12 Years..... _____	الفئة العمرية: 3-46 السنوات ..... _____	كم عمرك؟	5.
6.	What grade are you in? **[Note: If not in [grade being assessed], thank student and explain that you are only assessing [grades being assessed]**	Grade **2** ..... 1 Grade **3** 2	الصف **4** ..... 1 الصف **3** 2	ما صفك؟ **[ملحوظة: إذا كان الطالب من الصف الخاضع للتقييم، اشكر الطالب واشرح له أنك تقوم بتقييم الصف الخاضع للتقييم فقط.]**	6.

7.	What grade were you in last year? **[Do <u>not</u> verify by asking if child is repeating]**	Grade 1 ..... 1 Grade 2 ..... 2 Grade 3 ..... 3 Don't know/Refuse 888	الصف 1 ..... 1 الصف 2 ..... 2 الصف 3 ..... 3 عرف/رفض ا ب ة 888	ي أي صف كنت في العام الماضي؟ **[تحقق من الطالب إن كان معيداً الصف أم]**	7.
8.	Did you go to preschool or kindergarten?	No ..... 0 Yes ..... 1 Don't know/Refuse 888	0 ..... نعم ..... 1 عرف/رفض ا ب ة 888	هل التحقت بالروضة أو الحضانة؟	8.
	Now I would like to ask you some questions about your class.			**أود أن أطرح عليك بعض الأسئلة عن صفك.**	
9.	Do you have time to read books in your classroom or in your school library every day?	No ..... 0 Yes ..... 1 Don't know/Refuse 888	0 ..... نعم ..... 1 عرف/رفض ا ب ة 888	هل لديك وقت لقراءة كتب داخل صفك أو المكتبة المدرسية كل يوم؟	9.
10.	Do you bring home reading books from your classroom or from the school library?	No ..... 0 Yes ..... 1 Don't know/Refuse 888	0 ..... نعم ..... 1 عرف/رفض ا ب ة 888	هل تحضر معك إلى المنزل كتب القراءة من صفك أو المكتبة المدرسية؟	10.
	Now I would like to ask you some questions about your household.			أود ا ن أن أطرح عليك بعض الأسئلة عن منزلك.	
11.	During the week, does someone at home read to you? **If yes,** how often?	No, never ..... 0 Yes, once a week ..... 1 Yes, 2-3 times per week ..... 2 Yes, every day ..... 3 Don't know/Refuse 888	لا، أبداً ..... 0 نعم، مرة واحدة في ا ب وع ..... 1 نعم، 2-5 مرات في ا ب وع ..... 2 نعم، يومياً ..... 3 عرف/رفض ا ب ة 888	ل ا ب وع، هل يوجد شخص ما في المنزل يقرأ لك؟ **إذا كانت ا جابة نعم،** كم مرة تقريباً؟	11.

12.	Does someone at home help you with your homework when you need it?	No ..... 0 Yes ..... 1 Don't know/Refuse 888	0 ..... 1 ..... نعم 888 ..... عرف/رفض الإجابة	هل يساعدك أحد في المنزل عندما تحتاج إلى المساعدة في واجباتك المدرسية؟	12.
13.	Do you ever meet with other children in the community and listen to somebody read a story to you? **If yes**, how often?	No, never ..... 0 ➔ If no, skip to 18 Yes, just once ..... 1 Yes, not often/occasionally ..... 2 Yes, every week ..... 3 Don't know/Refuse 888	0 ..... أبداً، إذا كانت الإجابة ➔ انتقل إلى السؤال 18 1 ..... نعم، مرة واحدة 2 ..... نعم، ليس غالباً/أحياناً 3 ..... نعم، أسبوعياً 888 ..... عرف/رفض الإجابة	هل حدث وقت التقيت فيه مع أطفال آخرين في المجتمع المحلي واستمعت معهم وهو يقرأ لك قصة؟ إذا كانت الإجابة نعم، كم مرة؟	15.
14.	When you met with the other children in the community and listened to somebody read a story to you, did the person who read the story to you allow you to borrow a book and take it home?	No ..... 0 Yes ..... 1 Don't know/Refuse 888	0 ..... 1 ..... نعم 888 ..... عرف/رفض الإجابة	عندما التقيت مع أطفال آخرين من المجتمع المحلي واستمعت معهم وهو يقرأ لك قصة، هل سمح لك هذا الشخص باستعارة أي من الكتب وأخذها معك إلى المنزل؟	11.
15.	Are your parents: from Jordan, from Syria, from Iraq or from another country **Tick only ONE response.**	Jordan ..... 0 Syria ..... 1 Iraq ..... 2 Other ..... 3 Don't know/Refuse 888	0 ..... الأردن 1 ..... سوريا 2 ..... العراق 3 ..... أخرى 888 ..... عرف/رفض الإجابة	هل والداك من: الأردن، سوريا، العراق أو من بلدٍ آخر؟ **اختر إجابة واحدة فقط.**	13.
16.	Does your mother know how to read?	No ..... 0 Yes ..... 1 Don't know/Refuse 888	0 ..... 1 ..... نعم 888 ..... أعرف/رفض الإجابة	هل تستطيع والدتك القراءة؟	12.

<table>
<tr><td>17.</td><td>Does your father know how to read?</td><td>No 0
Yes.. 1
Don't know/Refuse .... 888</td><td>0 ....
1 ....نعم
888 .... لا أعرف/رفض الإجابة</td><td>هل يستطيع والدك القراءة؟</td><td>41.</td></tr>
<tr><td>18.</td><td>Does your family have the following items in your home?

Read answer options aloud.

🕮 Point to appropriate pictograms.</td><td colspan="2">
<table>
<tr><td></td><td>No</td><td>Yes</td><td>--</td><td>--</td><td>نعم</td><td></td><td></td></tr>
<tr><td>Dishwasher</td><td>0</td><td>1</td><td>888</td><td>888</td><td>1</td><td>0</td><td>غسالة صحون</td></tr>
<tr><td>Vehicle</td><td>0</td><td>1</td><td>888</td><td>888</td><td>1</td><td>0</td><td>سيارة</td></tr>
<tr><td>Computer</td><td>0</td><td>1</td><td>888</td><td>888</td><td>1</td><td>0</td><td>جهاز حاسوب</td></tr>
<tr><td>Laptop computer / Tablet</td><td>0</td><td>1</td><td>888</td><td>888</td><td>1</td><td>0</td><td>حاسوب محمول/حاسوب لوحي</td></tr>
<tr><td>Air conditioner</td><td>0</td><td>1</td><td>888</td><td>888</td><td>1</td><td>0</td><td>مكيف هوائي</td></tr>
<tr><td>Microwave</td><td>0</td><td>1</td><td>888</td><td>888</td><td>1</td><td>0</td><td>ميكروويف</td></tr>
</table>
</td><td>هل تمتلك عائلتك أي من الحاجيات التالية في المنزل؟

اقرأ خيارات الإجابة بصوت عالٍ.

🕮 أشر إلى الصورة المناسبة.</td><td>48.</td></tr>
<tr><td>19.</td><td>Where do you normally get your water from at home?

Read answer options aloud.
🕮 Point to appropriate pictograms.
Tick only ONE response.</td><td>Well or borehole 4
Communal tap 6
Water pipe / tap in your home 5
Water truck or tank............................1
Other..3
Don't know/Refuse ....888</td><td>4.. بئر ماء
6.............................. صنبور ماء عمومي
5..................أنبوب مياه/صنبور ماء في المنزل
1................................ خزان/صهريج ماء
3.. أخرى
888........................ لا أعرف/رفض الإجابة</td><td>من أين تحصل على الماء في منزلك عادة؟

اقرأ خيارات الإجابة بصوت عالٍ.

🕮 أشر إلى الصورة المناسبة.

اختر إجابة واحدة فقط.</td><td>49.</td></tr>
<tr><td>20.</td><td>Ending time
[Use 24-hour time HH:MM</td><td>*___ : ___*</td><td>*___ : ___*</td><td>وقت النهاية
[استعمل توقيت 24 ساعة – س س:دد]</td><td>60.</td></tr>
<tr><td></td><td>Thank you very much!</td><td></td><td></td><td>شكراً جزيلاً!</td><td></td></tr>
</table>

REVIEW OF USAID/JORDAN RAMP SURVEY AND IMPACT EVALUATION

FINAL REPORT

March 2019

This publication was prepared independently by Social Impact, Inc. at the request of the United States Agency for International Development. This publication is part of the Middle East Education, Research, Training, and Support (MEERS) activity.

REVIEW OF USAID/JORDAN RAMP SURVEY AND IMPACT EVALUATION
FINAL REPORT

March 2019

IDIQ: AID-OAA-I-14-00075

Task Order: AID-OAA-TO-17-00022 (Middle East Education Research, Training, and Support)

Submitted to:

Christine Capacci-Carneal, Contracting Officer's Representative
USAID/Middle East Bureau

CONTENTS

TABLES AND FIGURES

ACRONYMS & ABBREVIATIONS

AY	Academic Year
BL	Baseline
EGMA	Early Grade Mathematics Assessment
EGRA	Early Grade Reading Assessment
EL	Endline
MEERS	Middle East Education Research, Training, and Support
ML	Midline
MoE	Ministry of Education
MSI	Management Systems International
OLS	Ordinary Least Squares
PSM	Propensity Score Matching
PSW	Propensity Score Weight/Weighting
RAMP	Reading and Mathematics Project
RTI	Research Triangle International
SI	Social Impact, Inc.
USAID	United States Agency for International Development

INTRODUCTION

From 2015-2019, RTI International (RTI) implemented the Early Grade Reading and Mathematics Project (RAMP), supported by the United States Agency for International Development (USAID) and the United Kingdom's Department for International Development, in partnership with the Jordan Ministry of Education (MoE). RAMP aimed to improve reading and mathematics skills of Jordanian students from Kindergarten 2 through Grade 3 through a coordinated, national effort to distribute improved learning materials, build the capacity of educators and education administrators, promote community participation, and support adoption of relevant national policies, standards, curricula, and assessments. Although RAMP was implemented nationwide, it was rolled out sequentially in three sets of governorates or "cohorts," which produced periods of time that some governorates were participating in RAMP while others had yet to begin their participation. RTI conducted a midline survey to evaluate the program, and Management Systems International (MSI) conducted an independent impact evaluation of the program. Overall, these studies arrived at conflicting conclusions about the program's impact.

Through the Middle East Education Research, Training, and Support (MEERS) mechanism, USAID asked Social Impact (SI) to assess the validity of these two evaluations. The purpose of SI's assessment is to provide an independent, expert opinion to help USAID better understand what it should learn from these evaluations, how to reconcile differences in their findings, and how to use information from these assessments to improve its early grade reading and math programming in Jordan. This report begins with a summary of the methods, limitations, and findings of both RTI's and MSI's studies. We then provide our interpretation and reconciliation of the divergences. The following section explains additional analyses we conducted to test the sensitivity of MSI's findings to alternative models. We end with our conclusions, areas for further inquiry, and recommendations for future evaluations.

RTI'S MIDLINE SURVEY

SUMMARY OF METHODS

RTI's midline survey was nationally representative, using a stratified random sampling technique that ensured equal representation from all twelve governorates in the country. At the request of USAID and the MoE, RTI used data from a national survey for the 2013-2014 Intervention Pilot Research Activity to serve as a baseline for the RAMP midline survey. The full sample for the Intervention Pilot Research Activity survey included 156 schools and was nationally representative. From this full sample, RTI purposively selected 46 "treatment" schools from this group, excluding schools in districts that had fewer than three supervisors, schools that were participating in existing projects, and schools that were too far from each other for the MoE to provide adequate support. This set of treatment schools were excluded in the data RTI used to estimate the baseline for RAMP, under the assumption that the treatment schools benefited from the Pilot activity and were no longer representative as a baseline for the RAMP schools. Accordingly, this left a sample size of 2,159 students from 110 "control" schools to serve as the baseline for RAMP.

RTI confirmed that there were differences in average reading and math assessment scores between the Pilot control schools (or RAMP baseline schools) and a nationally representative sample, such that the average scores for the 110 "control" schools might be slightly biased downward from the true 2014 national average. To be clear, the data used to generate the baseline comes from a different sample of schools than the midline sample, and while the data is sufficient for making point estimates for assessment scores at the national, national by gender, and national by grade-levels, the sample size is not sufficient for generating accurate point estimates at the governorate level.

The midline survey included 4,679 students from 240 randomly selected schools, equally distributed among the twelve governorates. The sample frame included all schools in the country except those that had fewer than 20 G2 and G3 students combined, those that were participating in the MSI impact

evaluation, and 99 that were recently established. As RAMP was rolled out sequentially from 2015-2019 in three "cohorts" corresponding to three sets of governorates, the midline sample included various levels of exposure to RAMP corresponding to a student's grade and cohort. The duration of exposure to RAMP treatment varies by cohort and grade level, detailed in Table 1.

Table 1: Duration of exposure to RAMP treatment in midline survey sample

	Cohort 1		Cohort 2		Cohort 3	
Grade	G2	G3	G2	G3	G2	G3
Semesters of RAMP exposure	3	3	2	0	0	0
Percentage of academic career exposed to RAMP	75%	50%	50%	0%	0%	0%

The midline survey analysis compares the 2014 data from the Intervention Pilot Research Activity control schools to RAMP's own nationally representative midline survey from 2017. This methodology could thus be labeled a national longitudinal or pre-/post-study and enables a mostly defensible definition of the national trend for G2 and G3 student assessment scores from 2014 to 2017, with the possibility that 2014 figures may be biased slightly downward from the true 2014 national average. It also enables a comparison of assessment scores between students with and without exposure to RAMP for 2017 only.

RTI reported the following primary limitations of its methodology:

1. The baseline data only establish a true baseline at the kingdom, grade, and gender levels. It is not possible to assess trends with a high level of precision over time for other disaggregations, such as intervention status or governorate.
2. The weighting approach selected implies that larger non-intervention governorates such as Amman and Irbid can influence the overall mean to such an extent that impact in the intervention governorates may not be easily visible.

We underscore that even after the endline survey (not yet available), RTI's study enables ***only*** descriptive analysis of national, grade, gender, and governorate-level trends. It does not construct a counterfactual, which would permit the reader to understand what portion of the change in assessment scores over time, if any, is attributable to RAMP.

SUMMARY OF FINDINGS

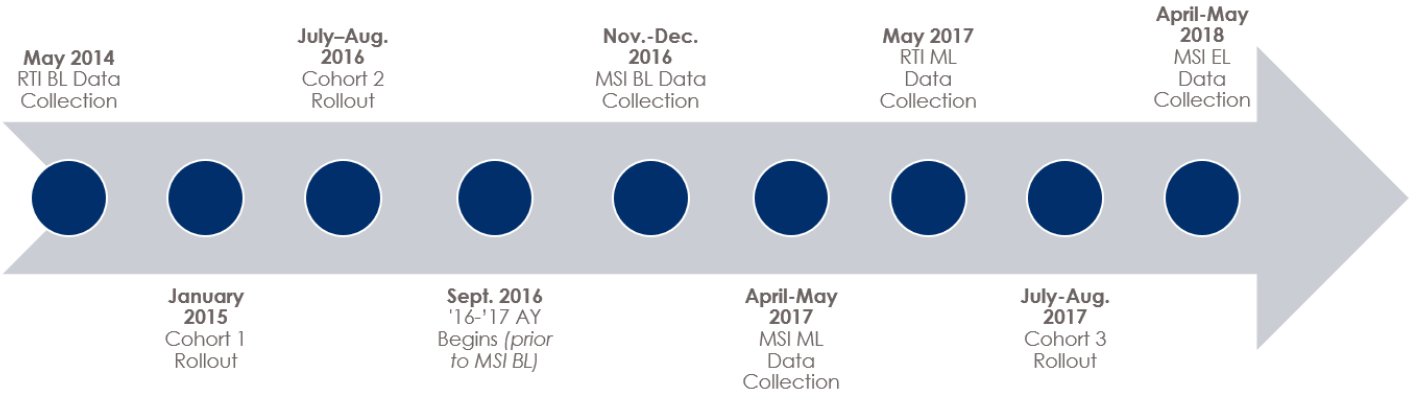
In comparing nationally representative samples of students from 2014 and 2017, RTI's midline survey found a statistically significant improvement in assessment scores for G2 and G3 students on pre-reading skills such as correctly identifying letter sounds, syllable sounds, and reading invented words. Changes in other reading skills, such as oral reading fluency and reading comprehension, were not statistically significant. The survey also found statistically significant improvement in conceptual mathematics skills, like number identification, quantity comparison, and Level 2 addition and subtraction, although other math skills showed insignificant changes. The improvements in reading and math scores were such that G2 students in 2017 often approached the performance of G3 students from 2014 or outperformed them, on average. The increased assessment scores significantly improved the proportion of children able to demonstrate grade-appropriate proficiency in reading and math according to benchmarks established as performance indicators for RAMP. On the question of whether RAMP might be responsible for these improvements, the survey found significantly higher 2017 Early Grade Reading Assessment (EGRA) scores for students in RAMP schools compared to those where RAMP had yet to begin, although baseline data were not available to discern the trends for these two groups over time. **The midline survey report concluded that: "On balance, the RAMP midline survey provides powerful evidence that RAMP is making a positive impact."**

MSI'S IMPACT EVALUATION

SUMMARY OF METHODS

MSI's evaluation used distinct methodologies for each of its evaluation questions. On G1 and G2 learning outcomes and G3 instructional practices, the evaluation relied on quasi-experimental methods to generate causal impact estimates for RAMP. To establish a valid baseline, Cohort 1 was excluded from the study, as implementation was already underway. Schools in governorates in Cohort 2 were considered "intervention" schools, while those in Cohort 3 were considered "comparison" schools. As the cohorts were not randomly selected, MSI used propensity score matching to select schools from Cohort 3 that were equivalent on observable characteristics to schools selected from Cohort 2. Upon discovering at baseline that students from matched intervention and comparison schools still had significantly different baseline assessment scores, MSI used student-lel propensity score weights to more heavily consider students from comparison schools with scores similar to typical students at intervention schools.

Figure 1: RAMP Implementation and Evaluation Timeline[1]



By comparing similar (on observable characteristics) students and teachers at similar treatment and comparison schools, the impact evaluation was able to consider not only how assessment scores changed over time in the treatment schools, but also **how assessment scores might have changed over time in the absence of RAMP (the counterfactual).** The program impact, or change in outcomes of interest that can be attributed to the program, was measured by (i) the "difference-in-differences" of the rate at which intervention and comparison teachers displayed desired instructional practices; and (ii) by an ordinary least squares (OLS) regression coefficient for the treatment term for student assessments, controlling for various other potential explanatory factors, including levels of baseline achievement. In both cases, the result of the analysis was the difference in outcomes of interest for the intervention group relative to the comparison group that can be explained by the program. In the latter case, MSI also tested for their model's sensitivity to excluding baseline levels of achievement from analysis.

MSI listed four core limitations of these quasi-experimental designs:

1. While propensity score matching accounts for observable differences between the intervention and comparison groups, unobservable pre-existing differences between the groups could remain. These unobservable differences could confound the interpretation of results.

[1] Note that "roll-out" refers to the commencement of Module 1 (reading instruction) and Module 2 (math instruction) training for K2 through G2 teachers. Other aspects of RAMP, or training for G3 teachers, were rolled out at later points of the year.

2. "Baseline" data collection occurred six to eight weeks into RAMP exposure for Cohort 2 students, whose teachers completed their module 1 (reading instruction) and 2 (math instruction) in-service training sessions in July before the academic year (AY) commenced (see Figure 1). This means that, for the teacher-level study, the observed instructional practices may have been influenced by this exposure to the program and any such influence would bias impact estimates downward.
3. The evaluation measures impact at a relatively short timeframe of exposure (six months and two years). If the program needed longer than this for impacts to take hold, the evaluation would not capture these impacts. However, the endline results for G1 students notably found only null or negative impacts after two years of program exposure, relative to one year of program exposure.
4. The evaluation confirmed some spillover of the intervention into comparison schools. Contamination of the comparison group can bias impact estimates downward, as both intervention and comparison schools would benefit in equal measure. However, MSI showed that the spillover was minimal and intervention schools received a substantially larger "dosage" (e.g. days of training, number of classroom visits, materials used, etc.)[2] of the program, such that program impacts would be expected to manifest if they did exist.

We observe one additional external validity limitation to MSI's impact evaluation, which is that its findings only apply to the impact of RAMP on Cohort 2 students and teachers. To the extent that conditions existed in Cohort 1 or Cohort 3 that were more conducive to RAMP generating impact, MSI's impact evaluation could not refute that a similar evaluation may have yielded different findings in those settings. However, there would have been no feasible to way to construct a counterfactual using these cohorts as the intervention group with the time MSI had available to conduct its impact evaluation. Cohort 1 was already well into implementation at baseline, and Cohort 3 was the final set of governorates in which RAMP was implemented. Furthermore, Cohort 2 included nearly half of the students in the country, so of the three cohorts, the impacts on Cohort 2 were arguably of the most interest in evaluating RAMP's impact.

SUMMARY OF FINDINGS

MSI's impact evaluation of RAMP was guided by three evaluation questions, which focused on (1) the effect of RAMP on teachers' instructional practices, (2) the impact of RAMP on students' proficiency in reading and math, and (3) differential impacts of RAMP on various sub-groups of interest (e.g. different types of students or schools). After tracking a panel of schools and G1 and G2 students in Cohort 2 and Cohort 3 governorates over about a year and a half, **the evaluation found few positive impacts *caused by RAMP* on teachers' instructional practices or students' proficiency.** RAMP appeared to shift some lesson time away from basic skills like writing activities, identifying written characters, number writing and number identification toward other content, although there was a significant positive effect on the lesson time spent on vocabulary. RAMP caused reading teachers to spend more time on whole-class instruction, although there was no significant impact on student engagement. Conversely, math teachers provided better feedback on student participation and written work with no significant program impacts on classroom structure.

Table 2: Interpretation of MSI student-level impact evaluation results

	Midline	Endline
Grade 1	Impact of 6 months of exposure (treatment)	Impact of 2 years of exposure (treatment)

[2] Specifically, 86% of G3 intervention teachers reported receiving RAMP training compared to 17% of comparison teachers. 80% of trained intervention teachers and 5% of comparison teachers reported that a RAMP coach observed their lesson during the current school year. MSI speculated that the true magnitude of contamination may be even less than these figures if comparison teachers are confusing MoE materials for RAMP materials.

	relative to no exposure (comparison)	relative to 1 year of exposure (comparison)
Grade 2	Impact of 6 months of exposure (treatment) relative to no exposure (comparison)	Impact of 2 years of exposure (treatment) relative to no exposure (comparison)

Regarding student proficiency (see Table 2 for interpretation), MSI's impact evaluation found mostly no impact of RAMP on reading scores, apart from positive impacts on midline G1 passage reading, midline G1 and G2 syllable segmentation, and endline G2 syllable segmentation. The evaluation found negative endline impacts on G1 and G2 letter sound knowledge. It finds mostly no impact of the program on math scores, except for negative impacts on midline G1 counting numbers and endline G1 addition facts. Key MSI results are presented in more detail in Annex II.

INTERPRETING AND RECONCILING FINDINGS

With these methods and limitations in mind, we interpret each evaluation's findings as follows:

RTI found that reading and math scores for G2 and G3 students mostly increased from 2014 to 2017. By statistically equating relevant sections of the tests used, we can be reasonably assured that the improvement is due to improvement in students' capabilities and not changes in the difficulty or students' understanding of the assessment. We can be confident that students in intervention schools in 2017 tended to score higher than peers in non-intervention schools. Given the rollout of RAMP in Cohorts 1 and 2 over this timeframe, the longer exposure of G2 students to RAMP than G3 students, and the particularly large increases in G2 assessment scores, it is a plausible conclusion that the increases observed might be caused by RAMP. However, **there is nothing in RTI's methodology that can establish RAMP as the cause of these increases.** Alternative explanations exist and cannot be refuted by RTI's data, sampling approach, or methods. As one example,[3] the MoE reportedly disseminated a new national curriculum booklet and conducted its own teacher training during the 2016/2017 school year, which could have influenced assessment scores for all students.

MSI found that RAMP had a mostly null or negative effect on instructional practices and student achievement. MSI's impact evaluation notably *does not refute* that assessment scores have increased in Jordan between 2016 and 2017, and even found evidence that scores increased for most students in their sample (see Annex I and Annex II). Rather, the evaluation asserts that comparing the achievement of similar comparison students from similar schools to analog students in analog intervention schools yields little evidence of significant, positive program impact from RAMP, with the exception, perhaps, of syllable segmentation. Indeed, there are a few sub-tasks (letter sounds, counting numbers, and level one addition) where comparison students' scores increased at a faster rate than their treatment peers who had been exposed to one to two years of the RAMP intervention.

We were able to replicate MSI's and RTI's results using their data and analysis files. Both studies adequately qualify confidence intervals within which true population averages may fall and specify for which population their samples are representative. Both studies are adequately powered for their respective purposes.[4] In each case, EGRAs and EGMAs (Early Grade Mathematics Assessments) were equated and/or validated such that the assessments validly measured underlying concepts and any changes in scores represented changes in aptitude, not changes in the difficulty of the test or adaptations in students' understanding of the test.[5] In sum, we have no reason to doubt either RTI's or MSI's findings

[3] According to MSI's Endline Report (p. 24).

[4] Since RTI's study did not aim to quantify program impact, representativeness is more relevant than statistical power. Subject to the limitations specified in their methods section, their study's sample was adequate to define national, grade-disaggregated, and sex-disaggregated differences in scores for their 2014 and 2017 samples. Although MSI did not update its ex-ante power calculation from the design phase in its endline report, Annex B of the endline report notes that attrition was lower than assumed in their "best-case scenario" power calculations which indicated the study would be able to detect an effect size of at least 0.26-0.27 standard deviations if such a program impact existed in fact.

[5] RTI tested the 2014 and 2017 EGRA and EGMA tools for reliability and found that each of the instruments showed good internal consistency, as judged by Cronbach's alpha values (α = 0.86 for EGRA and 0.85 for EGMA for the 2014 assessment; and α = 0.86 for EGRA and 0.91 for EGMA for the 2017

on the basis of mistakes in analysis, inadequate statistical power, or inadequate equating of assessment tools.

In reconciling the differences between RTI and MSI's conclusions, we must consider whether there is a situation that allows their findings to be true simultaneously. Our preliminary conclusion (which closely resembles the conclusions in MSI's memorandum reconciling the two studies), is that **MSI's findings and conclusions regarding the causal impact of RAMP, *over the time period measured*, are most credible, as their methodology was designed to discern these impacts and RTI's methodology was not. There is evidence of an underlying trend in Jordan of increasing assessment scores (see tables ES2 and ES4 of Annex I and tables O.2 through O.6 of Annex II), but causal impact attributable to RAMP is not indicated by either study.** While this conclusion does not negate RTI's findings, it does cast doubt on the suggestion in their midline report that that RAMP was at least a partial cause of these increasing scores. This conclusion favors and accepts MSI's interpretation of its results.

TESTING THE SENSITIVITY OF MSI'S FINDINGS TO ALTERNATIVE SPECIFICATIONS

After our review of RTI's and MSI's respective reports, datasets, and analysis files, we reviewed each entity's response to the other's report (see Annex IV and Annex V for these responses in their entirety). RTI raised five key points critiquing the credibility of MSI's results. These centrally argue that baseline differences in achievement between treatment and comparison groups are evidence of early program impact, not pre-existing differences in aptitude between students in statistically matched intervention and non-intervention schools. RTI requested that MSI's data be analyzed without using student-level propensity score weighting and/or baseline assessment scores as covariates in their model for estimating impact to eliminate these baseline differences. The purpose of this is to determine if the findings change in favor of suggesting program impacts. Prior to reaching final conclusions about the takeaway from these two studies, we believe RTI's argument should be addressed. In this section, we discuss what these analysis strategies aim to achieve, present the results of these alternative analyses, and then discuss how USAID should interpret their results.

PURPOSE OF MSI'S PROPENSITY WEIGHTING AND BASELINE ACHIEVEMENT COVARIATES

Governorates, and thus schools, were assigned to implementation cohorts in a non-random fashion. It is a standard and defensible procedure in this context to use statistical techniques such as propensity score matching and weighting to achieve balance between intervention and comparison groups on outcome variables and independent variables that might influence outcomes. Doing so increases confidence that differences observed after the intervention reflect true program impacts rather than pre-existing differences between the groups. MSI chose to approach this by: (i) applying propensity score matching to data from the MoE's Education Management Information System and their own primary verification data to ensure sampled schools were equivalent on observable characteristics; and (ii) weighting students in their midline and endline analyses based on their propensity to be in the intervention group according to primary baseline data on student characteristics, home environment,

assessment). The only portion of the 2017 assessments issued by RTI that required equating was the oral reading fluency sub-task, since the oral reading passage was changed. Otherwise, baseline ceiling effects on listening comprehension rendered that sub-task redundant in 2017, while the remaining subtasks simply systematically rearranged the items in each line of the stimulus sheet. Thus, these sub-tasks presented a different version of the same assessment and did not require equating. For most sub-tasks on MSI's EGRA and EGMA assessment, common or anchor items were identified from the baseline and midline tools and carried through to future tools to calibrate and adjust for the difficulty of the instruments. There were also anchor items on the G1 and G2 assessment tools at each data collection period so that G1 and G2 student capabilities could be assessed on the same scale. Thus, all person and item parameters at endline were on the same scale as at baseline and midline. In turn, MSI used a common examinee method to equate scores for the oral reading and comprehension sub-tasks between grades and over time.

reading and math habits, and reading and math performance.[6] In addition to these two techniques, MSI used baseline assessment scores as control variables in their final regression analysis of program impact. Where the first two approaches attempt to make the treatment and comparison groups more similar, the latter approach attempts to control for variation in endline scores that is explained by students' own baseline capabilities. The effect of this is the same as the propensity score weighting, however, which is to minimize the effect of differences in baseline levels of achievement between treatment and control students on midline and endline estimates of program impact.

Both student-level analysis choices (propensity score weighting and using baseline assessment scores as covariates) are valid ways to estimate program impact if one can be assured that differences in baseline assessment scores reflect true pre-intervention differences in aptitude between these two groups. However, the crux of RTI's concern is that this difference could have been caused by the first six to eight weeks of RAMP implementation. If this were true, weighting students in part by their similarity in baseline outcomes would unintentionally erase actual program impact. Similarly, it would be invalid to include baseline assessment scores as covariates in an OLS regression predicting endline assessment scores, since they would be causally related with the treatment term (whose coefficient yields the estimate of program impact). It is thus fair for RTI to request that these analysis parameters be omitted if they have reason to believe baseline differences in assessment scores between the treatment and comparison group were caused by RAMP.

RESULTS OF ALTERNATIVE MSI ANALYSIS STRATEGIES

We present in Table 3 the endline impact evaluation results according to four different sets of specifications in response to RTI's objections. The first column, titled "PSW & BL," incorporates propensity score weights and baseline assessment scores as covariates in the final OLS regression. This is the model presented in the body of MSI's report. The second column, titled "PSW only," includes the propensity score weights but omits baseline assessment scores as covariates. This is the model presented in Annex P of MSI's report. The third column, "BL only," omits propensity score weights but retains baseline assessment scores as covariates, while the fourth column, "neither," omits both propensity score weights and baseline assessment scores as covariates. These two models are not presented in MSI's endline report, and the latter model is the one RTI would like presented as an upper boundary for potential RAMP impacts, assuming the baseline difference in scores represents early program impact.

Table 3 summarizes the estimates of program impact yielded when we run these four versions of MSI's final OLS regression model. The coefficients listed are for the treatment term of the regression, indicating the estimated program impact, while the p-values indicate the percentage likelihood that the true program impact under the specifications of the model is non-zero. Coefficients highlighted in green are positive and significant at a 95% confidence level—we can be at least 95% confident that the coefficient is greater than zero. Coefficients highlighted in red are negative and significant at a 95% confidence level, while the remaining coefficients (not highlighted) are statistically insignificant. For cases of statistical insignificance, the program impact can be considered equivalent to zero. For example, according to the "PSW & BL" model, we can conclude that the RAMP program causes an improvement of nearly one syllable segment out of ten on the Grade 3 syllable segmentation sub-test, while causing no discernible change in Grade 2 phoneme isolation scores.

Taking these four models together, it is clear that accounting for baseline assessment scores in any way yields mostly null program impacts, even if propensity score weights are omitted and the model only accounts for the variation in endline assessment scores that is explained by a student's own baseline

[6] Propensity score estimation was done at both the student- and school-level using a boosted propensity score estimation method with a Kolmogorov-Smirnov statistic criterion.

assessment scores (BL only model). However, if neither propensity score weights nor baseline assessment score covariates are used, the model estimates significant, positive endline RAMP impacts across all program subtasks except for G2 letter sounds, counting objects, and addition facts and G3 letter sounds and reading comprehension. **This massive divergence in estimates yielded by the models is almost certainly explained by the difference in baseline assessment scores between the treatment and comparison group (see Table D.1 in Annex II).**

Table 3: Results of Alternative MSI Analysis Strategies

Grade	Outcome Variable	PSW & BL		PSW Only		BL Only		Neither	
		Coef.	*p-value*	Coef.	*p-value*	Coef.	*p-value*	Coef.	*p-value*
2	Phoneme isolation (out of 10)	-0.258	*0.194*	-0.162	*0.425*	-0.098	*0.525*	0.353	*0.030*
	Syllable segmentation (out of 10)	0.265	*0.338*	0.381	*0.191*	0.654	*0.004*	1.305	*0.000*
	Letter sound (out of 100, prorated score)	-3.787	*0.008*	-3.473	*0.028*	-1.465	*0.241*	0.278	*0.842*
	Reading vocabulary (out of 10)	0.034	*0.786*	0.118	*0.442*	0.077	*0.523*	0.502	*0.001*
	Passage reading (out of 52, prorated score)	-0.220	*0.817*	0.328	*0.749*	0.427	*0.495*	2.962	*0.000*
	Reading comprehension (out of 6)	-0.190	*0.162*	-0.119	*0.410*	-0.029	*0.756*	0.279	*0.008*
	Counting numbers (out of 60)	-0.289	*0.530*	-0.089	*0.860*	-0.170	*0.655*	0.950	*0.041*
	Counting objects (out of 10)	-0.042	*0.366*	-0.030	*0.516*	-0.068	*0.157*	-0.013	*0.769*
	Number identification (out of 20, prorated score)	-0.228	*0.791*	0.324	*0.758*	0.035	*0.957*	2.499	*0.002*
	Number discrimination (out of 10)	0.008	*0.950*	0.071	*0.620*	0.038	*0.724*	0.352	*0.007*
	Missing number (out of 10)	-0.101	*0.488*	0.030	*0.871*	0.029	*0.816*	0.667	*0.000*
	Addition facts-L1 (out of 20, prorated score)	-1.045	*0.013*	-0.806	*0.080*	-0.406	*0.226*	0.607	*0.098*
3	Syllable segmentation (out of 10)	0.784	*0.007*	0.876	*0.004*	0.831	*0.000*	1.304	*0.000*
	Letter sound (out of 100, prorated score)	-3.161	*0.025*	-3.631	*0.037*	-2.498	*0.051*	-1.033	*0.470*
	Non-word decoding (out of 50, prorated score)	0.755	*0.268*	0.784	*0.336*	0.349	*0.560*	1.875	*0.005*
	Reading vocabulary (out of 10)	-0.027	*0.784*	-0.025	*0.815*	-0.019	*0.860*	0.259	*0.048*
	Passage reading (out of 52, prorated score)	-0.552	*0.494*	-0.666	*0.513*	-0.472	*0.468*	1.657	*0.039*
	Reading comprehension (out of 6)	-0.043	*0.638*	-0.055	*0.599*	-0.125	*0.140*	0.113	*0.255*
	Number identification (out of 20, prorated score)	0.353	*0.589*	0.391	*0.727*	0.560	*0.337*	2.995	*0.000*
	Number discrimination (out of 10)	0.006	*0.954*	0.027	*0.852*	0.023	*0.790*	0.360	*0.002*
	Missing number (out of 10)	-0.039	*0.720*	0.001	*0.995*	-0.002	*0.983*	0.431	*0.002*
	Addition facts-L1 (out of 20, prorated score)	-0.259	*0.520*	-0.120	*0.787*	-0.244	*0.475*	0.733	*0.044*
	Addition facts-L2 (out of 5)	0.004	*0.972*	0.045	*0.759*	0.033	*0.761*	0.364	*0.003*
	Subtraction facts-L1 (out of 20, prorated score)	-0.261	*0.496*	-0.113	*0.783*	-0.216	*0.493*	0.677	*0.038*
	Subtraction facts-L2 (out of 5)	0.040	*0.764*	0.075	*0.631*	-0.007	*0.948*	0.293	*0.012*

INTERPRETING THE DIVERGENCE IN FINDINGS FROM ALTERNATIVE ANALYSIS SPECIFICATIONS

Given the way that treatment of baseline differences in assessment scores affects estimates of program impact in MSI's impact evaluation, the key question is whether this difference is likely due to different pre-existing aptitudes in these two groups or an initial impact from RAMP, which was then sustained over the remaining year and a half of the study**. If MSI's baseline survey were conducted eight to ten weeks earlier, would these differences still have manifested, or would the baseline assessment scores have been equivalent?**

In this section, we discuss the evidence in favor of each of these potential conclusions. With the data and documentation available to us, it is not possible to definitively determine which of these is the case. However, in deciding which is most likely to be true, we considered the following factors:

1. Is reliable secondary data available to assess the degree to which treatment and comparison schools had students with pre-existing differences in aptitude for early-grade reading and math?
2. From a theoretical point of view, is it expected that the program would achieve impacts within the first six to eight weeks of pupil exposure and then maintain these impacts over time? If so, are impacts expected on certain sub-tests more than others (e.g. letter sound identification and not reading comprehension), or are they expected across all sub-tests?
3. Is there evidence that RAMP was implemented according to plan, especially in the first six to eight weeks of implementation?

SECONDARY DATA ON PRE-EXISTING CONDITIONS

If standardized early-grade reading and math assessment scores were available at the Grade 1 and Grade 2 level for the schools included in the study, it would be possible to assess whether the difference in scores measured by MSI at baseline reflects or is atypical of existing trends. Although MSI's school-level propensity score matching incorporated historical reading and math scores for treatment and comparison schools, and the matched treatment and comparison schools were balanced on these scores at baseline, these assessment scores are reported by principals, are not standardized, and are reported at the Grade 3 level (see Table B.2, Annex II). In their comments on MSI's endline report, RTI argued that the balance between treatment and comparison students on most demographic characteristics in the baseline survey, prior to student-level propensity score weighting (see table D.1, Annex II), indicates that pre-existing assessment scores for these students were also likely to be balanced. However, there are many observable and unobservable student- and school-level characteristics that could explain the difference in scores, and this balance is not sufficient in the absence of historical standardized assessment scores to conclude that the difference in baseline scores must be caused by RAMP alone.

THEORETICAL TIMELINE OF RAMP IMPACT

To determine whether we believe the high baseline treatment assessment scores come from six to eight weeks of exposure to RAMP, we must assess whether the pattern of improved scores is likely given the program's theory of change. RTI asserts that a 2018 paper by Benjamin Piper et al. supports the notion that the gains in the first six to eight weeks of a similar program are among the most significant that the program achieves. The report does not include a complete citation, so we are unable to validate this assertion. We would need to critically assess the evidence supporting such early impacts to determine if it is generalizable to this context. If the assertion is made on the basis of curvilinear increases in scores of students benefitting from the Kenya Tusome and Malawi EGRA/MERIT programs, as is cited in the Technical Note in Annex B of MSI's Baseline Report, MSI notes that "the gains observed in these

projects are not measured relative to gains in a comparison group and, therefore, cannot be confidently attributed to the interventions alone."[7]

In terms of the program itself, RTI's midline report[8] indicated that the in-service training activity "is likely to have had the greatest direct impact on student reading and mathematics performance." The first two modules of this activity, which targeted improved reading and math instruction, were delivered before the school year began. In Annex B of their baseline report, MSI concedes that there are several factors that theoretically support early gains from RAMP, including energized teachers at the beginning of the school year who still have the training from the summer fresh in their minds.

Thus, the aspects of the program most likely to produce initial impacts were indeed implemented before MSI's baseline took place, although this does not guarantee that these aspects would yield results in a six to eight-week timeframe or that they were implemented to the quality and extent that is prescribed. Additionally, previous studies[9] have indicated that certain reading and math skills, like reading comprehension and conceptual mathematics, improve more slowly than others, like letter sounds and procedural mathematics. Given that Grade 1 and 2 students are at an early stage in their academic career, it stands to reason that program-caused improvements in assessment scores in the short-term would favor simpler sub-tasks like letter sounds and not yet manifest in more complex sub-tasks like reading comprehension. As displayed in Table D.1, Annex II, the baseline differences observed in assessment scores appear across all reading and math sub-tasks in both grades. Furthermore, as MSI suggests in Annex B of their Baseline Report, the magnitude of these differences seems excessively large given the short period of time and the presence of nine non-instructional days in this time period.

Thus, although it appears plausible that RAMP could have had an impact in the first six to eight weeks of implementation based on the way it was designed, the broad pattern of the difference between baseline assessment scores treatment and comparison schools may suggest that the differences are pre-existing and reflective of higher-performing students in treatment schools.

RAMP IMPLEMENTATION FIDELITY

As discussed in the previous section, concluding that RAMP had an impact in the first six to eight weeks presumes that the program was implemented according to plan—reaching the number of teachers, delivering the quality of in-service training and materials, and progressing at the pace that was prescribed. It is certainly possible that impacts could have been achieved in the context of sub-optimal implementation, but evidence of stronger implementation fidelity lends itself more to believing assertions of sizeable short-term impact over this timeframe.

RTI and MSI clearly have conflicting points of view on RAMP implementation fidelity. MSI concludes that poor implementation fidelity is an explanatory factor for mostly null program impacts and asserts that "interviews with teachers and principals offered no persuasive evidence to suggest that RAMP caused the observed differences in the short period between the beginning of the school year and baseline data collection for this evaluation."[10] In RTI's response to MSI's endline report, RTI laments that "[w]e are left without findings as to how often RAMP materials are being used, how well teachers plan using the RAMP approach, how often skills trained on by RAMP are used, nor how often RAMP is integrated into the daily instructional behavior of teachers," generally contending that the evaluation displays a limited understanding of the RAMP approach in the measures and analyses. With the documentation available to

[7] Pg. 65
[8] Pg. 21
[9] Jukes, M. C. H., Dubeck, M. M., Adelman, E., Sheppard, M., Jasti, C., & Turner, E. (2016). The impact of child-to-child reading on reading skills and motivation in Kenya. Manuscript submitted for publication.
Piper, B., Ralaingita, W., Akach, L., & King, S. (2016). Improving procedural and conceptual mathematics outcomes: Evidence from a randomised controlledtrial in Kenya. Journal of Development Effectiveness, 2016. https://doi.org/
10.1080/19439342.2016.1149502.
[10] Pg. 63

us and the complexity of the approach, we do not feel strongly positioned to make a definitive statement about which implementation fidelity contention is most accurate. However, given its more intimate knowledge of the implementation, we encourage USAID/Jordan to reflect on the implementation process to draw conclusions on implementation fidelity, which can inform whether it believes an initial impact from RAMP is detected in MSI's baseline survey.

CONCLUSIONS AND AREAS FOR FURTHER INQUIRY

There are two possible conclusions that would allow USAID to reconcile MSI and RTI's divergent findings regarding RAMP's impact:

1. RAMP had a mostly null impact on Grade 1 and 2 students' early grade reading and math capabilities, adding little to no effect to an existing positive trend in these capabilities in Jordan. This conclusion is valid and defensible if the observed baseline difference in assessment scores between treatment and comparison students are not caused by six to eight weeks of exposure to RAMP. Rather, these changes reflect a pre-existing trend between Grade 1 and Grade 2 students in these two sets of schools.
2. RAMP had a positive impact across nearly all early grade reading and math capabilities for Grade 1 and 2 students, contributing to an increasing trend in these capabilities in Jordan. This impact was realized almost in its entirety in the first six to eight weeks of implementation and maintained thereafter. This conclusion follows naturally if the observed baseline difference in assessment scores between treatment and comparison students are caused by six to eight weeks of exposure to RAMP.

There is insufficient evidence in the data and available documentation to conclude which of these possibilities occurred with absolute certainty. We can conclude confidently based on MSI's impact evaluation that there were little to no additional impacts realized by RAMP on Cohort 2 students after the first eight weeks of implementation, but their impact evaluation cannot prove that baseline differences in assessment scores between treatment and comparison students reflect pre-existing differences in these groups rather than initial program impacts. RTI's midline study does not allow them or us to assert that the second scenario is true either, as they do not construct a counterfactual that would causally link RAMP to increases in assessment scores.

Based primarily on (i) the lack of information on pre-existing Grade 1 and 2 standardized assessment scores in treatment and comparison schools and (ii) the imbalance in baseline scores across all reading and math sub-tasks (rather than only sub-tasks that would be likely to improve quickly) we are inclined to favor the first conclusion. However, we would likely abandon this if independent analysis of historical standardized assessment scores in treatment and comparison schools indicated that students scored similarly prior to the intervention beginning. If such data are publicly available, we highly encourage USAID to obtain a corresponding public-use dataset and use it to inform its final takeaways from these studies.[11]

RECOMMENDATIONS FOR FUTURE EVALUATIONS

Given the immense complications that arose from the inability to establish a true baseline for treatment and comparison students, we recommend the following measures for future evaluations, to facilitate USAID's ability to derive lessons learned from evaluation conclusions:

1. Ensure that baseline data collection can occur before the program begins.

[11] If USAID is able to secure this dataset and it is of interest, SI will analyze it and include the results in the revised final version of this report.

2. Strengthen collaboration between evaluator and implementer to improve measurement of implementation.

If it is not possible to collect baseline data completely before implementation begin, it may be possible to collect data from a smaller sample of students in the same schools that have not yet been affected by the program. This sample could comprise students in the neighboring grades and could indicate whether students in treatment and control schools start with similar aptitude in the absence of the program.

ANNEX I: KEY EXCERPTS FROM RTI'S MIDLINE REPORT[12]

Table ES2. EGRA results for 2014 and 2017 by grade

Subtask	Measure	2014 G2	2017 G2	2014 G3	2017 G3
Letter sound	fluency (correct letters per min.)	38.2 (33.9; 42.4)	47.3 *** (45.3; 49.3)	35.7 (31.6; 39.8)	48.6 *** (45.5; 51.7)
Syllable sound	fluency (correct syllables per min.)	22.1 (19.7; 24.6)	29.8 *** (27.5; 32.0)	28.7 (26.6; 30.8)	33.8 *** (31.5; 36.1)
Invented words	fluency (correct words per min.)	7.1 (6.2; 8.1)	10.8 *** (9.8; 11.8)	10.9 (9.8; 11.9)	13.7 *** (12.8; 14.7)
	% correct of items attempted	34.5% (30.2%; 38.8%)	48.2% *** (44.9%; 51.5%)	44.9% (41.3%; 48.5%)	50.4% (47.9%; 53.0%)
Oral reading	oral reading fluency (ORF)	19.1 (16.9; 21.3)	21.4 (20.0; 22.9)	35.0 (32.2; 37.8)	31.7 (29.7; 33.6)
	% correct of items attempted	56.7% (50.0%; 63.4%)	56.2% (52.6%; 59.8%)	71.6% (68.5%; 74.8%)	64.7% (61.4%; 68.1%)
Reading comprehension	% of students with 80% comp.	7.9% (4.7%; 11.2%)	11.4% (9.2%; 13.6%)	29.0% (23.6%; 34.4%)	31.5% (27.8%; 35.2%)

* p<.05; ** p<.01; *** p<.001

Table ES4. EGMA results for 2014 and 2017 by grade

Subtask	Measure	2014 G2	2017 G2	2014 G3	2017 G3
Addition and Subtraction L2	% correct	36.8% (33.8%; 39.9%)	42.5% ** (39.5%; 45.6%)	47.5% (44.1%; 50.9%)	51.2% (48.5%; 53.9%)
Missing Number	% correct	54.3% (50.7%; 57.9%)	60.2% ** (57.4%; 63.1%)	66.9% (63.9%; 70.0%)	68.9% (66.6%; 71.2%)

* p<.05; ** p<.01; *** p<.001

Table 9. Grade 2 EGRA results for 2014 and 2017, including 2017 intervention and non-intervention results

Subtask	Measure	2014 G2	2017 G2	2017 G2 Int	2017 G2 non-Int
Letter sound	fluency (correct letters per min.)	38.2 (33.9; 42.4)	47.3 (45.3; 49.3)	47.6 (45.3; 49.9)	46.2 (42.2; 50.1)
Syllable sound	fluency (correct syllables per min.)	22.1 (19.7; 24.6)	29.8 (27.5; 32.0)	31.5 (29.4; 33.6)	22.8 ** (17.9; 27.7)
Invented words	fluency (correct words per min.)	7.1 (6.2; 8.1)	10.8 (9.8; 11.8)	11.6 (10.5; 12.6)	7.9 *** (6.2; 9.5)
	% correct of items attempted	34.5% (30.2%; 38.8%)	48.2% (44.9%; 51.5%)	50.5% (46.9%; 54.1%)	39.0% *** (34.4%; 43.5%)
Oral reading	oral reading fluency (ORF)	19.1 (16.9; 21.3)	21.4 (20.0; 22.9)	22.6 (21.1; 24.0)	17.0 *** (14.1; 19.9)
	% correct of items attempted	56.7% (50.0%; 63.4%)	56.2% (52.6%; 59.8%)	58.5% (54.8%; 62.1%)	47.2% ** (39.7%; 54.7%)
Reading comprehension	% of students with 80% comp.	7.9% (4.7%; 11.2%)	11.4% (9.2%; 13.6%)	12.7% (10.2%; 15.3%)	6.3% ** (3.6%; 9.0%)

* p<.05; ** p<.01; *** p<.001

Int = intervention; non-Int = non-intervention

[12] Note that p-values are testing significance of difference of means relative to 2014, by grade for Tables ES2 and ES4 and between intervention statuses for Table 9

ANNEX II: KEY EXCERPTS FROM MSI'S ENDLINE REPORT

Table 13. Summary of RAMP Reading and Math Impacts at Midline and Endline, by Grade.

	Midline		Endline	
READING	G1 at baseline and midline	G2 at baseline and midline	G1 at base- and midline; G2 at endline	G2 at base- and midline; G3 at endline
1. Orientation to print	No impact	NA	NA	NA
2. Phoneme isolation	No impact	NA	No impact	NA
3. **Syllable segmentation**	Positive impact [S]	Positive impact [S]	No impact	Positive impact
4. **Letter sound knowledge**	No impact	No impact	Negative impact	Negative impact
5. Non-word decoding	NA	No impact [S]	NA	No impact
6. Reading vocabulary	No impact [S]	No impact	No impact [S]	No impact
7. **Passage reading**	Positive impact [S]	No impact [S]	No impact [S]	No impact
8. **Reading comprehension**	No impact [G]	No impact	No impact [S]	No impact
MATHEMATICS				
1. **Counting numbers**	Negative impact [S]	NA	No impact	NA
2. Counting objects (or enumerating quantities)	No impact	NA	No impact	NA
3. Number identification	No impact	No impact [G]	No impact	No impact
4. Number discrimination	No impact	No impact	No impact	No impact
5. Missing numbers	No impact	No impact	No impact	No impact [S]
6. **Addition facts - L1**	No impact	No impact	Negative impact	No impact
7. Addition facts – L2	NA	No impact	NA	No impact
8. Subtraction facts - L1	NA	No impact	NA	No impact [S]
9. Subtraction facts – L2	NA	No impact	NA	No impact [S]

Note: G denotes different impacts of RAMP for boys versus girl. S denotes different impacts of RAMP for students in single- versus double-shift schools. See details about differences below. L1 denotes ___ and L2 _____.

Table 14. Endline Reading Performance Scores for Grade 1 Students.

Variable (Total # of items)	Intervention (T)	Comparison (C)	Impact (T-C) *	*p*-value	Effect size	Number of Students
Phoneme isolation (out of 10)	4.6	4.9	-0.3	0.19	-0.1	1931
Syllable segmentation (out of 10)	6.8	6.6	0.3	0.34	0.1	1931
Letter sound knowledge (out of 100, prorated)	30.5	34.3	**-3.8***	0.008	-0.2	1931
Reading vocabulary (out of 10)	8.2	8.1	0.0	0.79	0.0	1931
Passage reading (out of 41, prorated score)	14.6	14.8	-0.2	0.82	0.0	1931
Reading comprehension (out of 6)	1.5	1.7	-0.2	0.16	-0.1	1931
Number of schools	**117**	**120**				

Source: RAMP Impact Study - Endline data 2018 Student assessments
Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Students' home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes are calculated as Hedge's g. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated to the midline scale.
*Difference in group means is statistically significant at the .05 level.

Table 15. Endline Reading Performance Scores for Grade 2 Students.

Variable (Total # of items)	Intervention (T)	Comparison (C)	Impact (T-C) *	*p*-value	Effect size	Number of students
Syllable segmentation (out of 10)	6.0	5.2	**0.8***	0.007	0.2	1931
Letter sound knowledge (out of 100, prorated)	33.6	36.7	**-3.2***	0.025	-0.1	1931
Non-word decoding (out of 50, prorated)	12.6	11.8	0.8	0.27	0.1	1931
Reading vocabulary (out of 10)	9.1	9.2	0.0	0.78	0.0	1931
Passage reading (out of 41, prorated score)	22.5	23.1	-0.6	0.49	0.0	1931
Reading comprehension (out of 6)	2.7	2.8	0.0	0.64	0.0	1931
Number of schools	**118**	**119**				

Source: RAMP Impact Study - Endline data 2018 Student assessments
Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Students' home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes are calculated as Hedge's g. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated to the midline scale.
*Difference in group means is statistically significant at the .05 level.

Table 20. Endline math Performance Scores for Grade 1 Students.

Variable (Total # of items)	Intervention (T)	Comparison (C)	Impact (T-C) *	*p*-value	Effect size	Number of students
Counting numbers (out of 40)	36.3	36.6	-0.3	0.53	0.0	1931
Enumerating quantities (out of 10)	9.8	9.9	0.0	0.37	-0.1	1931
Number identification (out of 20, prorated)	32.3	32.6	-0.2	0.79	0.0	1931
Number discrimination (out of 10)	9.0	9.0	0.0	0.95	0.0	1931
Missing numbers (out of 10)	6.8	6.9	-0.1	0.49	0.0	1931
Addition facts - L1 (out of 20, prorated)	5.1	6.1	**-1.0***	0.013	-0.2	1931
Number of schools	**117**	**120**				

Source: Ramp Impact Study - Endline data 2018 Student assessments
Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes are calculated as Hedge's g. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated to the midline scale.
*Difference in group means is statistically significant at the .05 level.

Table 21. Endline Math Performance Scores for Grade 2 Students.

Variable (Total # of items)	Intervention (T)	Comparison (C)	Impact (T-C) *	*p*-value	Effect size	Number of students
Number identification (out of 20, prorated)	35.3	35.0	0.4	0.59	0.0	1931
Number discrimination (out of 10)	8.6	8.6	0.0	0.95	0.0	1931
Missing numbers (out of 10)	8.3	8.4	0.0	0.72	0.0	1931
Addition facts - L1 (out of 20, prorated) timed	9.1	9.4	-0.3	0.52	0.0	1931
Addition facts - L2 (out of 5)	2.9	2.8	0.0	0.97	0.0	1931
Subtraction facts - L1 (out of 20, prorated)	7.0	7.2	-0.3	0.50	0.0	1931
Subtraction facts - L2 (out of 5)	2.2	2.2	0.0	0.76	0.0	1931
Number of schools	**118**	**119**				

Source: RAMP Impact Study - Endline data 2018 Student assessments
Note: Columns T and C present ordinary least squares regression-adjusted group means (or percentages). Regressions adjust for baseline reading and math scores as well students' demographic characteristics. Home language and whether the student attended preschool use midline or endline data, when baseline is missing. Errors are clustered at the school level. All regressions include student propensity score weights to improve baseline equivalence between the groups. Effect sizes are calculated as Hedge's g. The table indicates the number of items administered at baseline, because the endline scores were equated into the baseline scale. Subtasks that were not administered at baseline were equated into the midline scale.
*Difference in group means is statistically significant at the .05 level.

TABLE O.2. READING AND MATH DESCRIPTIVE STATISTICS FOR GRADE 1 STUDENTS IN COHORT 2

	Baseline					Midline					Endline				
Variable	Number of students	Mean	Sd	Min	Max	Number of students	Mean	Sd	Min	Max	Number of students	Mean	Sd	Min	Max
Reading															
Orientation to print	973	2.4	1.4	0	5	941	3.1	1.3	0	5	NA	NA	NA	NA	NA
Phoneme isolation	973	3.9	2.1	0	10	941	4.4	2.6	0	10	973	5	2.8	0	10
Letter sound	973	22.2	14.8	0	86	941	32.5	19.4	0	88	973	33.2	21.1	0	90
Syllable segmentation	973	4.9	4.3	0	10	941	6.1	4	0	10	973	7.2	3.6	0	10
Invented word decoding	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Reading vocabulary	973	4.4	3.1	0	10	941	7.1	2.6	0	10	973	8.4	2.2	0	10
Passage reading	NA	NA	NA	NA	NA	941	7.3	9.1	0	53.2	973	16.6	13.2	0	68.3
Reading comprehension	NA	NA	NA	NA	NA	941	1	1.1	0	6	973	1.7	1.7	0	6
Math															
Counting numbers	973	28.7	10.9	0	40	941	33.3	10.5	2	40	973	36.9	7.4	4	40
Counting objects	973	8.9	1.9	0	10	941	9.8	0.8	1	10	973	9.9	0.8	1	10
Number identification	973	28.4	13.5	0	76	941	24	7.5	1.3	57.1	973	33.8	12.5	0	85.7
Number discrimination	973	8	2.5	0	10	941	8.7	2.6	0	10	973	9.3	1.9	0	10
Missing number	NA	NA	NA	NA	NA	941	4.9	2.9	0	10	973	7.2	2.7	0	10
Addition facts-L1	NA	NA	NA	NA	NA	941	5.7	6.6	0	24	973	6	5.7	0	37.2

Source: RAMP Impact Study - Baseline, Midline, and Endline data 2018 Student assessments

Note: The table shows unadjusted, unweighted descriptive statistics (mean, standard deviation, and minimum and maximum values observed) for student reading and math scores. Sd stands for standard deviation. Midline and endline scores are equated into the baseline scale. Endline subtasks that were not administered at baseline are equated into the midline scale, and midline subtasks that were not administered at baseline are not equated. The analytic sample at all time points consists of students who were assessed at both baseline and endline. NA indicates that the subtask was not administered to students in that grade and time point.

TABLE O.3. READING AND MATH DESCRIPTIVE STATISTICS FOR GRADE 1 STUDENTS IN COHORT 3

	Baseline					Midline					Endline				
Variable	Number of students	Mean	Sd	Min	Max	Number of students	Mean	Sd	Min	Max	Number of students	Mean	Sd	Min	Max
Reading															
Orientation to print	979	2.1	1.5	0	5	923	2.8	1.4	0	5	NA	NA	NA	NA	NA
Phoneme isolation	979	3.4	2.1	0	9	923	3.8	2.4	0	10	979	4.5	2.7	0	10
Letter sound	979	19.8	13.1	0	62	923	29.1	18.2	0	80	979	32.7	20	0	90
Syllable segmentation	979	3	3.9	0	10	923	4.5	4.1	0	10	979	5.9	4	0	10
Invented word decoding	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Reading vocabulary	979	3.5	3	0	10	923	6.1	2.9	0	10	979	7.8	2.6	0	10
Passage reading	NA	NA	NA	NA	NA	923	4.7	7.1	0	46.8	979	13.1	12.1	0	61.6
Reading comprehension	NA	NA	NA	NA	NA	923	0.7	1	0	5	979	1.3	1.6	0	6
Math															
Counting numbers	979	25.6	11.3	0	40	923	31.3	11.4	1	40	979	35.8	8.8	0	40
Counting objects	979	8.4	2.3	0	10	923	9.6	1.1	1	10	979	9.9	0.6	1	10
Number identification	979	24.7	13.8	0	92.3	923	22.2	7.6	0	63.2	979	31.1	12.1	0	75
Number discrimination	979	7.1	2.9	0	10	923	8.1	3	0	10	979	8.9	2.3	0	10
Missing number	NA	NA	NA	NA	NA	923	4	2.8	0	10	979	6.4	2.9	0	10
Addition facts-L1	NA	NA	NA	NA	NA	923	3.9	5.7	0	20.7	979	5.2	5.7	0	36.4

Source: RAMP Impact Study - Baseline, Midline, and Endline data 2018 Student assessments

Note: The table shows unadjusted, unweighted descriptive statistics (mean, standard deviation, and minimum and maximum values observed) for student reading and math scores. Sd stands for standard deviation. Midline and endline scores are equated into the baseline scale. Endline subtasks that were not administered at baseline are equated into the midline scale, and midline subtasks that were not administered at baseline are not equated. The analytic sample at all time points consists of students who were assessed at both baseline and endline. NA indicates that the subtask was not administered to students in that grade and time point.

TABLE O.5. READING AND MATH DESCRIPTIVE STATISTICS FOR GRADE 2 STUDENTS IN COHORT 2

	Baseline					Midline					Endline				
Variable	Number of students	Mean	Sd	Min	Max	Number of students	Mean	Sd	Min	Max	Number of students	Mean	Sd	Min	Max
Reading															
Orientation to print	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Phoneme isolation	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Letter sound	966	29.3	19.3	0	84	936	34.2	21.7	0	105.3	966	35.1	24.1	0	101.7
Syllable segmentation	966	5.3	3.6	0	10	936	6.4	3.4	0	10	966	6.3	3.4	0	10
Invented word decoding	966	7.9	7.3	0	38	936	9.2	7.8	0	50	966	13.3	11.7	0	50
Reading vocabulary	966	6.9	2.9	0	10	936	8.1	2.5	0	10	966	9.3	1.8	0	10
Passage reading	966	12.2	13.1	0	100	936	18.1	11.8	0	82	966	23.8	14.1	0	68.3
Reading comprehension	966	1.4	1.3	0	6	936	2	1.8	0	6	966	2.8	1.8	0	6
Math															
Counting numbers	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Counting objects	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Number identification	966	24.1	12	1	60	936	30.7	13.7	0	80	966	37	14.5	3	80
Number discrimination	966	7.7	2.1	0	10	936	7.8	2.4	0	10	966	8.8	1.8	0	10
Missing number	966	5.9	2.8	0	10	936	7.3	2.8	0	10	966	8.6	2.2	0	10
Addition facts-L1	966	9.1	5.5	0	29.3	936	10.7	5.5	0	42.9	966	9.7	6.2	0	33.3

Source: RAMP Impact Study - Baseline, Midline, and Endline data 2018 Student assessments

Note: The table shows unadjusted, unweighted descriptive statistics (mean, standard deviation, and minimum and maximum values observed) for student reading and math scores. Sd stands for standard deviation. Midline and endline scores are equated into the baseline scale. Endline subtasks that were not administered at baseline are equated into the midline scale, and midline subtasks that were not administered at baseline are not equated. The analytic sample at all time points consists of students who were assessed at both baseline and endline. NA indicates that the subtask was not administered to students in that grade and time point.

TABLE O.6. READING AND MATH DESCRIPTIVE STATISTICS FOR GRADE 2 STUDENTS IN COHORT 3

	Baseline					Midline					Endline				
Variable	Number of students	Mean	Sd	Min	Max	Number of students	Mean	Sd	Min	Max	Number of students	Mean	Sd	Min	Max
Reading															
Orientation to print	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Phoneme isolation	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Letter sound	978	26.7	18	0	86	936	33	21.4	0	99	978	35.5	22.2	0	100.3
Syllable segmentation	978	3.4	3.6	0	10	936	4.5	3.5	0	10	978	4.9	3.5	0	10
Invented word decoding	978	6.1	6.3	0	34	936	7.1	7	0	35	978	11	10.6	0	44
Reading vocabulary	978	6	3.4	0	10	936	7.7	2.9	0	10	978	8.9	2.4	0	10
Passage reading	978	9.4	11.7	0	92.3	936	15.4	11.5	0	84.8	978	21.4	14.3	0	72.7
Reading comprehension	978	1.3	1.3	0	6	936	1.6	1.7	0	6	978	2.6	1.8	0	6
Math															
Counting numbers	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Counting objects	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Number identification	978	20.8	12.1	0	92.3	936	27.7	14.3	0	100	978	33.5	15.6	0	120
Number discrimination	978	6.9	2.7	0	10	936	7.4	2.6	0	10	978	8.3	2.2	0	10
Missing number	978	5.3	2.9	0	10	936	6.6	3.1	0	10	978	8.1	2.6	0	10
Addition facts-L1	978	7.3	5.4	0	27.3	936	9.6	5.7	0	26.1	978	8.8	6	0	34.3

Source: RAMP Impact Study - Baseline, Midline, and Endline data 2018 Student assessments

Note: The table shows unadjusted, unweighted descriptive statistics (mean, standard deviation, and minimum and maximum values observed) for student reading and math scores. Sd stands for standard deviation. Midline and endline scores are equated into the baseline scale. Endline subtasks that were not administered at baseline are equated into the midline scale, and midline subtasks that were not administered at baseline are not equated. The analytic sample at all time points consists of students who were assessed at both baseline and endline. NA indicates tat the subtask was not administered to students in that grade and time point.

TABLE B.2. MEANS AND STANDARDIZED BIAS AFTER MATCHING.

School characteristic	Intervention group mean	Comparison group mean	Absolute % bias after matching	Interpretation
Number of non-classrooms in school	7.5	7.6	1.1	Satisfies equivalency requirement
Percent of staff who are administrators	13.5%	13.4%	1.7	
Total early grade teachers	6.9	7.0	1.9	
G3 average language score in previous year	79.4	79.6	2.1	
Percent of schools receiving donor intervention (not training or infrastructure-related)	8.5%	9.2%	2.2	
Free meal provided	26.5%	27.5%	2.2	
G3 Math average score	79.2	79.0	2.3	
Sections of G3	1.9	1.9	3.5	
Percent of schools with kindergarten	31.6%	30%	3.5	
Index for school repairs	15.9	16.2	4.3	
Year school constructed	1987.5	1988.3	4.9	
Number of G3 students	55.3	53.2	4.9	
Number of G1 students	52.8	50.5	5	Requires statistical adjustment
Sections of G2	1.8	1.9	5.7	
Percent of early grade students who are female	51.3%	49.4%	6.3	
Sections of G3	1.9	1.9	6.5	
Number of G2 students	54.7	51.6	6.6	
Number of classrooms in school	13.1	12.7	6.7	
Maximum early grade teacher experience(in years)	88.9%	83.3%	7.2	
Early grade student-teacher ratio	22.7	22.0	7.3	
Index of water and sanitation non-availability in school	12%	14.1%	8.1	
Percent of schools where grade students use library	52.1%	46.7%	10.9	
Percent change in enrollment from beginning of previous school year	3.5%	1%	11.2	
Average level of overcrowding	114.6%	112.1%	11.8	
Percent of students who are Syrian	9.4%	5.8%	12.1	
Maximum grade in school	7.3	7.7	12.9	
G3 minimum language score in previous year	54.4	53.1	13.7	
Percent of schools receiving donor-funded training intervention	13.7%	9.2%	14.2	
Percent of schools that are urban	53%	45.8%	14.3	
Percent of students who are poor	33.7%	37.4%	14.4	

School characteristic	Intervention group mean	Comparison group mean	Absolute % bias after matching	Interpretation
Percent of schools where students have received their textbooks for the year	94%	90%	14.9	
Percent of early grade teachers with BA	95.1%	93.1%	16.9	
Percent of schools receiving donor-funded infrastructure intervention	46.2%	37.5%	17.5	
Percent of schools renovated in previous 5 years	29.9%	21.7%	18.9	
Percent of schools with donor-funded intervention (any type)	56.4%	46.7%	19.5	
Minimum early grade teacher experience (in years)	4.1	4.9	19.6	
Percent change in enrollment between beginning and end of previous year	1.4%	2.2%	19.8	
Percent of early grade teachers who are women	78.9%	70.3%	20.1	
Length of school day	5.5	5.6	20.3	
G3 minimum mathematics score in previous year	55.3	53.2	21.7	
None	-	-	-	Does not satisfy equivalency requirement

TABLE D.1. COMPARISON OF STUDENT CHARACTERISTICS BETWEEN INTERVENTION AND COMPARISON GROUPS AT BASELINE, BEFORE AND AFTER WEIGHTING

		Before Weighting			After Weighting		
		Means		Difference between means (effect size)	Means		Difference between means (effect size)
	G1 students	Intervention Group (T)	Comparison Group (C)		Intervention Group (T)	Comparison Group (C)	
G1 student characteristics							
1	Age	6.2	6.2	0.06	6.2	6.2	0.00
2	Female	55.4	50.9	0.09	58.3	50.8	0.15
3	Attended preschool	85.1	83.9	0.03	86.4	87.7	-0.04
4	Attended the same school the year before	17.3	23.3	-0.15	18.8	20.1	-0.03
5	Missed one day of school the prior week	51.2	53.7	-0.05	49.8	53.2	-0.07
6	Had a meal before getting to school	80.8	88.5	-0.22	81.1	83.5	-0.06
7	Arabic is primary language spoken at home	98.7	94.1	0.25	98.7	98.0	0.06
8	Parents knew last time student received a good grade	98.1	96.9	0.08	98.0	97.3	0.04
G1 reading habits and home environment							
9.	Likes to read	95.9	96.6	-0.04	95.7	96.4	-0.03
1(	Has books to read other than textbooks	60.3	56.7	0.08	60.0	60.0	0.00
1	Reads with other kids or parents at home	84.2	82.6	0.05	85.0	84.9	0.00
1:	Often reads aloud to another person at home	65.9	66.5	-0.01	65.1	65.7	-0.01
1:	Is read to at home	61.0	62.0	-0.02	61.1	61.2	0.00
1'	Does math problems at home	96.0	96.2	-0.01	95.8	96.4	-0.03
1!	Gets help with homework at home	88.7	88.8	0.00	89.3	89.1	0.01
1(	Takes private after-school lessons in reading or math	16.2	18.6	-0.06	17.2	16.7	0.01
G1 student reading at baseline							
1	Orientation to print raw score (out of 5)	2.4	2.1	0.25	2.5	2.4	0.07
1(	Phoneme isolation raw score (out of 10)	3.9	3.4	0.25	3.9	3.8	0.05
1'	Syllable sound raw score (out of 10)	4.8	3.1	0.40	4.8	4.4	0.09
2(	Letter sound raw score (out of 100)	22.0	19.6	0.17	22.8	22.4	0.02
2	Reading vocabulary raw score (out of 10)	4.5	3.4	0.35	4.5	4.4	0.05
G1 student math performance at baseline							
2	Count numbers raw score (out of 40)	28.0	25.3	0.25	28.6	28.0	0.06
2	Enumerating quantities raw score (out of 10)	9.0	8.4	0.29	9.0	8.9	0.04
2	Number identification raw score (out of 20)	28.3	25.0	0.24	28.8	28.2	0.04
2	Number discrimination raw score (out of 10)	8.0	7.2	0.31	8.1	8.0	0.05
Number of students		1,183	1,189	--	974	979	--
	G2 students						
G2 student characteristics							
1	Age	7.3	7.4	-0.15	7.32	7.32	-0.01
2	Female	56.3	50.0	0.13	61.79	53.41	0.17
3	Attended preschool	82.9	77.6	0.14	83.49	85.41	-0.05
4.	Attended the same school the year before	78.5	79.6	-0.03	79.93	79.53	0.01

		Before Weighting			After Weighting		
		Means		Difference between means (effect size)	Means		Difference between means (effect size)
		Intervention Group (T)	Comparison Group (C)		Intervention Group (T)	Comparison Group (C)	
5	Missed one day of school the prior week	44.7	48.7	-0.08	44.75	46.96	-0.04
6	Had a meal before getting to school	78.6	83.8	-0.14	78.22	80.62	-0.06
7	Arabic is primary language spoken at home	98.8	97.0	0.12	98.74	98.76	0.00
8.	Parents knew of last time student received a good grade	98.3	97.6	0.05	98.41	97.93	0.04
G2 reading habits and home environment							
9	Likes to read	96.6	96.4	0.01	97.1	98.7	-0.11
1	Has books to read other than textbooks	70.6	63.1	0.16	69.3	69.5	-0.01
1	Reads with other kids or parents at home	85.3	85.4	0.00	85.3	85.4	0.00
1	Often reads aloud to another person at home	63.3	61.2	0.05	61.8	63.5	-0.04
1	Is read to at home	57.0	66.1	-0.19	56.0	56.4	-0.01
1	Does math problems at home	98.6	96.4	0.14	98.4	98.4	0.00
1	Gets help with homework at home	89.8	90.0	-0.01	89.3	90.2	-0.03
1	Takes private after-school lessons in reading or math	19.8	20.1	-0.01	19.7	17.9	0.05
G2 student reading at baseline							
1	Syllable sound raw score (out of 10)	5.2	3.5	0.46	5.3	4.8	0.13
1	Letter sound raw score (out of 100)	28.3	26.8	0.08	29.3	30.0	-0.04
1	Invented word raw score (out of 50)	7.9	5.6	0.33	7.8	7.7	0.02
2	Reading vocabulary raw score (out of 10)	6.9	5.8	0.36	6.9	6.8	0.04
2	ORF passage raw score (out of 41)	12.1	8.4	0.30	12.0	11.8	0.02
2	Reading comprehension raw score (out of 6)	1.4	1.2	0.11	1.4	1.5	-0.09
G2 student math at baseline							
2	Number identification raw score (out of 20)	23.8	20.6	0.26	23.6	23.6	0.00
2	Number discrimination raw score (out of 10)	7.7	7.0	0.27	7.6	7.6	0.01
2	Missing number raw score (out of 10)	5.8	5.2	0.20	5.8	5.9	-0.03
2	Addition L1 raw score (out of 20)	8.9	7.4	0.29	9.0	8.7	0.06
2	Addition L2 raw score (out of 5)	2.3	1.9	0.23	2.4	2.2	0.08
2	Subtraction L1 raw score (out of 20)	6.2	5.0	0.29	6.3	6.0	0.06
2	Subtraction L2 raw score (out of 5)	2.2	1.8	0.20	2.3	2.1	0.06
	Number of students	1,181	1,168	--	965	978	

Source: RAMP Impact Study – Baseline 2016 student assessments

Note: The table presents ordinary least squares (OLS) regression-adjusted means (or percentages) with clustered standard errors. The last three columns on the right side show results from models including student propensity score weights for the analysis sample (students who were assessed in both baseline and endline). Effect sizes are the difference between the mean outcome of the intervention group and the mean outcome of the comparison group divided by the pooled within-group standard deviation of that outcome measure (WWC, 2017). Effect sizes > .05 (absolute value) are shaded. All comparisons before weighting include 240 schools. Comparisons after weighting include 238 schools. The table shows the maximum number of students included in the analyses. Sample sizes vary by outcome due to item non-response.

ANNEX III: FULL ALTERNATIVE SPECIFICATION REGRESSION OUTPUT

Grade	Model	Dependent Variable	coeffT	se	pval	meanT	meanC	fullN	rsquare
2	PSW Only	Phoneme isolation (out of 10)	-0.162	0.202	0.194	4.89	5.05	1931	0.026
2	PSW Only	Syllable segmentation (out of 10)	0.381	0.290	0.338	7.10	6.72	1931	0.019
2	PSW Only	Letter sound (out of 100, prorated score)	-3.473	1.573	0.008	31.65	35.12	1931	0.031
2	PSW Only	Reading vocabulary (out of 10)	0.118	0.153	0.786	8.41	8.29	1931	0.066
2	PSW Only	Passage reading (out of 52, prorated score)	0.328	1.023	0.817	16.21	15.88	1931	0.075
2	PSW Only	Reading comprehension (out of 6)	-0.119	0.144	0.162	1.66	1.78	1931	0.048
2	PSW Only	Counting numbers (out of 60)	-0.089	0.502	0.53	36.82	36.91	1931	0.024
2	PSW Only	Counting objects (out of 10)	-0.030	0.046	0.366	9.85	9.88	1931	0.016
2	PSW Only	Number identification (out of 20, prorated score)	0.324	1.051	0.791	33.81	33.49	1931	0.035
2	PSW Only	Number discrimination (out of 10)	0.071	0.144	0.95	9.23	9.16	1931	0.043
2	PSW Only	Missing number (out of 10)	0.030	0.182	0.488	7.15	7.12	1931	0.039
2	PSW Only	Addition facts-L1 (out of 20, prorated score)	-0.806	0.458	0.013	5.71	6.52	1931	0.033
3	PSW Only	Syllable segmentation (out of 10)	0.876	0.305	0.007	6.22	5.34	1930	0.030
3	PSW Only	Letter sound (out of 100, prorated score)	-3.631	1.727	0.025	33.63	37.26	1930	0.016
3	PSW Only	Non-word decoding (out of 50, prorated score)	0.784	0.814	0.268	12.97	12.19	1930	0.041
3	PSW Only	Reading vocabulary (out of 10)	-0.025	0.108	0.784	9.22	9.24	1930	0.036
3	PSW Only	Passage reading (out of 52, prorated score)	-0.666	1.016	0.494	23.00	23.66	1930	0.067
3	PSW Only	Reading comprehension (out of 6)	-0.055	0.104	0.638	2.79	2.84	1930	0.046
3	PSW Only	Number identification (out of 20, prorated score)	0.391	1.116	0.589	36.10	35.71	1930	0.010
3	PSW Only	Number discrimination (out of 10)	0.027	0.146	0.954	8.75	8.73	1930	0.018
3	PSW Only	Missing number (out of 10)	0.001	0.159	0.72	8.51	8.50	1930	0.026
3	PSW Only	Addition facts-L1 (out of 20, prorated score)	-0.120	0.445	0.52	9.47	9.59	1930	0.013
3	PSW Only	Addition facts-L2 (out of 5)	0.045	0.146	0.972	2.99	2.94	1930	0.019
3	PSW Only	Subtraction facts-L1 (out of 20, prorated score)	-0.113	0.409	0.496	7.34	7.45	1930	0.020
3	PSW Only	Subtraction facts-L2 (out of 5)	0.075	0.156	0.764	2.33	2.26	1930	0.011
2	PSW & BL	Phoneme isolation (out of 10)	-0.258	0.198	0.425	4.63	4.89	1931	0.192
2	PSW & BL	Syllable segmentation (out of 10)	0.265	0.276	0.191	6.82	6.55	1931	0.136
2	PSW & BL	Letter sound (out of 100, prorated score)	-3.787	1.419	0.028	30.53	34.32	1931	0.189

Grade	Model	Dependent Variable	coeffT	se	pval	meanT	meanC	fullN	rsquare
2	PSW & BL	Reading vocabulary (out of 10)	0.034	0.126	0.442	8.15	8.12	1931	0.350
2	PSW & BL	Passage reading (out of 52, prorated score)	-0.220	0.952	0.749	14.62	14.84	1931	0.426
2	PSW & BL	Reading comprehension (out of 6)	-0.190	0.135	0.41	1.46	1.65	1931	0.336
2	PSW & BL	Counting numbers (out of 60)	-0.289	0.460	0.86	36.27	36.56	1931	0.189
2	PSW & BL	Counting objects (out of 10)	-0.042	0.046	0.516	9.82	9.86	1931	0.065
2	PSW & BL	Number identification (out of 20, prorated score)	-0.228	0.860	0.758	32.32	32.55	1931	0.414
2	PSW & BL	Number discrimination (out of 10)	0.008	0.130	0.62	9.03	9.03	1931	0.270
2	PSW & BL	Missing number (out of 10)	-0.101	0.146	0.871	6.79	6.89	1931	0.436
2	PSW & BL	Addition facts-L1 (out of 20, prorated score)	-1.045	0.416	0.08	5.08	6.12	1931	0.280
3	PSW & BL	Syllable segmentation (out of 10)	0.784	0.288	0.004	6.03	5.24	1930	0.102
3	PSW & BL	Letter sound (out of 100, prorated score)	-3.161	1.401	0.037	33.53	36.69	1930	0.181
3	PSW & BL	Non-word decoding (out of 50, prorated score)	0.755	0.681	0.336	12.61	11.85	1930	0.401
3	PSW & BL	Reading vocabulary (out of 10)	-0.027	0.099	0.815	9.12	9.15	1930	0.284
3	PSW & BL	Passage reading (out of 52, prorated score)	-0.552	0.806	0.513	22.52	23.07	1930	0.538
3	PSW & BL	Reading comprehension (out of 6)	-0.043	0.090	0.599	2.72	2.77	1930	0.382
3	PSW & BL	Number identification (out of 20, prorated score)	0.353	0.653	0.727	35.34	34.99	1930	0.516
3	PSW & BL	Number discrimination (out of 10)	0.006	0.106	0.852	8.62	8.61	1930	0.406
3	PSW & BL	Missing number (out of 10)	-0.039	0.110	0.995	8.32	8.36	1930	0.459
3	PSW & BL	Addition facts-L1 (out of 20, prorated score)	-0.259	0.402	0.787	9.11	9.37	1930	0.411
3	PSW & BL	Addition facts-L2 (out of 5)	0.004	0.114	0.759	2.85	2.85	1930	0.386
3	PSW & BL	Subtraction facts-L1 (out of 20, prorated score)	-0.261	0.383	0.783	6.97	7.23	1930	0.395
3	PSW & BL	Subtraction facts-L2 (out of 5)	0.040	0.131	0.631	2.22	2.18	1930	0.365
2	Neither	Phoneme isolation (out of 10)	0.353	0.162	0.030	4.94	4.59	1931	0.020
2	Neither	Syllable segmentation (out of 10)	1.305	0.246	0.000	7.19	5.89	1931	0.049
2	Neither	Letter sound (out of 100, prorated score)	0.278	1.392	0.842	33.20	32.92	1931	0.015
2	Neither	Reading vocabulary (out of 10)	0.502	0.148	0.001	8.40	7.89	1931	0.068
2	Neither	Passage reading (out of 52, prorated score)	2.962	0.779	0.000	16.39	13.43	1931	0.078
2	Neither	Reading comprehension (out of 6)	0.279	0.105	0.008	1.66	1.39	1931	0.048
2	Neither	Counting numbers (out of 60)	0.950	0.464	0.041	36.84	35.89	1931	0.021
2	Neither	Counting objects (out of 10)	-0.013	0.044	0.769	9.85	9.86	1931	0.003

Grade	Model	Dependent Variable	coeffT	se	pval	meanT	meanC	fullN	rsquare
2	Neither	Number identification (out of 20, prorated score)	2.499	0.801	0.002	33.74	31.24	1931	0.034
2	Neither	Number discrimination (out of 10)	0.352	0.129	0.007	9.29	8.94	1931	0.026
2	Neither	Missing number (out of 10)	0.667	0.169	0.000	7.16	6.49	1931	0.042
2	Neither	Addition facts-L1 (out of 20, prorated score)	0.607	0.366	0.098	5.92	5.31	1931	0.015
3	Neither	Syllable segmentation (out of 10)	1.304	0.228	0.000	6.30	5.00	1931	0.053
3	Neither	Letter sound (out of 100, prorated score)	-1.033	1.428	0.470	34.84	35.87	1931	0.014
3	Neither	Non-word decoding (out of 50, prorated score)	1.875	0.656	0.005	13.14	11.26	1931	0.042
3	Neither	Reading vocabulary (out of 10)	0.259	0.130	0.048	9.22	8.96	1931	0.046
3	Neither	Passage reading (out of 52, prorated score)	1.657	0.800	0.039	23.48	21.82	1931	0.069
3	Neither	Reading comprehension (out of 6)	0.113	0.099	0.255	2.79	2.68	1931	0.052
3	Neither	Number identification (out of 20, prorated score)	2.995	0.835	0.000	36.83	33.84	1931	0.044
3	Neither	Number discrimination (out of 10)	0.360	0.116	0.002	8.76	8.40	1931	0.038
3	Neither	Missing number (out of 10)	0.431	0.135	0.002	8.57	8.14	1931	0.051
3	Neither	Addition facts-L1 (out of 20, prorated score)	0.733	0.362	0.044	9.64	8.91	1931	0.017
3	Neither	Addition facts-L2 (out of 5)	0.364	0.122	0.003	3.05	2.68	1931	0.037
3	Neither	Subtraction facts-L1 (out of 20, prorated score)	0.677	0.324	0.038	7.49	6.81	1931	0.027
3	Neither	Subtraction facts-L2 (out of 5)	0.293	0.115	0.012	2.38	2.09	1931	0.025
2	BL only	Phoneme isolation (out of 10)	-0.098	0.154	0.525	4.72	4.81	1931	0.181
2	BL only	Syllable segmentation (out of 10)	0.654	0.227	0.004	6.87	6.21	1931	0.187
2	BL only	Letter sound (out of 100, prorated score)	-1.465	1.246	0.241	32.33	33.79	1931	0.190
2	BL only	Reading vocabulary (out of 10)	0.077	0.120	0.523	8.18	8.11	1931	0.358
2	BL only	Passage reading (out of 52, prorated score)	0.427	0.625	0.495	15.13	14.70	1931	0.446
2	BL only	Reading comprehension (out of 6)	-0.029	0.093	0.756	1.51	1.54	1931	0.332
2	BL only	Counting numbers (out of 60)	-0.170	0.381	0.655	36.28	36.45	1931	0.225
2	BL only	Counting objects (out of 10)	-0.068	0.048	0.157	9.82	9.89	1931	0.055
2	BL only	Number identification (out of 20, prorated score)	0.035	0.651	0.957	32.50	32.47	1931	0.420
2	BL only	Number discrimination (out of 10)	0.038	0.106	0.724	9.13	9.10	1931	0.256
2	BL only	Missing number (out of 10)	0.029	0.124	0.816	6.84	6.81	1931	0.435
2	BL only	Addition facts-L1 (out of 20, prorated score)	-0.406	0.335	0.226	5.41	5.82	1931	0.277
3	BL only	Syllable segmentation (out of 10)	0.831	0.232	0.000	6.07	5.24	1931	0.135

Grade	Model	Dependent Variable	coeffT	se	pval	meanT	meanC	fullN	rsquare
3	BL only	Letter sound (out of 100, prorated score)	-2.498	1.272	0.051	34.10	36.60	1931	0.193
3	BL only	Non-word decoding (out of 50, prorated score)	0.349	0.597	0.560	12.37	12.02	1931	0.391
3	BL only	Reading vocabulary (out of 10)	-0.019	0.107	0.860	9.08	9.09	1931	0.313
3	BL only	Passage reading (out of 52, prorated score)	-0.472	0.648	0.468	22.41	22.88	1931	0.543
3	BL only	Reading comprehension (out of 6)	-0.125	0.085	0.140	2.68	2.80	1931	0.408
3	BL only	Number identification (out of 20, prorated score)	0.560	0.582	0.337	35.61	35.05	1931	0.545
3	BL only	Number discrimination (out of 10)	0.023	0.086	0.790	8.59	8.57	1931	0.445
3	BL only	Missing number (out of 10)	-0.002	0.101	0.983	8.35	8.35	1931	0.482
3	BL only	Addition facts-L1 (out of 20, prorated score)	-0.244	0.340	0.475	9.15	9.39	1931	0.427
3	BL only	Addition facts-L2 (out of 5)	0.033	0.109	0.761	2.88	2.85	1931	0.397
3	BL only	Subtraction facts-L1 (out of 20, prorated score)	-0.216	0.315	0.493	7.04	7.26	1931	0.415
3	BL only	Subtraction facts-L2 (out of 5)	-0.007	0.106	0.948	2.23	2.24	1931	0.379

ANNEX IV: MSI COMPARISON BRIEF OF RAMP STUDIES

Comparison Table: RTI and MSI Results	
RAMP Implementation (RTI)	**RAMP Impact Evaluation (MSI)**
Study Design	
• Performance evaluation • No counterfactual • Cross-sectional studies	• **Quasi-experimental impact evaluation** • Intervention and comparison groups • Counterfactual allows attribution to investigate change over time • Longitudinal: Three points to measure change over time
Sampling	
• Different samples for the two studies • 2014: From 156 schools (nationally representative), purposively selected 43 schools for pilot • Unclear if remaining 110 schools nationally representative • 2016: 240 schools • Small schools excluded • Cannot draw conclusions from different samples	• **Following same schools over time** • Randomly selected schools based on RAMP rollout (120 intervention, 120 comparison schools for total of **240** schools) • Statistically representative of Cohort 2 and Cohort 3 • Propensity Score Matching at 2 levels (schools and students) to ensure baseline equivalency • Following same students, schools, teachers over time (G3) • Thus, differences can confidently be attributed to intervention
Tools / Instruments	
• Used same tool for G2 and G3 students, which doesn't acknowledge differences in curriculum and grade level standards • 2014 study: based on 2012 curriculum; 2016/2017 study: revised, unclear if based on 2016 curriculum	• **Separate tools for G2 and G3** • Both baseline and midline tools based on 2016 curriculum for each grade • Tools calibrated to standards at beginning and end of school year • Developed in a collaborative process with MoE
Inferences: What does the approach tell us?	
• Descriptive info on students' outcomes • Identification of foundational skills students struggle with; Determine focal areas to help students meet benchmarks • If the sample of 110 schools is nationally representative, study could provide data on trends, but not causal changes • National level results include students who have not been exposed to RAMP	• Descriptive info on students in Cohorts 2 & 3 • Allow us to draw causal conclusions about impacts and, with some confidence, attribute changes to intervention • Student learning in reading and math • Teacher practices in G3 teachers

Comparison Brief of RAMP Studies:

RAMP Implementation vs. RAMP Impact Evaluation

In the second half of 2017, RTI and MSI submitted reports aimed at establishing whether the early grade reading and math program (RAMP) had an *impact* on students learning outcomes. Per the request of USAID, to clarify why findings and conclusions differ between the two studies, this document (1) describes key requirements to estimate causal impacts of an intervention like RAMP, (2) identifies key differences in the study designs that RTI and MSI used, (3) notes differences in instrumentation, and (4) explains what inferences can (and cannot) be drawn from each study.

What is required to establish whether RAMP *had an impact* on student outcomes?

Both RTI's and MSI's studies claim to establish whether RAMP had an *impact* on students' reading and math outcomes. Assessing the differences between the two studies necessitates understanding the requirements of an *impact evaluation*; that is, an evaluation that allows answering causal questions about an intervention's impacts.

USAID's Evaluation Policy (2011, updated in 2016) defines impact evaluations as those that "measure change in a development outcome that is attributable to a defined intervention" (p. 3). To attribute changes in student outcomes to RAMP or to any other intervention, impact evaluations need a *counterfactual*, that is, a measure of "what would have happened in the absence of RAMP" (USAID, 2013). "Impact evaluations in which comparisons are made between beneficiaries that are randomly assigned to either a treatment or control group provide the strongest evidence of a relationship between the intervention and the outcome measured" (USAID 2011, p. 3).

The key challenge of any impact evaluation is constructing a counterfactual that allows a comparison of students with and without an intervention so that differences in outcomes can be compared. However, we cannot observe the same students (or group of students) simultaneously in the two conditions (with and without RAMP). Thus a credible impact evaluation rests on its ability to find a plausible approximation for comparison. The comparison group must be rigorously assessed to ensure that it is an appropriate counterfactual or comparison to the intervention group. Without a credible counterfactual, it is not appropriate to attribute changes in outcomes to the program.

According to USAID's Evaluation Policy (2011), *performance evaluations* differ from impact evaluations in that they "often incorporate before-after comparisons but lack a rigorously defined counterfactual" (p. 3). As such, *they can only demonstrate whether change has occurred, but cannot establish what actually caused the observed change* (USAID, 2013; emphasis added). Without a rigorous approximation to the counterfactual, there are a myriad of alternative explanations for the observed differences. Therefore, performance evaluations cannot claim that the intervention caused observed changes.

Based on the above definitions, *RTI's study is a performance evaluation*, whereas *MSI's study is an impact evaluation*. The remaining of this document centers on describing the counterfactual of each study and its implications for the interpretation of results.

RTI's "Early Grade Reading and Mathematics Initiative: Midline Survey Report"

RTI aims to estimate the counterfactual with data collected from G2 and G3 students in 110 schools in 2014, as part of the Intervention Pilot Research Activity implemented in 2013-2014. RTI selected 110 schools out of a nationally representative sample of 156 schools. From this sample, 43 schools were *purposely* selected to receive the Pilot treatment. The report does not specify whether the remaining schools were still a nationally representative sample once the 43 pilot schools were excluded. The authors refer to data collected from those 110 schools as "baseline" data.

Next, in 2016/2017, RTI collected "midline" data from G2 and G3 students in a sample of 240 nationally representative schools, between the end of 2016 and the beginning of 2017. The sample excluded schools without at least 20 students in G2 and G3 (combined), 240 schools sampled for MSI's evaluation, and 99 recently established schools.

RTI's main analysis compares outcomes for students in the "baseline" and "midline" samples. However, this comparison cannot be used to draw causal conclusions about the impacts of RAMP because it is comparing outcomes of two samples at two different points in time. While it is possible to estimate differences in indicators between the two samples across the two periods, the difference cannot be attributed to the intervention given that there may be other causes for the changes that cannot be ruled out. Alternative explanations (or confounding factors) for the differences observed can include "interventions run by other donors (or the Ministry of Education), natural events (e.g., rainfall, drought, earthquake, etc.), government policy changes, or natural changes that happen in an individual or community over time" (USAID 2013, p. 2).

Two additional issues threaten the validity of RTI's conclusions: (1) It is unclear whether the sample of 110 schools is nationally representative and (2) the midline sample includes students with varying levels of exposure to RAMP, including children that have not been exposed to RAMP. We elaborate on these issues below.

Again, it is unclear whether the 110 schools in the baseline sample constitute a nationally representative sample. Even if we assumed that no other factors could have led to differences in students' outcomes between 2014 and 2016/17, the 110 schools could have differed from schools in the midline sample before the introduction of RAMP. Comparing two samples from dissimilar populations is problematic because differences could be related to the sample composition, rather than to the intervention. Importantly, impact evaluations follow rigorous procedures to ensure that samples are similar for all or most measurable characteristics so that the only difference will be exposure to the intervention. For example, the school selection used for the Pilot activity could have produced a sample of 110 schools that, on average, were poorer or had a larger share of underperforming students than schools in the nationally representative sample. If that were the case, it would be reasonable to expect students in the 2016/2017 sample to perform better than students in the 2014 sample. Further, the report indicates that 43 schools were *purposively* selected (as opposed to *randomly* selected) to receive the Pilot activity in 2014. The rationale for the purposive sampling was not provided but there could be something about these schools that makes them non-representative of the typical Jordanian school. There is no analysis of the pilot group or the larger sample to determine whether they are representative of all Jordanian schools.

The report also states that approximately 50% of the students in the study had not been exposed to RAMP by the time of the midline assessment (p. 4). This means that the results presented in Tables ES3 and ES4—even *if* attribution were possible from a performance evaluation—RAMP only reached half the sample. The report further states that, due to the variation in exposure, significant differences observed in G2 are encouraging and null findings in G3 are not reason for concern. Yet, the overall positive trends that RTI attributes to RAMP include significant differences for untreated students (see Tables 9 and 14 for example), which *indicates that factors other than RAMP* underlie the findings.

Despite not being an impact evaluation, *if* RTI were able to provide evidence that the 110 schools are indeed a nationally representative sample of Jordanian schools, then the study could make important contributions to understanding students' early reading and mathematics. The study, better characterized as a *performance evaluation*, provides useful descriptive information about national trends in G2-G3 students' outcomes from 2014 to 2017. Moreover, it identifies discrete foundational reading and math skills where students need additional support to meet performance benchmarks. Finally, it offers descriptive information about Syrian students, which is not available from other sources and can be useful for education authorities in the region.

MSI's "RAMP Impact Evaluation: Estimating Impacts of Early-Grade Reading and Math Project (RAMP) in Jordan": Midline report.

In the most rigorous impact evaluation design—a randomized control trial—schools would be randomly assigned to the intervention or comparison groups before implementation (USAID, 2011). Random assignment is the best method to create a credible approximation to the counterfactual, because it increases the likelihood that that the groups being compared are similar before introduction of the intervention. MSI was unable to randomize schools for this impact evaluation, because RTI planned to implement RAMP at the governorate level and had already determined that cohort 2 governorates would receive the intervention in 2017.

The evaluation used a quasi-experimental design (QED) to approximate the counterfactual, using *propensity score matching* to create groups that were as similar as possible at baseline. USAID (2013) notes that matching is "the most common means for selecting a comparison group, wherein the evaluator selects a group of similar units based on observable characteristics that are thought to influence the outcome" (p. 5). MSI's evaluation used a two-step propensity score matching process with school-level characteristics and later with student data. The midline report shows that, after matching, impact inferences are based on groups of students that are statistically equivalent at baseline, according to strict criteria set by the Institute of Education Science's What Works Clearinghouse[13]. Having equivalent groups minimizes the risk that between-group differences observed at midline (or endline) are due to factors other than RAMP.

MSI's evaluation is not without limitations. A key limitation is that matching cannot account for differences in unobservable (or unmeasured) characteristics. There is a risk that statistically equivalent groups may differ in variables not included in the matching process, which could lead to erroneous conclusions about the intervention's impacts. Most relevant to this document is

[13] The What Works Clearinghouse (WWC) is an initiative of the U.S. Institute of Education Sciences to evaluate studies on the effectiveness of programs, policies and practices. WWC Standards Briefs lay out rules to assess the quality of studies and are highly regarded in the field of program evaluations.

that while RTI's 2016/2017 are generalizable across the country, results from MSI's evaluation are generalizable to cohort 2 schools only (although there were differences in the EGRA/EGMA instrumentation for the two studies.) The reasons are that (1) cohort 1 governorates were not included in the study because the implementation of RAMP was already underway in those governorates, and (2) propensity score matching created a sample of cohort 3 schools and students that was as similar as possible to cohort 2 schools to credibly approximate the counterfactual.

Instrumentation

There were also differences in the EGRA/EGMA instrumentation used to measure students' early grade reading and math outcomes. First, RTI used the same tool for grade 2 and grade 3 students. In contrast, the MSI evaluation used separate EGRA/EGMA tools for students in grades 2 and 3. These tools were developed in a collaborative process with the Ministry of Education. The MSI tools were matched to the Ministry's most recent 2016 curriculum.

Conclusion

The reports prepared by RTI and MSI summarize different studies aimed at estimating students' reading and mathematics outcomes. However, as per USAID's Evaluation Policy each RTI and MSI differ in their studies. RTI's study can be characterized as a performance evaluation, and cannot be used to make inferences about the impact of RAMP on students' outcomes directly. MSI's study, in contrast, is an impact evaluation draws causal conclusions about the impacts of RAMP (or lack thereof). Differences in the designs of the two studies, and differences in instrumentation, explain the lack of convergence in the results and conclusions.

References

USAID (United States Agency for International Development), 2011. *USAID Evaluation Policy*. U.S. Agency for International Development, Washington D.C. Retrieved on December 5, 2017, from https://www.usaid.gov/sites/default/files/documents/1870/USAIDEvaluationPolicy.pdf

USAID (United States Agency for International Development), 2013. Impact Evaluations Technical Note. *Monitoring and Evaluation Series*. Bureau for Policy, Planning and Learning: Retrieved on December 5, 2017, from https://usaidlearninglab.org/sites/default/files/resource/files/ie_technical_note_2013_0903_final_2.pdf

ANNEX V: RTI COMMENTS ON MSI ENDLINE IE REPORT

Baseline Differences

Annex X describes the differences in student level learning outcomes for the weighted and unweighted sample of learners at the baseline. The differences are non-trivial, with Table D.1 in the Annex showing up to a .46 standard deviation difference between the treatment and control group. As we have argued in previous discussions on this topic, we think that it is entirely inappropriate to remove these baseline assessment differences. Surprisingly, the endline report makes no mention of these differences, though RTI has shared this concern about various versions of the midterm draft several times. Our argument comes from several sources, summarized below.

1. The RAMP baseline assessment was undertaken up to weight weeks after the intervention began in Cohort 1 schools. During this period some of the more meaningful gains are likely to have occurred. Our work elsewhere shows that initial gains are some of the largest impacts that these sorts of programs have (Piper et al, 2018).
2. The RAMP school-level matching exercise shows that very few of the comparisons between treatment and control schools were statistically significantly different. After utilizing the school level matching that the RAMP evaluation employed, we think it is very likely that the school level matching accounted for the differences at the baseline, and using the student level matching after the intervention started is removing some evidence of the treatment effect, if it exists
3. The RAMP external evaluation is in the minority in the employing this student-level matching strategy. There are many evaluations of program impact employed by USAID programs focused on reading, and none of the other evaluations are employing a student level matching program, let alone one employed after the intervention started. We would recommend that the RAMP external evaluators provide evidence of other evaluations employing this methodology amongst USAID's reading programs.
4. The period between RAMP implementation beginning and the baseline is a substantial proportion of the available instructional time prior to the midterm assessment, given the condensed timeframe for implementation within that first year of implementation. Removing that large a portion of instructional improvement is non-trivial.
5. Table D.1 shows much smaller differences between treatment and comparison groups for student background characteristics than in learning outcomes. The argument that none of these differences in learning outcomes are attributable to the RAMP intervention is simply hard to make, as if the differences were due to prior skills differences, we would expect differences at a similar magnitude to the differences between the school level or student background characteristics. The argument that the only place where large differences exist between the treatment and control schools is in learning outcomes, rather than student background characteristics or school characteristics, is hard to believe, particularly when a more plausible alternative explanation exists.

We recommend the following solution to this baseline assessment problem. This is the same solution we suggested in our responses for the midterm report, which was not employed at that point, but we believe should be employed here. We recommend that the MSI report either present impacts of RAMP using the un-adjusted baseline scores for the treatment group or present the range of RAMP impacts depending on baseline scores are employed. Table 1 present a range of impacts of RAMP rather than a solitary estimate. The interpretation of the RAMP impact is obviously very different based on whether the baseline weighting for treatment schools is applied, and we would like to provide the MOE and the program with more policy relevant information.

Assessing RAMP Approach

In addition to our concerns about how the RAMP impact estimate was derived based on how the baseline differences were treated, we are also concerned about how the tools employed by the external evaluation measured the RAMP approach. We were hoping to be able to use the findings of the RAMP evaluation to implement improvements to our approach and design. Despite our attempts to suggest how the tools at the midterm and endline could be improved, the usefulness of the endline assessment for the program implementer to improve delivery is lessened because of what appears to be a limited understanding of the RAMP approach in the measures and analyses. We are glad to see that in Table 1, there are analyses on whether the RAMP strategies were implemented, and whether RAMP and MOE curriculum are integrated. But several of the other measures presented in the findings table are evaluating RAMP on what are not relevant measures that RAMP is focused on.

We would have liked the RAMP evaluation to focus on more fine-grained analyses of the specifics of the RAMP approach rather than looking at more general tasks that are not related to what RAMP was implementing. We are left without findings as to how often RAMP materials are being used, how well teachers plan using the RAMP approach, how often skills trained on by RAMP are used, nor how often RAMP is integrated into the daily instructional behavior of teachers. All of these are important points that would have been helpful for the program to know as it works to improve its outcomes in the last year.

Instead, the RAMP evaluation utilizes a set of classroom assessment and teacher interview tools that are unrelated to the core instructional approach of RAMP. For example, a great deal of the findings presented in Table 1 are evaluating RAMP's impact on a variety of measures that RAMP was never designed to impact. Including a table of somewhat random measures and whether RAMP impacts them is inappropriate, as we have argued since the midterm findings report.

Specific comments

12 – We wonder whether the differences between the single-shift schools and double-shift schools are affected by what we think is the over-fit model of the external evaluator. Adjusting for school level baseline characteristics and student level characteristics might make it difficult to determining the impact of RAMP on these measures. We wonder how more parsimonious models would evaluate these findings

13 – We worry that the difficulty that the external evaluator had in understanding RAMP has led them to recommend changes to the program design that are based on the lack of their understanding of the approach rather than that RAMP should redesign their theory of change. This is not to say that RAMP is not working to improve clarity of implementation efforts, in fact, RAMP has revised its focus for Semester 2 2019 in response to specific implementation challenges.

13 – the recommendation on program implementation seems to be written by people unaware of the pilot work that preceded RAMP and unaware of the coaching and mentoring visit data available from RAMP implementation

14 – The statement that "implementers should guard against bias during data collection due to over sampling from stronger regions or schools" should be substantiated or removed. Both the national EGRA and LQAS efforts use sampling weights to ensure that the results are externally representative of the populations of interest

14 – The comment that "it would have been beneficial to have access to program materials at the onset of the project to inform measurement decisions" seems to imply that RAMP has not shared materials with the external evaluator which is clearly not the case

14 – We agree that the condensed time period for the RAMP implementation is problematic. Given that it is only in Semester 1 of the 2018-2019 academic year that the RAMP program has had its materials

and structures embedded in the MOE teachers' guide, we think that additional time is needed to determine the impact of the program now implemented within the MOE daily instructional program

19 – this section shows that the external evaluator is aware of the studies that preceded RAMP. Could this information have been included in the sections above where the evaluator is asking about the process of piloting and the instructional design?

22 – The teacher study design figure suggests that there were no teacher study analyses done in the endlne. We wonder why there are findings on the teacher study part of the report, then? The report does not do a good job explaining that the findings on classrooms are from previous assessments, if that is how the study was done

22- Given the reduced timeline for this student assessment and this cohort analysis, we reiterate our recommendation that the differences at the baseline that were observed should not be removed from program impact

24 – The text box does not address the key question of what an external evaluation should do when the "baseline" assessments are undertaken after the intervention began, and we are surprised to see such a strong statement about this issue given prior discussions about how concerned we are about how the baseline differences were dealt with. The statement that "IEs rely on comparing similar groups at baseline, prior to an intervention" was clearly not implemented in this case, given the timing of the baseline assessment. The final statement that "differences were due to pre-existing between-group differences-can be ruled out" is difficult in this circumstance, and the report should respond to that reality.

25 – the footnote does not share with the reader that the 6 months of exposure include some time period prior to the baseline

25 – the explanation for G1 is probably not how we would have presented it. For individual children, both the intervention group and the comparison group have received one year of intervention, since they have only been in primary school for one year. The 2 years of intervention for the intervention group refers to the school level, not the student level. This should be accounted for in the analysis and the explanation

26 – the explanation for how the comparison schools were selected is clear and logical. We think that the baseline corrections should have stopped there, use the matching exercise that you have done to select similar schools to serve as the baseline correction rather than overfitting the estimates by also having student level corrections made to the estimate, given the timing of the baseline assessment vis a vis the intervention start.

27 – what proportion of Grade 3 teachers were not available in the endline observation study?

27 – we note that there is not much included in this study on the perceptions of the community and MOE on RAMP, which as we have seen from previous presentations from RAMP was quite positive.

27 – Table 4 does not indicate anywhere that the baseline data was collected after the intervention began

27 – utilization of the COTI in the Jordanian context seems to be part of the problem. The COTI was not fit for purpose for the RAMP intervention and includes a wide range of interesting information that is completely unrelated to the specifics of RAMP. This is a large part of why the RAMP external evaluation of classroom practices appears to show little impact, it is because a tool used in different countries and on different projects was pre-determined to be applied to RAMP, when RAMP's design required a specific set of tools to be newly developed to apply to the particularities of RAMP's theory of change. RAMP's design and theory of change has very little connection to many of the areas that COTI assesses. We have pointed this out in several previous discussions on the external evaluation. It is where

the COTI was dramatically adjusted to respond to the RAMP realities where the assessment was able to identify RAMP program impact.

30 – How do the OLS regressions work for the RAMP impact analysis? Were the baseline differences removed using a DID model, or controlled for in the regression?

30 – the textbox once again does not address the issues of the baseline differences coming after the intervention began. These are therefore not "pre-existing differences between the groups" as the textbox says. We think that you are over-fitting your model and you can use the school level matching and do simple comparisons at the midline and the endline, depending on the matching of the schools rather than removing the baseline differences of student learning outcomes since we think this was due to actual learning gains from the program

31 – we are concerned that this is simply a rehash of analyses done at the midline, when the tools were not fit for the RAMP intervention

32 – as discussed above, the instructional practices in the COTI tool were not related to RAMP, so it is expected that there wouldn't be differences in areas where RAMP isn't working to make a difference. Writing the report this way it appears that there is a lack of impact of RAMP on instruction, when RAMP wouldn't have an impact on these areas

32 – for this classroom analysis, you did not use baseline analysis, which may be fine because you worked to have the school level analysis, as we have discussed elsewhere

32 – as discussed above, the Grade 3 classroom study is still evaluating RAMP teachers on instructional practices that are not part of the RAMP intervention

33 – what are the implications of having 17% of comparison teachers being trained? Maybe this is from transfers?

34 – We think it's worth including the spillover issue earlier. Large scale programs like RAMP often suffer from this problem, and it's worth indicating earlier in the report that this could reduce the identifiable program impact. Your point that the number of coaching visits were lower than planned for ignores that the amount of coaching was quite substantial compared to other large scale programs in the region

34 – Table 6 represents the key parts of the RAMP intervention and we think that this should be given a much earlier discussion, including in the Executive Summary, rather than the presentation of assessments on tools unrelated to RAMP.

36 – "there were no clear patterns indicating that the program was implemented with fidelity" this statement seems contradictory to Table 6. In addition, given that the effect size of the teacher demonstrating RAMP strategies and integrating MOE and RAMP were both above .53 SD, we think this should be reworded.

36 – when comparing baseline and endline on the lesson content, keep in mind that the time of year was different, meaning that different skills are emphasized at different parts of the year. So much of this analysis in Table 7 is probably misleading

37-38 – the comparisons presented in Table 8 are largely unrelated to the RAMP intervention and we are unsure why these sort of comparisons remain in the report, particularly presented the way they are, as if RAMP would change these outcomes. These outcomes come from COTI, not fit for the RAMP context.

39 – stating that the RAMP intervention has both positive and negative impacts on classroom practice is inappropriate given how much of the evaluation depends on tools not developed for the RAMP

intervention. We recommend focusing on the parts of the intervention that are related to the RAMP intervention.

39-40 – Table 10 presents the key instructional methods that RAMP was working on, and we think that this should be given much more emphasis in the executive summary and in this section. For the number identification and writing finding, note that the time of year influences what is being done.

41 – Table 11 again presents findings on a tool that we do not think is related to the RAMP intervention. Including it here makes the reader think that RAMP was designed to improve these outcomes and it didn't, rather than that the external evaluator decided to use a tool that was unrelated to the intervention and then assess the intervention using it.

44 – we disagree with many statements on this page. Once again, the evaluation is not accounting for the fact that the baseline happened after the intervention began. We do not think that the evidence shows there were "minimal impacts on teachers' instructional practices" given what your findings show, as there were consistently large impacts on the elements that you assess that RAMP was supposed to do. We recommend rewording this discussion

44-45 – we think that the textbox should be updated after consideration is given to the baseline differences, as should Table 13

45 – Table 13 should be more specific about how the measures are derived. When the table says that, for the endline, G2 at base and midline and G3 at endline, are these just the endline measures? Or not? This is confusing, particularly when considering the complexity of how the baseline differences are accounted for

46 - Table 14, we would recommend refitting to show the results without adjusting for the baseline scores (as you describe in your note), or showing the results from both analysis methods. We note also that a couple of schools seem to have been removed from the intervention sample, can you discuss this please?

47-48 – We really like these figures. Could you create them for models that do not adjust for the baseline differences as well?

48 – Table 15 needs to have the same baseline adjustment removed, or at least the comparisons of impacts when they are removed from when they are not

49-50 – please recreate these figures for when the baseline differences are not adjusted

51 – Table 16 accounts for baseline reading and math scores, but does it make sense to do it that way when this is an entirely new cohort of children at Grade 1?

52-53 – when you are doing your single shift vs. double shift analysis, we wonder if you should refit without the school level matching and controls, since one or two schools of particular groups could have a significant impact on these estimates.

54-62 – we have the same comments for Mathematics as we do for the Reading results. Given our suggestions in previous discussions to include the range of outcomes, we are surprised that didn't happen in this version of the report

63 – you cite Schochet (2008) but we do not think you can cite any other evaluations of USAID funded reading or math interventions that overfit their models with both school level and student level propensity score matching to remove the baseline differences. We think it is important to share the results depending on the school level matching and not making the baseline score correction that you are making. The fact that you are presenting the baseline data collection problem on page 63 with no mention of it earlier is very concerning. Finally, you mention that the results presented in Annex O show that the conclusions do not change when baseline scores are excluded. Having read Annex O, I do

not see evidence of that. And if that was the case, why not present the more parsimonious model with the conservative estimate rather than having the RAMP evaluation be the only one using this technique, ignoring the gains that could have ensued from RAMP in the initial key part of the program. When I read Annex P, it seems that the results do change significantly when the baseline results are not included, and we would recommend using the unadjusted model, since the model is not overfit and the most plausible explanation for the differences at baseline after the school level propensity matching was used.

63 – we agree that the short time period of the intervention makes it difficult to detect impacts. We would encourage the use of future analyses to see if the program has an impact

63 – Better understanding the impact of the 17% of teachers who were in RAMP trainings and the potential impact is important. We think a follow-up study of spillover might help explain things.

66 – "Incentives for teachers to implement RAMP are not aligned with MOE curricula" is a statement that we think is unclear. Is the point about the lack of incentives for program implementation, or whether the RAMP materials are aligned with the MOE.

68 – These are useful discussions of how RAMP has been implemented. It will be worth investigating if this changes as the RAMP program's methods are incorporated in the MOE teachers' guide in 2018-19

68 – 15 days of training is not considered as short, in comparison to other programs

70 – these are practical discussions of how teacher onboarding and turnover affects the program

72 – the discussion of the impact of teachers is using an inappropriate measure of impact ("teachers' use of instructional time, student engagement or classroom management" We would suggest changing this paragraph to focus on what RAMP was actually designed to do with teachers. The evaluation remains critical of RAMP's ability to impact outcomes that were not part of its design, but where part of the COTI tool that was not designed for RAMP.

74-77 – the comments elsewhere in the document should be applied to the recommendations

77 – we agree with the need for more time to estimate the impact of RAMP

Annex P – thank you for presenting the unadjusted scores, without corrections from the baseline, though we request some clarifications:

- We assume that these are the raw outcomes from the children at various levels at the endline only, correct?
- If so, these are the outcomes we recommend sharing in the report, rather than the adjusted ones.
- We disagree with the idea that there is no "persuasive evidence to suggest that RAMP had an early impact on students' outcomes after only 6 to 8 weeks of implementation."
- We are actually more concerned about how the scores should account for the 2 schools that appear to have dropped out of the evaluation, and how that is adjusted for (should the matched schools in the other comparison group also be removed)?
- It appears that you are still potentially overfitting the models by including the home language and attending preschool at the student level, as that should be accounted for by the school level matching.
- What are the raw results without any of the student level control variables?
- It appears that the vocabulary measure has a ceiling effect, and it looks like it should not be included in the estimates as a result.

U.S. Agency for International Development
Middle East Bureau
1300 Pennsylvania Avenue NW
Washington, D.C. 20004