SCHOOL-TO-SCHOOL INTERNATIONAL

UNIVERSITY OF MALAWI

PILOT IMPACT EVALUATION ENDLINE REPORT

NATIONAL NUMERACY PROGRAMME IN MALAWI

March 2023

CAMBRIDGE EDUCATION

UNITY AND FREEDOM

UKaid from the British people

Disclaimer:

This document is issued for the party which commissioned it and for specific purposes connected with the above-captioned project only. It should not be relied upon by any other party or used for any other purpose.

We accept no responsibility for the consequences of this document being relied upon by any other party, or being used for any other purpose, or containing any error or omission which is due to an error or omission in data supplied to us by other parties.

This document contains confidential information and proprietary intellectual property. It should not be shown to other parties without consent from us and from the party which commissioned it.

CONTENTS

ACRONYMS

CERT	Centre for Educational Research and Training
EGMA	Early Grade Mathematics Assessment
FCDO	United Kingdom's Foreign, Commonwealth & Development Office
KII	Key Informant Interview
PPS	Probability proportionate to size
MOS	Measure of size
MDES	Minimum Detectable Effect Size
MoE	Malawi Ministry of Education
NNP	National Numeracy Programme
PPS	Probability Proportionate to Size
STS	School-to-School International
TLC	Teacher Learning Circles
USAID	United State Agency for International Development

LIST OF TABLES

LIST OF FIGURES

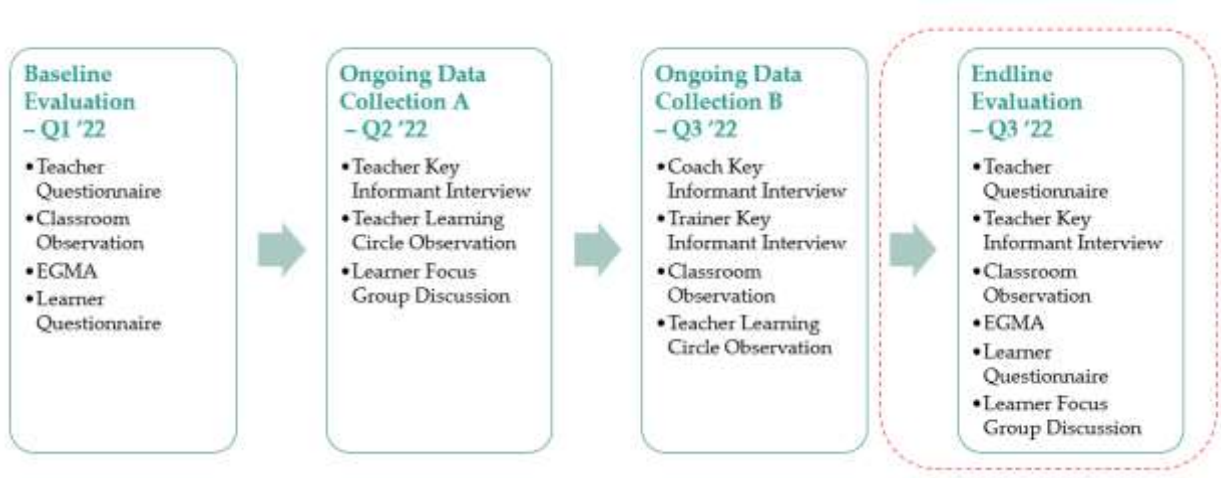
EXECUTIVE SUMMARY

EVALUATION PURPOSE

The evaluation of the National Numeracy Programme (NNP) pilot aims to gather information on participants' views about the programme, assess the mathematics skills of learners in treatment schools participating in the NNP pilot compared with learners in comparison schools, examine how teachers have changed their approach to mathematics instruction and provide insight into the efficacy of NNP materials and the in-service teacher training methodology. Findings will inform the strengthening of the NNP prior to scaling it nationwide, especially in terms of the NNP's materials and teacher training components.

School-to-School International (STS), in partnership with its Malawi-based partner, the Centre for Educational Research and Training (CERT), is conducting an independent evaluation of the NNP pilot at four different time points, as displayed in Figure 1—baseline, ongoing data collection periods, and endline. The endline occurred from 15–26 August 2022.

Figure 1: Timeline for the pilot evaluation[1]



Nine questions guide the pilot evaluation:

1. Have pilot activities improved learner performance?
    a. *Under what conditions have pilot activities improved learner engagement and performance?*
2. Have the pilot activities changed learner engagement?[2]
3. How are learners engaging with and using the workbooks independently?
    a. *Are there language issues that impede on learner engagement with the materials?*

[1] The classroom observation tool was revised for ongoing data collection B and endline.

[2] In comparison with the observed learner behaviours in the pre-study classroom observations, learners are (a) doing more "independent" work; and (b) working in the workbooks. Comparison is with scoping study produced prior to start of project.

   b. *Is learner engagement with the learning materials equitable with regard to gender and learners with learning difficulties?*

4. To what extent have the pilot intervention's activities changed classroom practices?[3]
5. In what ways are face-to-face teacher trainings changing teacher practices in the classroom?
   a. *Under what conditions are teacher trainings being implemented as intended?*
6. Are the pilot coaching sessions and Teacher Learning Circles (TLC) working as intended?
   a. *Under what conditions are the coaching sessions effective?*
7. In what ways are the teacher guides supporting teachers to understand the methodology/approach being implemented?
8. In what ways do the teacher guides support teachers in implementing the intended methodology/approach and use learner materials effectively and with fidelity?
9. How are the training videos being used?
   a. *Are the training videos perceived as a useful training resource?*
   b. *How could they be improved and made more useful?*

PROJECT BACKGROUND

Led by the Malawi Ministry of Education (MoE) and funded by the United Kingdom's Foreign, Commonwealth & Development Office (FCDO), the four-year NNP aims to improve outcomes in mathematics so girls and boys have a solid foundation in basic skills to succeed in the rest of their schooling and fulfil their potential. The NNP aims to develop a new vision for teaching and learning mathematics in Malawi in which children experience mathematics as a meaningful, sense-making and problem-solving activity. Learners will be expected not only to know mathematics but also to understand the mathematics they know, apply it to solve unfamiliar problems, and reason and argue using the mathematics they develop.

The NNP is being implemented in three phases: inception, pilot, and scale-up. During the initial inception phase, the project developed a vision for mathematics teaching and learning in Malawi; new teaching and learning materials for standards 1–4 for piloting; a plan for revising the curriculum; and a training methodology for piloting. The pilot phase, which took place during the 2022 school year, involved piloting new materials and teacher training approaches in 200 schools across all divisions of Malawi. The project will culminate with the nationwide scale-up phase.

[3] Intervention activities include teacher guides; learner materials; face-to-face training; and school based CPD, such as TLCs & coaching.

EVALUATION METHODS AND LIMITATIONS

STS is evaluating the effectiveness of the NNP pilot phase by employing a mixed-methods, pre-post approach using a quasi-experimental design. STS provides technical leadership and oversight of all components of the pilot evaluation, while STS's Malawian counterpart, CERT, manages all in-country logistics for training and data collection.

The pilot evaluation employs a diverse set of instruments targeted at different stakeholders and participants. Finalized during a pretest in November 2021, these instruments include learning and knowledge assessments, interviews, classroom observations, and questionnaires:

- Quantitative inquiry — deductive approach
    - Early Grade Mathematics Assessment (EGMA)
    - Teacher and learner demographic questionnaires
    - Classroom observation form[4]
- Qualitative inquiry — inductive approach
    - Teacher, coach, and trainer key informant interviews (KIIs)
    - Teacher Learning Circle (TLC) observation form
    - Learner focus group discussion (FGD)

All quantitative tools were administered at endline, as well as several qualitative tools—KIIs with teacher KIIs and FGDs with learners. The EGMA captures learners' knowledge of numeracy skills and is largely based on the version developed in 2010 by the United States Agency for International Development (USAID) and the MoE.[5] It includes two versions—one for standards 1–2 and the other for standards 3–4—and includes the following subtasks—number identification, addition and subtraction level 1, quantity discrimination, pattern completion, and problems. The standard 3–4 EGMA also includes addition and subtraction level 2 subtasks. The classroom observation tool administered at endline was revised from baseline to update the items used for indicator scoring and to include items on the quality of mathematics instruction.

For the pilot evaluation, treatment and comparison schools were sampled using a three-stage clustering random method. For the first stage, schools were randomly selected using a probability proportionate to size (PPS) without replacement approach. When visiting selected schools, data collection teams used a simple random approach to randomly select one classroom per standard and then a sample of boys and girls in each classroom.

[4] The classroom observation tool administered at endline was revised from baseline to update the items used for indicator scoring and to include items on the quality of mathematics instruction.

[5] USAID/Malawi and MoEST. USAID Funded Malawi Teacher Professional Development Support (MTPDS) Activity 2010 Early Grade Mathematics Assessment (EGMA): National Baseline Report 2010.Washington, DC: USAID, 2010.

The endline data collection training utilised a cascade model. STS conducted remote training with the principal researchers for three days from 3-5 August 2022. Following the remote training, the principal researchers conducted a five-day training with the research assistants in-person in Lilongwe from 8–12 August 2022. The NNP Technical Lead facilitated training of the classroom observation form.

Visiting one school per day, eight teams of four—one supervisor and three research assistants—conducted the endline study from 15–26 August 2022. The teams visited 40 comparison schools and 35 treatment schools in 17 districts across Malawi. A total of 296 teachers and 1,489 learners from standards 1–4 participated in the endline study.

The pilot study's evaluation questions guided the analysis. EGMA and questionnaire data were coded and analysed in Stata following best practices outlined in the EGMA toolkit guidance.[6] All items or questions were analysed individually, with means, standard deviations, and frequencies produced for each variable. In addition, data was aggregated, as needed, to respond to each evaluation question. To correct for the unequal probability of selection due to clustering of the sample, survey weights were computed with a two-step procedure and included in all analyses.

The pilot evaluation includes some limitations. The short time frame for the evaluation between baseline and endline—nearly seven months that amount to no more than two-thirds of a typical school year in Malawi—may result in less nuance and variation in the data across timepoints. Additionally, the high reliance on self-reports and on stakeholders' viewpoints carries an inherent risk of bias.

FINDINGS AND CONCLUSIONS

One of the two main indicators calculated for the pilot evaluation focused on overall EGMA scores. In each standard, the average gain in EGMA scores for learners in treatment schools was compared with the respective gain in comparison schools from baseline to endline—which was a span of approximately 24 weeks.[7] The EGMA scores for learners in treatment schools increased by 3.154 (43% of the baseline score) in standard 1; -0.789 (-6% of the baseline score) in standard 2; 4.056 (43% of the baseline score) in standard 3; and 4.084 (52:% of the baseline score) in standard 4[8]. With the exception of standard 2, where no impact was detected, this represents an average[9] of 3.765 (46%) across standards 1, 3 and 4. These differences are, however, not all statistically significant at the 0.05 level. The only statistically

[6] See https://shared.rti.org/content/early-grade-mathematics-assessment-egma-instrumentkit

[7] The baseline occurred 24 January–2 February 2022, followed by the endline 15–26 August 2022. A typical Malawian school year lasts 42 weeks (three terms of 14 weeks each), but only 37 weeks of teaching usually occur due to orientation and assessment taking up the rest of the school calendar. Therefore, no more than two-thirds of the typical school calendar—24 weeks—passed between baseline and endline.

[8] Without controlling for appropriate variables: Std 1: 3.154; Std 2: -0.789; Std 3: 3.991; and Std 4: 3.642. When controlling for appropriate variables: Std 1: 2.462; Std 2: -1.112; Std 3: 4.056; and Std 4: 4.084.

[9] As much as this average is for all the three standards, it should be noted that the version of the EGMA used Standard 1was different to the version used with the Standard 3s and 4s.

significant difference[10] were found in standard 3 (both with and without controlling for appropriate variables[11]) and in standard 4 (when controlling for appropriate variables) (Table 1). The average gain for standard 3 learners in treatment schools was nearly four points greater than the average gain for their peers in comparison schools. In other words, the gains that standard 3 learners in treatment schools achieved would have taken more than 42 percent longer for their counterparts in control schools to attain.[12] In standard 4, the gains that learners realized in treatment schools would have taken more than 47 percent longer for their peers in control schools to reach.[13] Although the gain of roughly three points in standard 1 is not statistically significant, it would nonetheless also have taken more than 47 percent longer for the peers in control schools to reach[14].

Achievement targets for the NNP were developed in terms of Cohen's D[15]. Based on the short duration of the intervention (24 weeks) the NNP set a target of a mild to moderate impact in at least two standards. The target was achieved in three standards: standards 1, 3 and 4.

Table 1: Difference-in-difference results for overall EGMA mean scores at baseline and endline by standard

Standard	Baseline		Endline		DID (no covariates)		Covariates	
	Comparison	Treatment	Comparison	Treatment	Coefficient	p-value	Coefficient	p-value
Standard 1	6.91	8.25	14.14	18.63	3.154	0.132	2.462	0.246
Standard 2	19.29	19.24	31.01	30.17	-0.789	0.735	-1.12	0.592
Standard 3	28.92	27.45	38.44	40.96	3.991	0.017**	4.056	0.025**
Standard 4	43.09	40.63	50.81	52.00	3.642	0.074*	4.084	0.04**

Note: Two asterisks (**) denote differences between baseline and endline are statistically significant at the $p < 0.05$ level. One asterisk (*) denotes differences between baseline and endline are statistically significant at the $p < 0.1$ level.[16]

[10] Results that are statistically significant at the $p < 0.05$ level are referred to as "statistically significantly" lower or higher in the text.

[11] The study of the impact of this program is based on a quasi-experimental research design. As such, systematic differences between the comparison and treatments groups need to be controlled for to have a better measure of the real impact of the program on student's performance on EGMA.

[12] The overall average EGMA score for standard 3 learners in treatment schools increased by 13.51 points from baseline to endline, or 0.56 points per week, considering that approximately 24 weeks of instruction took place between baseline and endline. The overall average EGMA score for counterparts in control increased by 9.52 points over that span, or 0.40 points per week. Therefore, learners in control schools would need about 34 weeks to achieve the same gain that learners in treatment schools—more than 40 percent longer.

[13] The overall average EGMA score for standard 4 learners in treatment schools increased by 11.37 points from baseline to endline, or 0.47 points per week, considering that approximately 24 weeks of instruction took place between baseline and endline. The overall average EGMA score for counterparts in control increased by 7.72 points over that span, or 0.32 points per week. Therefore, learners in control schools would need more than 35 weeks to achieve the same gain that learners in treatment schools—more than 47 percent longer.

[14] The overall average EGMA score for standard 1 learners in treatment schools increased by 10.38 points from baseline to endline, or 0.43 points per week, considering that approximately 24 weeks of instruction took place between baseline and endline. The overall average EGMA score for counterparts in control increased by 7.23 points over that span, or 0.30 points per week. Therefore, learners in control schools would need more than 35 weeks to achieve the same gain that learners in treatment schools—more than 47 percent longer

[15] Cohen's D is often used to determine the size of the effect observed during an experiment. Typically, the values of Cohen's D are categorized as small (D = 0.2), moderate (D= 0.5) or large (D= 0.8 or above). The sample size used in this independent evaluation only allows the detection of an effect that is 0,35 SD or greater (i.e. between a small and moderate treatment effect).

[16] Due to this evaluation being a pilot, differences at the $p < 0.1$ level are denoted.

Analysis of the teacher and learner demographic questionnaires and classroom observation data also found that:

- **Teachers and learners at treatment schools similarly reported more positive views about their engagement with NNP materials.** A greater proportion of both teachers and learners at endline than baseline reported that workbooks were enjoyable for learners to use, as well as easy to use. This finding may be related to several factors, including that classes have covered more material at endline and/or that teachers' and learners' increased familiarity has made them more confident in their engagement with materials.
- **Some learners continue to have difficulty understanding English in their workbooks.** Learners' self-reported issues with understanding the language in the workbook at baseline persisted at endline.
- **Nearly all teachers at treatment schools said they believe they are teaching mathematics differently due to NNP.** Only 0.7 percent of teachers at endline disagreed or strongly disagreed with the statement, 'The NNP has changed my approach to teaching mathematics'.
- **The vast majority of teachers in treatment schools are delivering lessons according to the NNP structure.** According to classroom observation data, 87.5 percent of observed teachers in treatment schools at endline implemented the NNP methodology with fidelity.
- **Teachers in treatment schools displayed statistically significant better quality of instruction at endline than their counterparts in control schools in three of five categories.** According to classroom observation data, teachers in treatment schools demonstrated better instruction than teachers in comparison schools with regard to the extent to which they discussed mathematical methods and procedures and explained why they worked (methods/procedures); the extent to which they connected individual problems or examples (connections); and the extent to which they asked learners to provide responses in class (justification of learner response).
- **Some statistically significant differences in classroom practices emerged between teachers at comparison and treatment schools while they led reflection with learners.** Most notably, statistically significantly fewer teachers in treatment schools than comparison schools told learners what they expected them to notice, and statistically significantly more teachers in treatment schools than comparison ones built on learner responses by asking them to explain how their observation helped them to complete the task. These findings indicate that reflection activities may be more profound in treatment schools, with active listening and learning from others more encouraged.
- **In KIIs conducted at endline, teachers credited trainings with providing them with the knowledge they needed to implement the NNP.** Trainers corroborated teachers' positive views about the effectiveness of trainings during an earlier research period.

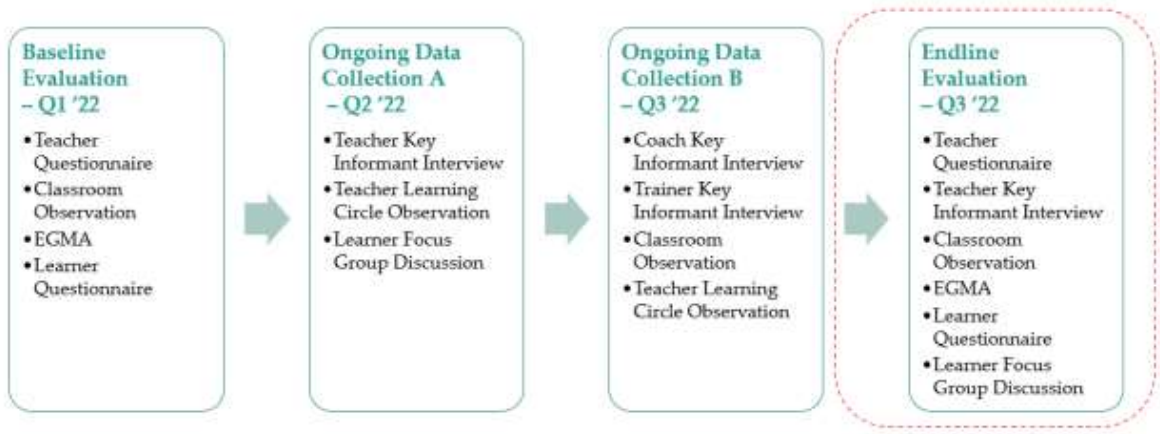
- **Overall, a greater proportion of teachers at treatment schools at endline than baseline strongly agreed with positive statements about the NNP teacher guide such as it is 'easy to use'.** For example, nearly half of the teachers at endline—47.8 percent—strongly agreed that the teacher guide 'provides sufficient guidance' on implementing NNP lessons, a statistically significant increase from 28.4 percent at baseline.
- **Questions remain for teachers at treatment schools about NNP implementation, though they reported having fewer at endline than baseline.**
- **Most teachers at treatment schools said they felt prepared to implement the NNP.** At endline, 66.9 percent of teachers said they felt 'very well prepared' to implement the new program, which was a statistically significant increase from the proportion of teacher who reported so at baseline (49.6 percent).
- **While nearly all teachers viewed the training videos as an asset, many teachers explained how the videos were not representative of their classrooms, with the videos featuring small classes with many high-performing learners.**

EVALUATION PURPOSE

EVALUATION PURPOSE

The National Numeracy Programme (NNP) pilot evaluation aims to gather information on participants' views about the programme and assess the mathematics skills of learners in treatment schools participating in the NNP pilot compared with those in comparison schools. Findings will inform the strengthening of the NNP prior to scaling it nationwide, especially in terms of the NNP's materials and teacher training components. School-to-School International (STS), in partnership with its Malawi-based partner, the Centre for Educational Research and Training (CERT), is conducting an independent evaluation of the NNP pilot at four different time points—baseline, ongoing data collection periods, and endline (Figure 2).[17]

Figure 2: Pilot evaluation timeline and tools[18]



The purpose of the pilot evaluation is to:

- Provide insight into the efficacy of the materials developed by the project—teacher guides and learner materials, including workbooks—in contributing to the achievement of improved numeracy outcomes in the lower primary phase.
- Provide insight into the efficacy of the project's in-service teacher training methodology in supporting teachers to implement the project's approach with fidelity.
- Provide insight into the impact the project has on improving learners' learning in mathematics.
- Gather participants' views to ensure further refinements to the program.

[17] A strategy to ensure that School-to-School can plan work, report findings and make recommendations independently, without any perceived or actual influence by the managing partner Cambridge Education is described in the draft Project Governance document (July 2020).

[18] The classroom observation tool was revised for ongoing data collection B and endline.

- Determine how best to scale to a national level in a cost-effective, sustainable manner that ensures both the maximum and sustained impact of the project.

PILOT EVALUATION QUESTIONS

Nine questions guide the pilot evaluation, as detailed in Table 2. Cambridge Education and STS developed the evaluation questions and evaluation methodologies for each question based on numerous discussions.

Table 2: Pilot evaluation questions and learning areas

Evaluation question	Domains[19] (Pilot evaluation instruments are listed in Table 3)
1. Have pilot activities improved learner performance? a. *Under what conditions have pilot activities improved learner engagement and performance?*	• Learner mathematics learning outcomes
2. Have the pilot activities changed learner engagement?[20]	• Teacher perceptions of materials/content quality • Learner experience with materials and activities
3. How are learners engaging with and using the workbooks independently? a. *Are there language issues that impede on learner engagement with the materials?* b. *Is learner engagement with the learning materials equitable with regard to gender and learners with learning difficulties?*	
4. To what extent have the pilot intervention's activities changed classroom practices?[21]	• Teacher perceptions of materials/content quality • Ease of adoption • Classroom routine/sequence
5. In what ways are face-to-face teacher trainings changing teacher practices in the classroom? a. *Under what conditions are teacher trainings being implemented as intended?*	• Stakeholder perceptions of training quality • Teacher knowledge • Classroom routine/sequence
6. Are the pilot coaching sessions and Teacher Learning Circles (TLC) working as intended? a. *Under what conditions are the coaching sessions effective?*	• Coaches' perceptions of effectiveness • Teacher perceptions of coaching session and TLC quality
7. In what ways are the teacher guides supporting teachers to understand the methodology/approach being implemented?	

[19] Domains refers to what the evaluation instruments capture.
[20] In comparison with the observed learner behaviours in the pre-study classroom observations, learners are (a) doing more "independent" work; and (b) working in the workbooks. Comparison is with scoping study produced prior to start of project.
[21] Intervention activities include teacher guides; learner materials; face-to-face training; and school based CPD, such as TLCs & coaching.

Evaluation question	Domains[19] (Pilot evaluation instruments are listed in Table 3)
8. In what ways do the teacher guides support teachers in implementing the intended methodology/approach and using learner materials effectively and with fidelity?	• Teacher perceptions of materials/content quality • Ease of adoption • Classroom routine/sequence
9. How are the training videos being used? a. *Are the training videos perceived as a useful training resource?* b. *How could they be improved and made more useful?*	• Teacher perceptions of training videos quality and usefulness • Coaches' perceptions

PROJECT BACKGROUND

Led by the Malawi Ministry of Education (MoE) and funded by the United Kingdom's Foreign, Commonwealth & Development Office (FCDO), the four-year NNP aims to improve outcomes in mathematics, so girls and boys have a solid foundation in basic skills to succeed in the rest of their schooling and fulfil their potential.

The programme responds to key findings from a scoping exercise commissioned by FCDO and MoE that investigated what factors are hindering outcomes in mathematics for learners in lower primary (standards 1–4). The research identified two overarching themes about the mathematics landscape in primary schools in Malawi—'limited and limiting expectations of learners; and the focus of teaching…on form over substance'.[22] Building on the scoping study's findings and recommendations, the NNP aims to develop a new vision for teaching and learning mathematics in Malawi in which children experience mathematics as a meaningful, sense-making, and problem-solving activity. Learners will be expected not only to know mathematics but also to understand the mathematics they know, apply the mathematics to solve unfamiliar problems, and reason and argue using the mathematics that they develop.

The NNP's key objectives are to:

- Revise the mathematics curriculum for lower primary.
- Develop teaching and learning materials aligned to the revised curriculum.
- Create a sustainable system to train teachers and other school officials, including school-based support structures.
- Institutionalise the new mathematics curriculum and training approach in MoE systems.
- Rigorously pilot the new materials and training strategies.
- Measure the impact of the pilot and refine the materials and training strategies.
- Oversee the national scale-up.

The NNP is being implemented in three phases: inception, pilot, and scale-up. In the initial inception phase, the has developed a vision for mathematics teaching and learning in Malawi, a plan for the revision of the curriculum, new teaching and learning materials for standards 1–4 for piloting, and a training methodology for piloting.

The pilot phase, which took place during the 2022 school year, included the following activities:

[22] Brombacher, Aarnout. 'Research to Investigate Low Learning Achievement in Early Grade Numeracy (Standards 1–4) in Malawi: The Victory of Form Over Substance.' Oxford, U.K.: HEART, 2019.

- Rigorous piloting of the new materials and teacher training approach in 200 schools across all divisions of Malawi.
- Development of a scale-up mechanism.
- Refinement of materials based on the learnings of the pilot activity.

The project will culminate with the nationwide scale-up phase. The project will produce and distribute materials, provide in-service training, and facilitate ongoing school- and classroom-based support for all standard 1–4 teachers.

EVALUATION METHODS AND LIMITATIONS

STUDY DESIGN

STS is evaluating the effectiveness of the NNP pilot phase by employing a mixed-methods, pre-post approach using a quasi-experimental design. A mixed-methods design involves selecting a number of schools to participate in the quantitative portion of the evaluation and selecting a subsample of schools to participate in the qualitative portion of it. The pilot evaluation includes a baseline, two ongoing collection periods, and an endline.

The evaluation team is comprised of U.S. and Malawi-based experts working together to design, conduct, analyse, and report on pilot phase learnings. STS provides technical leadership and oversight of all components of the pilot evaluation. STS also provides quality control over data collection. STS's Malawian counterpart, CERT, manages all in-country logistics for baseline assessor training and data collection.

TOOLS

The pilot evaluation employs a diverse set of instruments targeted at different stakeholders and participants. These instruments include learning and knowledge assessments, interviews, classroom observations, and questionnaires. The instruments were designed to respond to the evaluation questions and were all finalized after a pretest was conducted in November 2021 except for the classroom observation form, which was updated after baseline. Due to the mixed-methods nature of this evaluation, instruments are used for quantitative and qualitative types of inquiry.

Quantitative Inquiry – Deductive Approach

- **The teacher questionnaire** captures demographic information about teachers, their teaching experience, and their experience with and perceptions about the project.
- **The learner demographic questionnaire** captures background and demographic information about learners.
- **The Early Grade Mathematics Assessment (EGMA)** captures learners' knowledge of numeracy skills and is largely based on the version developed in 2010 by the United States Agency for International Development (USAID) and the MoE.[23] It includes two versions—one for standards 1–2 and the other for standards 3–4. The specific tasks are detailed in Table 3.
- **The classroom observation form** primarily measures teachers' quality of instruction, as well as the extent to which they were delivering mathematics lessons with fidelity. Prior to ongoing data collection B, the original classroom observation form was revised to better understand the proportion of teachers implementing the NNP

[23] USAID/Malawi and MoEST. USAID Funded Malawi Teacher Professional Development Support (MTPDS) Activity 2010 Early Grade Mathematics Assessment (EGMA): National Baseline Report 2010.Washington, DC: USAID, 2010.

methodology with fidelity as well as providing high quality of instruction in mathematics. This revised form was pretested at ongoing data collection B and then used at endline.

Table 3: EGMA subtasks

Subtask	Number of items	Skill	Description Learner was asked to …	Differences between Administrations by Standards	
				EGMA for Standards 1-2	EGMA for Standards 3-4
Subtasks that assess more procedural (recall) type of knowledge					
Number Identification	10	This task requires knowledge of the number symbols.	… select a given number from three different numbers provided. (*Untimed subtask*)	Items include numbers ranging from 7 to 456.	Items include numbers ranging from 7 to 1,200.
Addition and Subtraction (Level 1 [basic facts])	20 per subtask	This subtask requires knowledge of and confidence with basic addition and subtraction facts. It is expected that learners should develop some level of automaticity and fluency with these facts because they need them throughout mathematics.	… mentally solve addition and subtraction problems, with sums and differences below 20. The problems ranged from those with only single digits to problems that involved the bridging of the 10. (*Timed subtask*[24])	No difference between items used for Standards 1-2 and 3-4.	
Subtasks that assess more conceptual (application) type of knowledge					

[24] Learners in standards 1 and 2 had 2 minutes to complete the 20 subtask items, while learners in standards 3 and 4 had 1 minute.

Subtask	Number of items	Skill	Description Learner was asked to …	Differences between Administrations by Standards	
				EGMA for Standards 1-2	EGMA for Standards 3-4
Quantity Discrimination (number comparison)	10	This subtask requires the ability to make judgments about differences by comparing quantities represented by numbers.	… identify the larger of a pair of numbers. For standards 1 and 2 learners, the number pairs included two pairs of single-digit numbers and eight pairs of double-digit numbers. For standard 3–4 learners, the number pairs included two pairs of single-digit numbers, three pairs of double-digit numbers, three pairs of three-digit numbers, and two pairs of four-digit numbers. (*Untimed subtask*)	Items include numbers ranging from 3 to 91.	Items include numbers ranging from 3 to 5,002.
Pattern Completion (number and shape patterns)	5	This subtask requires the ability to discern and complete number and shape patterns.	… determine the missing number or shape in a pattern of four numbers, one of which is missing, or four or more shapes, one of which is missing. Patterns used included counting forward by ones, twos, and fives and identifying sequences with triangles, circles, and/or diamonds. (*Untimed subtask*)	Items include patterns counting forward by ones and twos and identifying sequences with triangles, circles, and diamonds.	Items include patterns counting forward by twos and fives and identifying sequences with triangles and lines.
Addition and Subtraction (Level 2)[25]— only for standard 3 and 4 learners	5 per subtask	This subtask requires the ability to use and apply the procedural addition and subtraction knowledge assessed in the Level 1 subtask to solve more complicated addition and subtraction problems.	… solve addition and subtraction problems that involve the knowledge and application of the basic addition and subtraction facts assessed in the Level 1 subtask. The problems extended to the addition and subtraction of two-digit and three-digit numbers involving bridging. (*Untimed subtask*).	Not administered to Standards 1-2	

[25] The addition and subtraction (level 2) subtasks are more conceptual than the addition and subtraction (level 1) subtasks because a learner must understand what he or she is doing when applying the level 1 skills. Although the level 2 subtasks are

Subtask	Number of items	Skill	Description Learner was asked to …	Differences between Administrations by Standards	
				EGMA for Standards 1-2	EGMA for Standards 3-4
Problems	5	This subtask requires the ability to interpret a situation (presented orally to the learner), make 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. Because the focus of this subtask was on assessing the learners' abilities to interpret a situation, make a plan, and solve a problem, 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. (*Untimed subtask*).	No difference between items used for Standards 1-2 and 3-4.	

Qualitative Inquiry – Inductive Approach

- **The teacher key informant interview** (KII) captures teachers' perceptions about the effectiveness of the teacher training, materials, coaching, and Teacher Learning Circles (TLC), as well as their perceptions of learners' engagement. Questions are mostly open-ended.
- **The coach KII** captures coaches' perceptions about the effectiveness of the coaching and TLCs, as well as thoughts on teachers' implementation of materials and classroom activities. Questions are mostly open-ended.
- **The trainer KII** captures trainers' perceptions about the effectiveness of the teacher training, coaching, and TLCs. Questions are mostly open-ended.

not purely conceptual, because, with time, learners will develop some automaticity with the items in these subtasks, they are more conceptual than the level 1 subtasks, especially so for standard 2 learners.

.

- **The TLC observation form** captures information on the implementation of TLCs and the extent to which section heads are implementing the TLC activities with fidelity. Questions are open and closed-ended.
- **The learner focus group discussion** (FGD) captures learners' perceptions of and experience with the classroom materials.

All quantitative and qualitative tools were administered at endline except for the coach and trainer KIIs and the TLC observation form. The time points for the administration of the tools are detailed in Table 4.

Table 4: Timepoints of tools' administration

Instrument	Respondent group	Evaluation question(s)	Time point			
			Baseline	Ongoing Data Collection A	Ongoing Data Collection B	Endline
Teacher questionnaire*	Teachers (St 1-4)	1, 2, 3, 4, 5, 8	X			X
Teacher key informant interview	Teachers (St 1-4)	1, 2, 3, 4, 5, 6, 7, 8		X		X
Coach key informant interview	Coaches (Head teachers and section heads)	7, 8			X	
Trainer key informant interview	DEMs PEAs Others	6, 7, 8			X	
Classroom observation form*	N/A	1, 2, 3, 4, 5, 7	X			
Revised classroom observation form** - fidelity of implementation and quality of mathematics instruction					X	X
Teacher Learning Circle observation form	Teachers (St 1-4) and section heads			X	X	
Early Grade Mathematics Assessment (EGMA)*	Learners (St 1-4)	9	X			X

Instrument	Respondent group	Evaluation question(s)	Time point			
			Baseline	Ongoing Data Collection A	Ongoing Data Collection B	Endline
Learner demographic questionnaire*	Learners (St 1-4)	4,5	X			X
Learner focus group discussion	Learners (St 3-4)	4,5		X		X

Note: Tools with an asterisk (*) were administered in both comparison and treatment groups. Tools without an asterisk were administered in only treatment schools.

SAMPLING

For the pilot evaluation, treatment and comparison schools were sampled using a three-stage clustering random method. For the first stage, schools were randomly selected using a probability proportionate to size (PPS) without replacement approach. At this stage, schools were stratified by group—treatment and comparison—and a PPS approach was used to select 40 schools in each stratum randomly. Schools from both groups came from zones participating in the pilot. EMIS data was used to determine the measure of size (MOS) of each school in the population. In addition, 12 schools from each group were identified as replacement schools.

For the second stage, when data collection teams visited selected schools, they randomly selected one classroom per standard using a simple random approach.

Lastly, for the third stage, learners in each selected classroom were stratified by gender. Data collection teams then used a simple random approach to select a sample of five boys and girls to participate in EGMA and learner questionnaire. Eighty schools were initially part of the baseline sample, but disruptions brought about by Cyclone Ana during baseline data collection resulted in only 75 schools being visited.

TRAINING AND DATA COLLECTION

Principal Researcher and Research Assistant Training

The endline data collection training utilised a cascade model. STS conducted remote training with the principal researchers for three days from 3-5 August 2022. The principal researchers have over a decade of experience in conducting quantitative and qualitative research, including similar mathematics studies.

Following the remote training, the principal researchers conducted a five-day training with the research assistants in-person in Lilongwe from 8–12 August 2022. The principal researchers utilised training materials provided by STS, which helped ensure the quality of

the training. The NNP Technical Lead facilitated training of the classroom observation tool. After leading each day of training, the principal researchers met with STS to discuss which topics were covered and resolve any outstanding questions. Research assistants were selected based on their previous experience in conducting quantitative research, including those from the Ministry of Education and retired Primary Education Advisors (PEAs).

Both training phases reintroduced the six endline instruments—the EGMA, learner demographic questionnaire, teacher questionnaire, classroom observation form, teacher KII, and learner FGD—and addressed administration and scoring. In addition to these data collection protocols, training for research assistants also included sessions on human protection policies, ethics, COVID safety, ethical handling and storage of data, and the project's 'Do No Harm' policy, as well as one day of in-school practice with the instruments. Principal researchers also examined research assistants' scoring accuracy for each instrument and interrater reliability (IRR) during training.

The principal researchers and some of the research assistants were already familiar with the instruments because they participated in baseline data collection in January and February 2022.

IRR Procedure and Scores

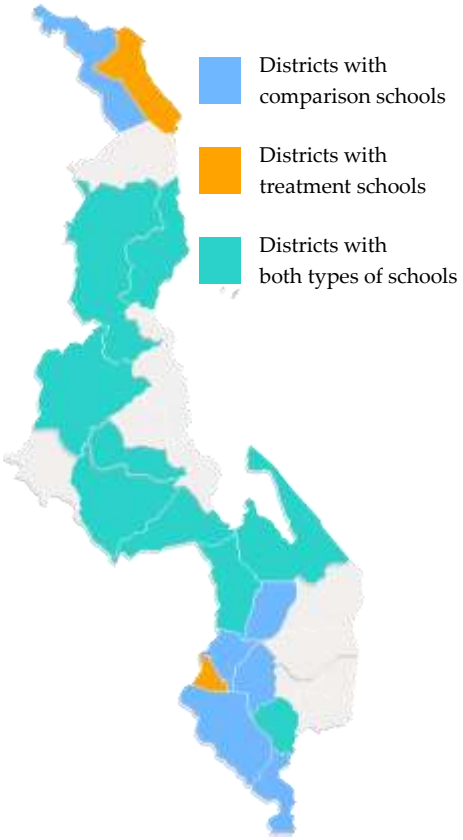
IRR measures were conducted during the research assistant training. Only one measure was conducted during the endline training. The research assistants engaged at endline were also engaged at baseline, received baseline and endline training, completed previous IRR measures, and conducted EGMAs in the field at baseline. Therefore, these enumerators had extensive experience with this EGMA.

During the IRR measure during the endline research assistant training, the principal researchers simulated both the individual and group administration of the EGMA—one principal researcher played the role of a learner while the other played the role of an assessor. The research assistants observed the simulation and marked the scores in their tablets. The data was uploaded and analysed. All research assistants scored above the 90.0 percent-agreement threshold during training. As such, no additional IRR measures were conducted.

Data Collection

Visiting one school per day, eight teams of four—one supervisor and three research assistants—conducted the endline study from 15–26 August 2022. The supervisor handled logistics; introduced the team to the head teacher at the start of each school visit; ensured spaces for conducting assessments and surveys were available and appropriate; provided oversight of assessments and surveys; administered surveys and assessments as needed; and reported their progress to CERT. Two research assistants from each team administered the EGMA and learner questionnaires, and the third research assistant completed the classroom observations and teacher questionnaires. The three principal researchers, who led the training, supervised the eight data collection teams and met with STS on a regular basis to report on the progress of data collection.

Figure 3: Map of baseline study sample



Each team visited nine or 10 schools over two weeks, with teams visiting all 75 schools assessed at endline. Eighty schools were initially part of the baseline sample, but disruptions brought about by Cyclone Ana during baseline data collection resulted in only 75 schools being visited.

At endline, the teams visited 40 comparison schools and 35 treatment schools from 17 districts across Malawi (Figure 3). A total of 296 teachers and 1,489 learners from standards 1–4—an average of 5–6 learners per standard per school—participated in the endline study (Figure 4).

Figure 4: Baseline sample

75 Schools	4 Standards per School	5-6 Learners per Standard	1 Teacher per Standard
40 comparison schools 35 treatment schools	EGMAs differed for learners in Standards 1–2 & 3–4	1,489 learners total	296 teachers total

Data Cleaning

Prior to conducting analysis, STS cleaned data based on certain criteria. To be considered 'clean', data had to be (1) complete, (2) accurate, and (3) internally consistent. STS used

multi-stage data cleaning plans to ensure all data values were within the allowable range and that reserve codes were used appropriately.

ANALYTIC METHODS

The pilot study's evaluation questions guided the analysis. EGMA and questionnaire data were coded and analysed in Stata following best practices outlined in the EGMA toolkit guidance.[26] All items or questions were analysed individually, with means, standard deviations, and frequencies produced for each variable. In addition, data were aggregated, as needed, to respond to each evaluation question.

Statistical Weighting

To correct for the unequal probability of selection due to clustering of the sample, survey weights were computed with a two-step procedure and included in all analyses. In the first step, base weights were computed for each dataset. In the second step, adjustment factors were applied to correct for the non-participation of the selected learners as well as a selection within the school.

The probability of inclusion of each learner in strata s is:

$$\pi^s = \left[\frac{m_i^s * n^s}{M^s}\right] * \left[\frac{s^s}{m_i^s}\right] = \frac{s^s * n^s}{M^s}$$

Where

- g_l^{ab} is the total enrolment of school i in strata s
- $M^s$is the total enrolment of all schools in strata s
- n^s is the total number of schools sampled in strata s
- s^s is the number of learners sampled per school in strata s

Thus, the school weight—or the inverse probability of selection—for strata s is:

$$W_s = {}^1/_{\pi^s} = \frac{M^s}{s^s * n^s}$$

To calculate the learner weight for the second stage, the probability of selecting a standard 2 learner of gender g = {male, female} at a school i is:

$$\pi^{g,i} = \frac{s^{g,i}}{M^{g,i}}$$

Where

- $s^{g,i}$ is the number of learners of gender g sampled from standard 2 of school i
- $M^{g,i}$ is the total number of learners of gender g in school i

[26] See https://shared.rti.org/content/early-grade-mathematics-assessment-egma-instrumentkit

Thus, the adjustment factor (inverse probability of selection) for learners is:

$$A_{g,i} = {}^{1}/_{\pi^{g,i}} = \frac{M^{g,i}}{S^{g,i}}$$

Adjustment factors are multiplied by the respective school weight when weighting each observation.

Characteristics of Assessment Tool

The Cronbach alpha—an estimate of reliability of a subtask's scores—was calculated for each EGMA subtask to assess its psychometric qualities. Cronbach alpha scores were computed separately for subtasks on the standard 1–2 EGMA and those on the standard 3–4 EGMA.

The Cronbach alpha estimates for subtasks on the standard 1–2 EGMA ranged from 0.56 for pattern recognition to 0.90 for subtraction level 1. On the standard 3–4 EGMA, the estimates ranged from 0.48 for pattern recognition to 0.92 for subtraction level 1.

Similar to baseline, the lower estimates for some of the tasks are most likely a reflection of learners' lack of familiarity with the mathematics that the task is assessing. Pattern recognition and problems are topics that traditionally do not get much attention in mathematics in the early years in Malawi. As such, many learners were unable to answer any question for those subtasks correctly. Rasch analysis of the data in terms of the person separation index for these subtasks confirms a low person variance thereby confirming the explanation for the lower alphas on these tasks.

A complete list of Cronbach alpha values can be found in Annex III.

Generation of Findings

For overall EGMA scores, findings were generated by using a difference-in-difference analysis approach. Appropriate covariates were also used in the analysis to control for the heterogeneity between the two groups. For individual EGMA subtasks, descriptive analysis for each standard and both groups was done using percent correct and zero scores. For all other tools, proportions were calculated for categorical variables and means computed for continuous variables. Specific descriptive results have also been disaggregated by standard and school status (treatment or comparison) depending on the evaluation questions. To study the relationships between specific indicators and contextual factors, Pearson correlations were calculated, and linear regressions were conducted. For all analyses, survey design and weights were considered.

LIMITATIONS

- The short time frame for the evaluation may result in less nuance and variation in the data between baseline and endline—nearly seven months that amount to no more than two-thirds of a typical school year in Malawi.[27]
- The modification of the classroom observation tool limits the analysis of the change in teacher practices over time.
- Findings on EGMA need to be interpreted with caution due to the method of administration. First, because the tool was administered to learners in a group-setting, it is difficult to interpret accuracy scores on timed subtasks. Second, there are concerns regarding group administration with younger learners in standard 1 and 2 due to challenges with understanding task instruction and producing a written output.
- The high reliance on self-reports and on stakeholders' viewpoints carries an inherent risk of bias. However, in the case of the teacher survey, the anonymous nature of the questionnaire reduces the risks of social desirability. In addition, using multiple instruments and gathering information from multiple stakeholders on similar issues enables the researchers to identify areas of consistency or lack thereof. The level of consistency between different responses and stakeholders provides a degree of confidence in the strength of the findings.

[27] The baseline occurred 24 January–2 February 2022, followed by the endline 15–26 August 2022. A typical Malawian school year lasts 42 weeks (three terms of 14 weeks each), but only 37 weeks of teaching usually occur due to orientation and assessment taking up the rest of the school calendar. Therefore, when accounting for vacation between school terms, no more than two-thirds of the school calendar passed between baseline and endline.

STUDY FINDINGS

Research from the endline study is presented by evaluation question. Evaluation questions 7 and 8 were combined due to their similarity, and evaluation questions 5 and 6 include some data collected from Ongoing Data Collection B due to certain tools not being administered at endline, including coach and trainer KIIs and TLC observation forms. Results that are statistically significant at the $p < 0.05$ level are referred to as "statistically significantly" lower or higher in the text.

EQ 1. HAVE PILOT ACTIVITIES IMPROVED LEARNER LEARNING? EQ 1.A. UNDER WHAT CONDITIONS HAVE PILOT ACTIVITIES IMPROVED LEARNER ENGAGEMENT AND LEARNING?

This section first compares the difference in learners' overall gains from baseline to endline by subtasks in each standard in both the treatment and comparison groups, followed by a comparison of the difference in gains in overall EGMA scores by standard. For all results, using a method called difference-in-differences (DID) analysis, analysts compared changes between the outcomes of the treatment and comparison groups to determine if they were statistically significant.

EGMA Scores by Subtask and Standard

The subtask results reported in this section include several measures. First, the total percentage correct for each subtask is presented—except for addition and subtraction level 1, for which the number of correct items per minute is reported.[28] Second, the percentage of learners who did not answer any items correctly on a subtask—also known as a zero score—is detailed. A decrease from baseline to endline in the proportion of zero scores on a subtask illustrates that more learners have developed the skills to answer the types of items in a subtask. Further, an exploratory analysis was conducted to examine whether EGMA performance (by subtask) of learners was meeting the minimum proficiency levels described in the Global Proficiency Framework (GPF). Given the exploratory nature of this investigation, the findings are presented in Annex I.

In standard 1, the gains for learners at treatment schools from baseline to endline were statistically significantly greater than the gains for their counterparts at comparison schools on three subtasks—number identification, pattern completion, and problems—as displayed in Table 5. On number identification, learners in comparison and treatment schools had similar results at baseline—25.1 percent and 27.8 percent, respectively. At endline, however, learners in treatment schools answered 54.9 percent of questions correctly, while their peers in comparison schools only answered 37.8 percent correctly. Learners in treatment schools displayed similar improvements in pattern completion and problems.

[28] Results for this measure should be viewed with some caution. The addition and subtraction subtasks for this EGMA were administered in a group setting. Enumerators only considered a learner to have completed the subtask before time elapsed if a learner had written down an answer for all items on the subtask. Therefore, it was not possible to discern if learners did not answer a question either because a) they attempted it, but provided no response (i.e., left it blank) because they did not know the answer, or b) they ran out of time on the subtask and did not have a chance to attempt the item.

These statistically significant gains for standard 1 learners in treatment schools suggest that they are developing mathematics abilities in a variety of areas, including foundational skills (represented by gains in number identification), conceptual knowledge (represented by gains in pattern completion), and problem solving (represented by gains in word problems).

Table 5. Standard 1 EGMA scores by subtasks

Subtask	Group	% correct / total		
		Baseline	Endline	p-value for DiD (no covariates)
Number Identification**	Comparison	25.1%	37.8%	0.008
	Treatment	27.8%	54.9%	
Quantity Discrimination	Comparison	25.6%	36.3%	0.338
	Treatment	31.6%	47.5%	
Pattern Completion**	Comparison	4.0%	11.0%	0.027
	Treatment	6.4%	22.4%	
Problems**	Comparison	7.0%	20.8%	0.019
	Treatment	7.2%	28.8%	
Subtask	**Group**	**Number of correct items per minute**		
		Baseline	**Endline**	**p-value for DiD (no covariates)**
Addition (Level 1)	Comparison	0.34	1.44	0.859
	Treatment	0.53	1.67	
Subtraction (Level 1)	Comparison	0.29	1.18	0.684
	Treatment	0.32	1.34	

Note: Two asterisks (**) indicates that the difference between comparison and comparison schools is statistically significant at $p < 0.05$.

As for zero scores in standard 1, learners in treatment schools had statistically significant lower proportions of zero scores on two subtasks—pattern completion and addition (level 1)—as detailed in Table 6. While 64.1 percent of learners in comparison schools at endline were not able to answer any questions correctly on the pattern completion subtasks, only 42.9 percent in treatment schools could not do so. On the addition (level 1) subtask, 38.6 percent of learners in comparison schools had a zero score, while only 24.2 percent did in treatment schools.

Table 6: Differences in percentages of EGMA zero-scores between comparison and treatment schools at endline, standard 1

Subtask	Comparison		Treatment		Difference
	N	% Zero-Scores	N	% Zero-Scores	(treatment - comparison) p-value
Number Identification	199	9.6%	175	6.9%	0.47
Quantity Discrimination	199	16.8%	175	15.1%	0.70
Pattern Completion**	199	64.1%	175	42.9%	0.01
Addition (Level 1)**	199	38.6%	175	24.2%	0.03
Subtraction (Level 1)	199	44.7%	175	44.7%	0.99
Problems*	199	43.2%	175	33.5%	0.08

Note: Two asterisks (**) denote differences between baseline and endline are statistically significant at the $p < 0.05$ level. One asterisk (*) denotes differences between baseline and endline are statistically significant at the $p < 0.1$ level.[29]

[29] Due to this evaluation being a pilot, differences at the $p < 0.1$ level are denoted.

In contrast with standard 1 learners, no statistically significant differences emerged in the gains from baseline to endline between standard 2 learners in comparison and treatment schools, as shown in Table 7. For example, on the pattern completion subtask, the proportion of questions that learners in comparison and treatment schools answered correctly was similar at baseline— 17.6 percent, and 19.6, respectively—and at endline—35.1 percent and 38.5 percent, respectively.

Table 7. Standard 2 EGMA scores by subtasks

Subtask	Group	% correct / total		
		Baseline	Endline	p-value for DiD (no covariates)
Number Identification	Comparison	52.8%	71.3%	0.152
	Treatment	49.7%	74.4%	
Quantity Discrimination	Comparison	47.6%	68.8%	0.705
	Treatment	41.1%	64.5%	
Pattern Completion	Comparison	17.6%	35.1%	0.701
	Treatment	19.6%	38.5%	
Problems	Comparison	23.2%	35.8%	0.482
	Treatment	27.2%	43.8%	
Subtask	**Group**	**Number of correct items per minute**		
		Baseline	**Endline**	**p-value for DiD (no covariates)**
Addition (Level 1)	Comparison	2.13	3.56	0.191
	Treatment	2.24	3.21	
Subtraction (Level 1)	Comparison	1.60	3.17	0.185
	Treatment	1.89	2.87	

As with the subtask results for standard 2 learners, no statistically significant differences emerged at endline in the proportion of zero scores between learners in comparison and treatment schools, as displayed in Table 8. For instance, the proportion of zero scores on subtraction (level 1) was similar for learners in comparison schools (19.0 percent) and treatment schools (18.3 percent).

Table 8: Differences in percentages of EGMA zero-scores between comparison and treatment schools at endline, standard 2

Subtask	Comparison		Treatment		Difference
	N	% Zero-Scores	N	% Zero-Scores	(treatment - comparison) p-value
Number Identification	200	2.6%	175	1.2%	0.47
Quantity Discrimination	200	5.2%	175	6.6%	0.61
Pattern Completion	200	15.3%	175	17.8%	0.65
Addition (Level 1)	200	6.7%	175	5.8%	0.78
Subtraction (Level 1)	200	19.0%	175	18.3%	0.88
Problems	200	14.7%	175	16.8%	0.67

In standard 3, the gains for learners at treatment schools from baseline to endline were statistically significantly greater than the gains for their counterparts at comparison schools on one subtask—addition (level 2)—as shown in Table 9. Learners in comparison and

treatment schools had similar results at baseline on addition (level 2)—23.2 percent and 25.0 percent, respectively. At endline, however, learners in treatment schools answered 46.5 percent correctly, while their peers in comparison schools only answered 34.2 percent correctly.

Table 9. Standard 3 EGMA scores by subtasks

Subtask	Group	% correct / total		
		Baseline	Endline	p-value for DiD (no covariates)
Number Identification	Comparison	60.8%	75.6%	0.220
	Treatment	57.5%	77.8%	
Quantity Discrimination	Comparison	55.4%	67.9%	0.312
	Treatment	53.6%	71.0%	
Pattern Completion	Comparison	12.0%	25.1%	0.107
	Treatment	19.0%	38.1%	
Addition (Level 2)**	Comparison	23.2%	34.2%	0.023
	Treatment	25.0%	46.5%	
Subtraction (Level 2)	Comparison	18.4%	27.3%	0.240
	Treatment	20.4%	34.3%	
Word problems	Comparison	37.2%	54.7%	0.583
	Treatment	37.2%	57.3%	
Subtask	**Group**	**Number of correct items per minute**		
		Baseline	**Endline**	**p-value for DiD (no covariates)**
Addition (Level 1)	Comparison	3.72	4.91	0.259
	Treatment	3.37	5.01	
Subtraction (Level 1)*	Comparison	2.86	3.63	0.095
	Treatment	2.27	3.64	

Note: Two asterisks (**) denote differences between baseline and endline are statistically significant at the $p < 0.05$ level. One asterisk (*) denotes differences between baseline and endline are statistically significant at the $p < 0.1$ level.[30]

As for zero scores in standard 3, learners in treatment schools had statistically significant lower proportions of zero scores on one subtask—pattern completion—as detailed in Table 10. While 29.8 percent of learners in comparison schools at endline were not able to answer any questions correctly on the pattern completion subtask, only 16.4 percent in treatment schools could not do so.

Table 10: Difference in percentages of EGMA zero-scores between comparison and treatment schools at endline, standard 3

Subtask	Comparison		Treatment		Difference
	N	% Zero-Scores	N	% Zero-Scores	(treatment - comparison) p-value
Number Identification	200	0.0%	170	0.0%	n/a
Quantity Discrimination	200	5.2%	170	2.5%	0.261
Pattern Completion**	200	29.8%	170	16.4%	0.006
Addition (Level 1)	200	1.9%	170	2.7%	0.635
Subtraction (Level 1)	200	15.8%	170	11.9%	0.412
Addition (Level 2)	200	1.9%	170	2.7%	0.635

[30] Due to this evaluation being a pilot, differences at the $p < 0.1$ level are denoted.

Subtask	Comparison		Treatment		Difference
	N	% Zero-Scores	N	% Zero-Scores	(treatment - comparison) p-value
Subtraction (Level 2)	200	15.8%	170	11.9%	0.412
Problems	200	4.3%	170	5.3%	0.666

Note: Two asterisks (**) indicates that the difference between comparison and comparison schools is statistically significant at $p < 0.05$.

In standard 4, the gains from baseline to endline were statistically significantly higher for learners in treatment schools than their peers in comparison schools on one subtask—pattern completion—as shown in Table 11. While learners answered a similar proportion of items correctly at baseline—27.4 percent at comparison schools and 30.2 percent at treatment schools—learners at treatment schools at endline answered nearly half of the items correctly (49.4 percent), compared with learners at comparison schools correctly answering 36.0 percent.

Table 11. Standard 4 EGMA scores by subtasks

Subtask	Group	% correct / total		
		Baseline	Endline	p-value for DiD (no covariates)
Number Identification	Comparison	79.8%	88.0%	0.477
	Treatment	74.8%	85.4%	
Quantity Discrimination	Comparison	77.9%	80.0%	0.673
	Treatment	74.8%	78.8%	
Pattern Completion**	Comparison	27.4%	36.0%	0.026
	Treatment	30.2%	49.4%	
Addition (Level 2)	Comparison	40.6%	50.8%	0.232
	Treatment	45.4%	63.2%	
Subtraction (Level 2)	Comparison	34.6%	46.0%	0.629
	Treatment	39.4%	48.2%	
Problems	Comparison	51.2%	61.3%	0.973
	Treatment	54.4%	64.5%	
Subtask	**Group**	**Number of correct items per minute**		
		Baseline	Endline	p-value for DiD (no covariates)
Addition (Level 1)	Comparison	5.44	6.69	0.133
	Treatment	4.87	6.82	
Subtraction (Level 1)	Comparison	4.41	5.53	0.437
	Treatment	3.95	5.49	

Note: Two asterisks (**) indicates that the difference between comparison and comparison schools is statistically significant at $p < 0.05$.

As for the proportion of zero scores in standard 4, no statistically significant differences emerged, as displayed in Table 12, primarily because so few learners were unable to answer any items correctly on each subtask. For instance, the proportion of zero scores on problems was close to zero for both learners in comparison schools (1.3 percent) and treatment schools (2.2 percent).

Table 12: Difference in percentages of EGMA zero-scores between comparison and treatment schools at endline, standard 4

Subtask	Comparison		Treatment		Difference
	N	% Zero-Scores	N	% Zero-Scores	(treatment - comparison) p-value
Number Identification	200	0.00%	170	0.00%	n/a
Quantity Discrimination	200	0.13%	170	0.03%	0.392
Pattern Completion*	200	17.50%	170	9.60%	0.09
Addition (Level 1)	200	0.00%	170	0.01%	n/a
Subtraction (Level 1)	200	0.07%	170	0.04%	0.175
Addition (Level 2)	200	0.00%	170	0.01%	n/a
Subtraction (Level 2)	200	6.90%	170	3.60%	0.175
Problems	200	1.30%	170	2.20%	0.489

Note: One asterisk (*) denotes differences between baseline and endline are statistically significant at the $p < 0.1$ level.[31]

Overall EGMA Scores

One of the two main indicators calculated for the pilot evaluation focused on overall EGMA scores. In each standard, the average gain in EGMA scores for learners in treatment schools was compared with the respective gain in comparison schools from baseline to endline—which was a span of approximately 24 weeks.[32] The EGMA scores for learners in treatment schools increased by 3.154 (43% of the baseline score) in standard 1; -0.789 (-6% of the baseline score) in standard 2; 4.056 (43% of the baseline score) in standard 3; and 4.084 (52:% of the baseline score) in standard 4[33]. With the exception of standard 2, where no impact was detected, this represents an average[34] of 3.765 (46%) across standards 1, 3 and 4. These differences are, however, not all statistically significant at the 0.05 level. The only statistically significant difference[35] were found in standard 3 (both with and without controlling for appropriate variables[36]) and in standard 4 (when controlling for appropriate variables) (Table 1). The average gain for standard 3 learners in treatment schools was nearly four points greater than the average gain for their peers in comparison schools. In other words, the gains that standard 3 learners in treatment schools achieved would have taken more than 42 percent longer for their counterparts in control schools to attain.[37] In standard 4, the gains that learners realized in treatment schools would have taken more than 47 percent longer for their

[31] Due to this evaluation being a pilot, differences at the $p < 0.1$ level are denoted.

[32] The baseline occurred 24 January–2 February 2022, followed by the endline 15–26 August 2022. A typical Malawian school year lasts 42 weeks (three terms of 14 weeks each), but only 37 weeks of teaching usually occur due to orientation and assessment taking up the rest of the school calendar. Therefore, no more than two-thirds of the typical school calendar—24 weeks—passed between baseline and endline.

[33] Without controlling for appropriate variables: Std 1: 3.154; Std 2: -0.789; Std 3: 3.991; and Std 4: 3.642. When controlling for appropriate variables: Std 1: 2.462; Std 2: -1.112; Std 3: 4.056; and Std 4: 4.084.

[34] As much as this average is for all the three standards, it should be noted that the version of the EGMA used Standard 1was different to the version used with the Standard 3s and 4s.

[35] Results that are statistically significant at the $p < 0.05$ level are referred to as "statistically significantly" lower or higher in the text.

[36] The study of the impact of this program is based on a quasi-experimental research design. As such, systematic differences between the comparison and treatments groups need to be controlled for to have a better measure of the real impact of the program on student's performance on EGMA.

[37] The overall average EGMA score for standard 3 learners in treatment schools increased by 13.51 points from baseline to endline, or 0.56 points per week, considering that approximately 24 weeks of instruction took place between baseline and endline. The overall average EGMA score for counterparts in control increased by 9.52 points over that span, or 0.40 points per week. Therefore, learners in control schools would need about 34 weeks to achieve the same gain that learners in treatment schools—more than 40 percent longer.

peers in control schools to reach.[38] Although the gain of roughly three points in standard 1 is not statistically significant, it would nonetheless also have taken more than 47 percent longer for the peers on control schools to reach[39].

Achievement targets for the NNP were developed in terms of Cohen's D[40]. Based on the short duration of the intervention (24 weeks) the program set a target of a mild to moderate impact in at least two standards. The target was achieved in three standards, standards 1, 3 and 4.

Table 13: Difference-in-difference results for overall EGMA mean scores at baseline and endline by standard

Standard	Baseline		Endline		DID (no covariates)		Covariates	
	Comparison	Treatment	Comparison	Treatment	Coefficient	p-value	Coefficient	p-value
Standard 1	6.91	8.25	14.14	18.63	3.154	0.132	2.462	0.246
Standard 2	19.29	19.24	31.01	30.17	-0.789	0.735	-1.12	0.592
Standard 3	28.92	27.45	38.44	40.96	3.991	0.017**	4.056	0.025**
Standard 4	43.09	40.63	50.81	52.00	3.642	0.074*	4.084	0.04**

Note: Two asterisks (**) denote differences between baseline and endline are statistically significant at the $p < 0.05$ level. One asterisk (*) denotes differences between baseline and endline are statistically significant at the $p < 0.1$ level.[41]

EQ 2. HAVE THE PILOT ACTIVITIES CHANGED LEARNER ENGAGEMENT?

The new features of the NNP seem to have changed some aspects of learner engagement, based on learners' responses and classroom observations (see Annex V for detailed tables). A statistically significantly higher proportion of learners in treatment schools at endline (75.5 percent) than baseline (57.2 percent) said they asked their teachers to explain parts of a lesson again if they did not understand what their teachers said.

Further, more teachers in treatment schools than comparison schools were observed using manipulatives appropriately and effectively for the specific page in the lesson—75.9 percent to 50.9 percent, respectively—which was a statistically significant difference.[42] In addition, although more than five learners were involved in reflection in most classrooms in both treatment and comparison schools—61.2 percent and 85.4 percent, respectively—the nature of the learners' reflection differed. Teachers in treatment schools encouraged learners to

[38] The overall average EGMA score for standard 4 learners in treatment schools increased by 11.37 points from baseline to endline, or 0.47 points per week, considering that approximately 24 weeks of instruction took place between baseline and endline. The overall average EGMA score for counterparts in control increased by 7.72 points over that span, or 0.32 points per week. Therefore, learners in control schools would need more than 35 weeks to achieve the same gain that learners in treatment schools—more than 47 percent longer.

[39] The overall average EGMA score for standard 1 learners in treatment schools increased by 10.38 points from baseline to endline, or 0.43 points per week, considering that approximately 24 weeks of instruction took place between baseline and endline. The overall average EGMA score for counterparts in control increased by 7.23 points over that span, or 0.30 points per week. Therefore, learners in control schools would need more than 35 weeks to achieve the same gain that learners in treatment schools—more than 47 percent longer

[40] Cohen's D is often used to determine the size of the effect observed during an experiment. Typically, the values of Cohen's D are categorized as small (D = 0.2), moderate (D= 0.5) or large (D= 0.8 or above). The sample size used in this independent evaluation only allows the detection of an effect that is 0,35 SD or greater (i.e. between a small and moderate treatment effect).

[41] Due to this evaluation being a pilot, differences at the $p < 0.1$ level are denoted.

[42] What specific manipulatives were used in classrooms during classroom observations is unknown. Enumerators only recorded if teachers were effectively using manipulatives corresponding to the lesson page being taught; they were not asked to record the specific manipulatives being used for the lesson.

respond to their probing questions, but in comparison schools, teachers mostly told learners how they should have responded to problems presented during the lesson.

EQ 3. HOW ARE LEARNERS ENGAGING WITH AND USING THE WORKBOOKS INDEPENDENTLY?

Teachers' views on how learners at treatment schools were engaging with their workbooks positively changed from baseline to endline, with some statistically significant differences (see Annex V for detailed tables).

- **Learners' enjoyment of workbooks:** Nearly all teachers at treatment schools agreed or strongly agreed with the statement, 'learners enjoy working with the workbooks', including a statistically significant increase of those who strongly agree—from 22.5 percent at baseline to 50.0 percent at endline.
- **Difficulty of workbooks:** Statistically significantly fewer teachers at treatment schools at endline than baseline reported that the learner workbooks were too difficult—14.2 percent and 38.4 percent, respectively. At endline, more than three in four teachers said the difficulty level was adequate.
- **Learners' engagement with workbooks:** At baseline, while 58.5 percent of teachers at treatment schools said they had some learners who did not engage with workbooks during lessons, only 40.3 percent of teachers at endline reported they did. This decrease was statistically significant.

In KIIs, teachers at treatment schools shared details about the workbooks' strengths and limitations. A standard 2 teacher praised the new materials for improving learner engagement. 'Workbooks have simplified mathematical lessons since learners can visualise the work and the workbooks bring pleasure to learners; hence they are actively engaged in lessons', she said. Many teachers also noted that they thought the workbooks reduced absenteeism.

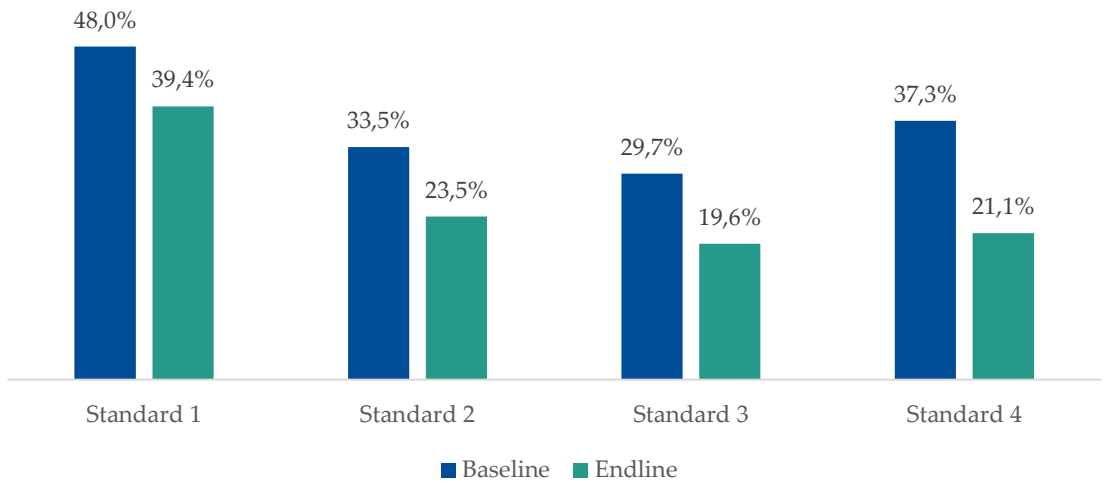
Teachers at treatment schools did mention some drawbacks to the workbooks. Some reported that learners and schools lacked sufficient tools, such as shapes, to be able to properly implement the math methodologies. Each learner is supposed to have a set of shapes provided by the parents or community, not by the teacher, school, or NNP. It is apparent from teacher KIIs that some parents and communities are not providing ample resources, notably in rural areas. 'Some learners fail to accomplish what they have been asked to do due to lack of resources', a standard 2 teacher said. 'Most learners tend to forget the resources they were told to bring for the particular learner activity'.

Multiple trainers and coaches also reported that some parents had issues produced shapes during an earlier research period—ongoing data collection point B. One trainer said, 'The provision of materials should be improved because it is difficult to get the right materials locally, especially the shapes. … They should be plastic in nature or laminated so that they are durable and stand the test of time. Parents are not able to trace the shapes properly, and this gives problems for learners to use such shapes in class'.

Teachers at treatment schools also related how learners engage with their workbooks outside of the classroom. Some teachers said that they allowed their learners to take the workbooks home, which led to mixed results. They described that while some learners benefited from practising what they had learned and revising their work, others had their parents and other family members fill in answers and exercises for them. 'Some parents take their learner's workbook and write answers', a standard 1 teacher said. 'Mostly those answers are always wrong since they do not follow the right strategies'. Learners at treatment schools had more positive views about their workbooks at endline compared with baseline, but the changes were not as pronounced as the changes in teachers' observations (see Annex V for detailed tables). While 12.0 percent of learners at baseline said the workbook was not at all 'easy to use', only 5.2 percent said so at endline. In addition, the proportion of learners who 'completely agreed' that the workbook was 'fun to work in' increased from 72.7 percent at baseline to 77.8 percent at endline.

More learners at treatment schools found it easier to work with the workbook at endline than baseline, as shown in Figure 5. Notably, while 37.0 percent of standard 4 learners at baseline said that the workbook was too difficult to use, only 21.1 percent said so at endline, which was a statistically significant decrease.

Figure 5: Learners reporting the workbook is too difficult to use in treatment group

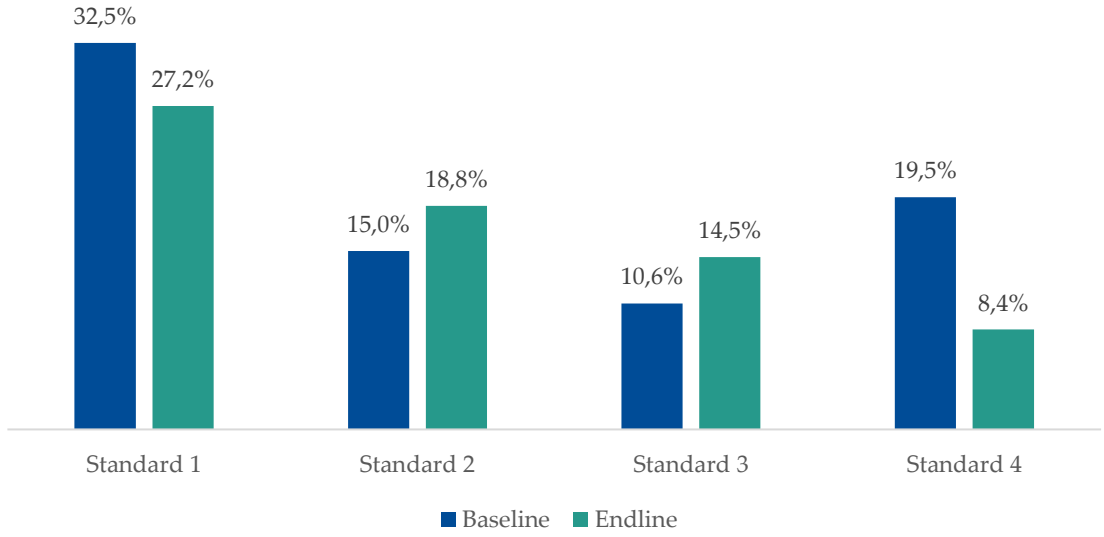


EQ 3A. ARE THERE LANGUAGE ISSUES THAT IMPEDE ON LEARNER ENGAGEMENT WITH THE MATERIALS?

Learners at treatment schools continue to have language issues that may impede their engagement with materials. Although fewer learners at endline than baseline in all four standards reported that the workbook was too difficult for them, learners' self-reported issues with understanding the language in the workbook at baseline persisted at endline, as illustrated in Figure 6. At endline, 27.2 percent of standard 1 learners said they did not understand the language in the workbook, while the proportion of standard 2 and 3 learners reporting similar issues increased from baseline to endline. By contrast, the proportion of

standard 4 learners who said they did not understand the language in the workbook decreased from 20.0 percent at baseline to 8.4 percent at endline.

Figure 6: Learners reporting that they do not understand the language in the workbook in treatment group



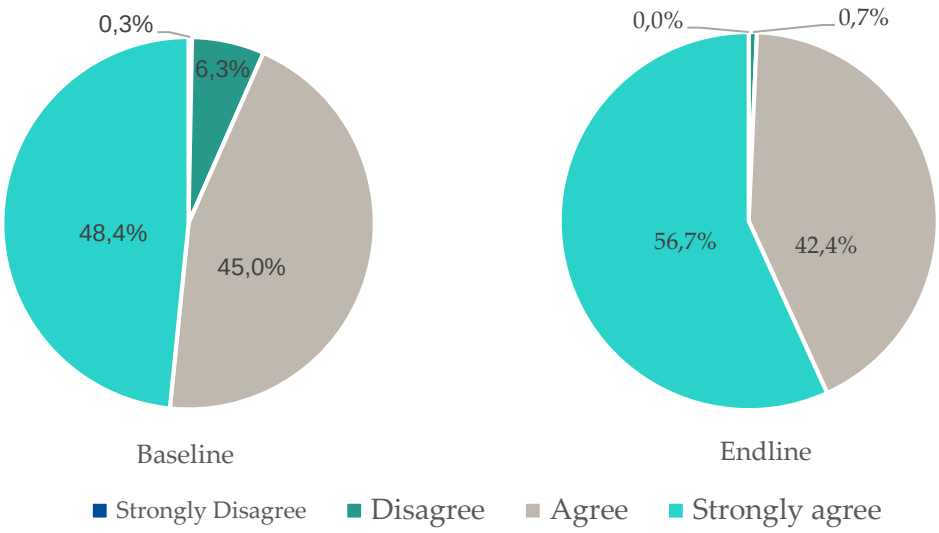
Learners at treatment schools corroborated this finding in FGDs, explaining how teachers mix English and Chichewa or another language like Chitumbuka to help those in their classrooms who struggle with English. A learner in an FGD summed up the difficulty expressed by learners across FGDs: 'No one speaks good English in our class and that can make someone to struggle in mathematics class'.

EQ 4. TO WHAT EXTENT HAVE THE PILOT INTERVENTION'S ACTIVITIES CHANGED CLASSROOM PRACTICES?

In addition to teachers' self-reported responses, the changes in teachers' classroom practices during the pilot were quantified two different ways. First, teachers in treatment schools were observed to determine the extent to which they were delivering mathematics lessons with fidelity—that is, according to NNP guidance. Second, teachers in both comparison and treatment schools were observed to calculate the quality of their mathematics instruction by using a rubric with five categories. These two measures are distinct. Although teachers at treatment schools could follow NNP guidance perfectly by delivering a lesson exactly as they were trained to do, the quality of their instruction could be poor. In other words, while the structure of these lessons would be considered sound, the substance would be flawed.

Nearly all teachers at treatment schools said they believe they were teaching mathematics differently due to NNP. Only 0.7 percent of teachers at endline disagreed or strongly disagreed with the statement, 'The NNP has changed my approach to teaching mathematics', as displayed in Figure 7, while at baseline, a statistically significantly higher percentage of teachers—6.6 percent—disagreed or strongly disagreed with the statement.

Figure 7: Teachers reporting that NNP changed their teaching approach in treatment group



Teachers at treatment schools detailed how their teaching approach has changed in KIIs. A standard 2 teacher described how not only her teaching had improved, but also her understanding of mathematics:

> 'Well, I am really different now, I am a better teacher now. Previously, I could teach even without understanding the concept by just copying the examples in the [former teacher's] guide, but now through mastery of any concept because it is tackled in different perspectives, I can teach without consulting other teachers. It has helped me to be self-reliant'.

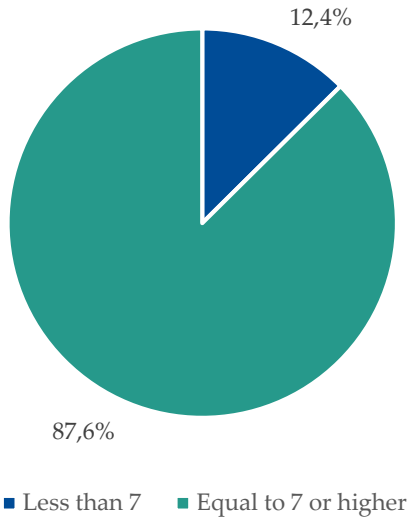
Another standard 2 teacher explained how the NNP had helped her, and her colleagues, be more resourceful—with teachers now able to write a lesson plan without consulting the teacher guide—as well as more proficient with time management because they have to complete one page per day.

Fidelity of Implementation

The vast majority of teachers in treatment schools have changed their practices to implement lessons according to the NNP structure. According to classroom observation data, 87.6 percent of observed teachers in treatment schools at endline implemented the NNP methodology with fidelity, as shown in Figure 8, including 40.9 percent of teachers who scored a perfect 11 out of 11 points on the fidelity composite.[43]

[43] To calculate the proportion of teachers implementing the NNP methodology with fidelity, select items from the classroom observation form were used to compute an 11-point composite. A teacher had to score at least a 7 out of 11 to be considered to be implementing the methodology with fidelity. The composite items included if the lesson plan was informed by the learner workbook; if teacher-led activities prepared learners to do to the page in the workbook; if the teacher set tasks from the workbook that learners must work on independently; the proportion of the lesson during which learners worked independently; if the teacher provided feedback to learners during independent work; if the teacher led reflection activities; what happened during reflection; the number of learners the teacher engaged in discussion during reflection; and what the teacher did if learners made a mistake; and if the lessons was aligned to the page in the workbook.

Figure 8: Proportion of teachers' fidelity scores in treatment schools at endline



Quality of Instruction

To measure quality of instruction, enumerators rated teachers in five different categories on a scale of zero to three—artefacts/manipulatives, writing, methods/procedures, connections, and justification of learner responses. Teachers in treatment schools displayed statistically significant better quality of instruction at endline than their counterparts in control schools in three categories, as displayed in Table 14

- **Methods/procedures:** This category measured the extent to which teachers discussed mathematical methods and procedures and explained why they worked. Teachers scored a three if they provided multiple methods and procedures, including learner production, for the same task, including explanations of why they worked and their advantages.
- **Connections:** This category measured the extent to which teachers connected individual problems or examples. Teachers scored a three if they discussed the connections between different representations of the examples or tasks, including previous examples or tasks that were similar, as well as the manipulatives and/or writing used in the lesson.
- **Justification of learner response:** This category measured the extent to which teachers asked learners to provide responses in class. Teachers scored a three if they invited learner responses and evaluated them not only in terms of being correct or incorrect, but also how and why there were correct or incorrect.

Table 14: Quality of mathematics instruction rubric scores by category

Category	Comparison	Treatment	p-value (total)
	Total	Total	
Artefacts/manipulatives	1.96	2.23	0.085
Writing	2.39	2.45	0.680
Methods/procedures	1.47	1.80	0.004**
Connections	1.81	2.07	0.015**
Justification of learner response	2.17	2.43	0.003**
Total	9.80	10.98	

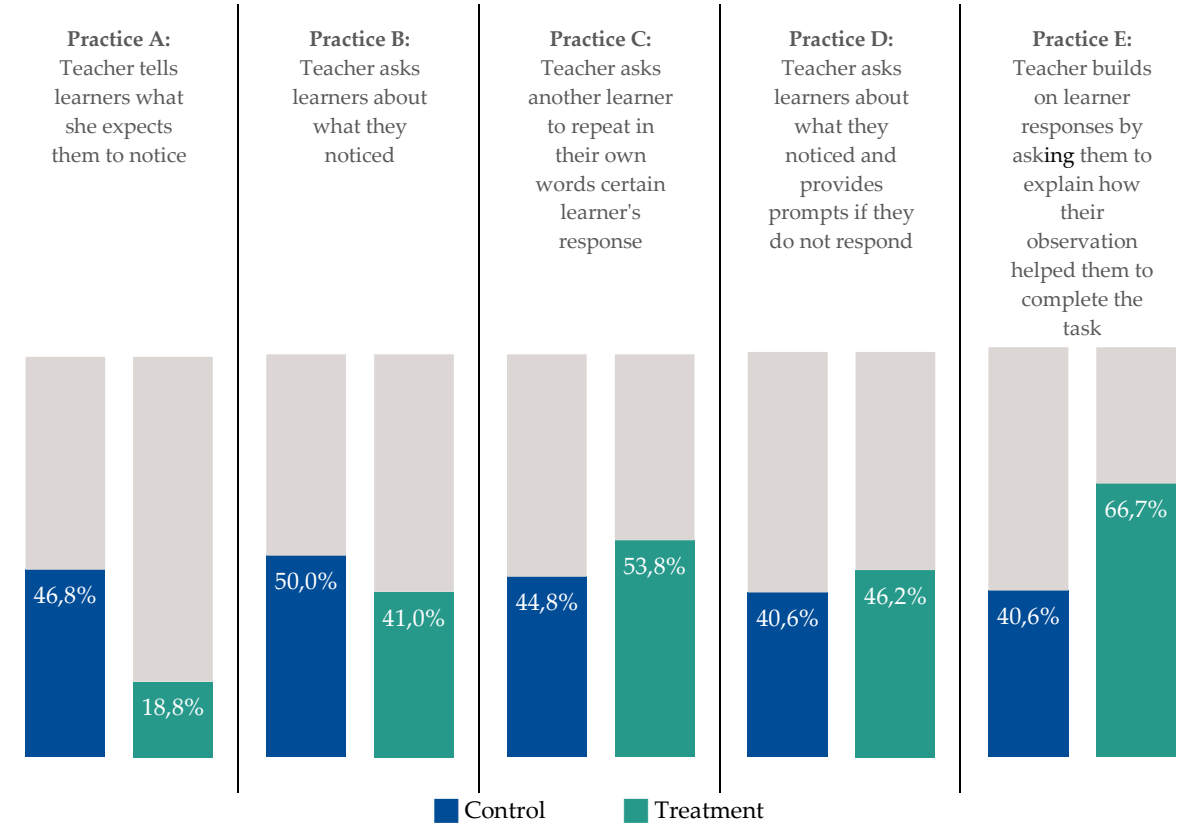
The overall total quality of instruction scores for each standard are shown in Table 15. For scores by category for each standard and a detailed breakdown of how each category was scored, please see Annex IV.

Table 15: Quality of mathematics instruction overall rubric score by standard

Standard	Comparison	Treatment
Standard 1	9.19	10.57
Standard 2	10.00	10.72
Standard 3	9.86	11.02
Standard 4	10.15	11.83

In addition, some statistically significant differences emerged between the practices that teachers at treatment schools demonstrated when leading classroom reflection compared to their counterparts at control schools, as pictured in Figure 9. While observing classrooms, enumerators recorded if teachers used five different practices while leading classroom reflection on tasks and activities learners had just completed. Many teachers employed more than one practice, as detailed in Annex II, with Practices C, D, and E considered more desirable than Practices A and B. Teachers in treatment schools less frequently demonstrated Practices A and B and more frequently used Practices C, D, and E than their counterparts in control schools. Most notably, statistically significantly fewer teachers in treatment schools (18.8 percent) than comparison schools (46.8 percent) told learners what they expected them to notice—one of the two less preferable practices—and statistically significantly more teachers in treatment schools (66.7 percent) than comparison ones (40.6 percent) built on learner responses by asking them to explain how their observation helped them to complete the task—one of the more desirable practices.

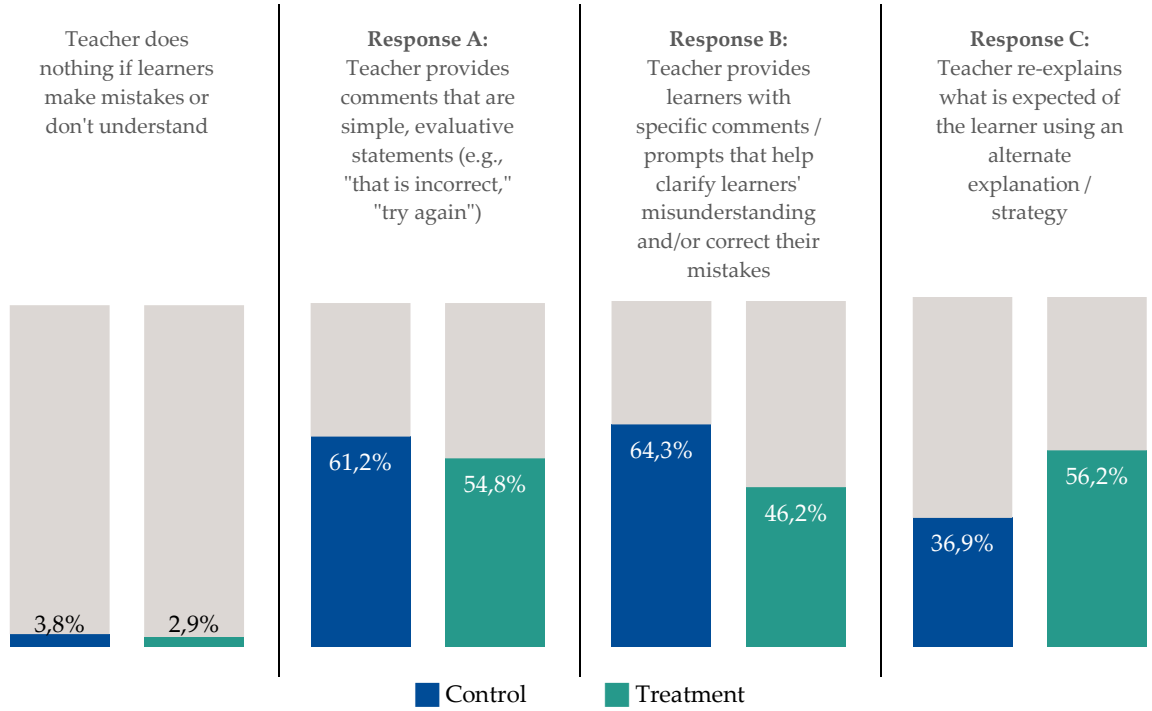
Figure 9: Teaching practices observed in both treatment and control groups during reflection on tasks/activities with learners[44]



When responding to learners who made mistakes, the reaction of teachers at treatment schools differed from their counterparts at comparison schools, as illustrated in Figure 10. While observing classrooms, enumerators recorded how teachers responded when learners made mistakes. Many teachers responded in more than one way, as detailed in Annex II, with Response C considered more desirable than Response A. Teachers in treatment schools responded to mistakes by providing simplistic, evaluative statements (Response A) less frequently than their counterparts in control schools—54.8 percent to 61.2 percent, respectively. By contrast, statistically significantly more teachers in treatment schools responded in a desirable fashion by re-explaining what was expected of learners using an alternate explanation or strategy (Response C) than those in comparison schools—56.2 percent to 36.9 percent, respectively.

To calculate the proportion of teachers implementing the NNP methodology with fidelity, select items from the classroom observation form were used to compute an 11-point composite. A teacher had to score at least a 7 out of 11 to be considered to be implementing the methodology with fidelity. The compo

Figure 10: Teachers' responses to learners' mistakes in observed classrooms in both control and treatment groups[45]



Teachers' interactions with learners during lessons slightly changed from baseline to endline, as detailed in Table 35 in Annex V. The proportion of teachers at treatment schools who reported that they made learners work by themselves five days a week increased from 82.2 percent at baseline to 90.5 percent at endline, but it declined at comparison schools (79.7 percent at baseline and 77.1 percent at endline). As for learners' responses, a similar percentage of learners at both treatment and comparison schools at endline said their teachers treated learners with functional difficulties the same as other children (81.9 percent and 79.8 percent, respectively), as well as provided extra support to struggling learners in math (70.9 percent and 72.4 percent, respectively). In addition, more learners at treatment schools at endline than baseline reported that teachers did not provide extra support to struggling learners—29.1 percent and 18.0 percent, respectively.

EQ 5. IN WHAT WAYS ARE FACE-TO-FACE TRAININGS CHANGING TEACHER PRACTICES IN THE CLASSROOM? EQ 5.A. UNDER WHAT CONDITIONS ARE TEACHER TRAININGS BEING IMPLEMENTED AS INTENDED?

In KIIs conducted at endline, teachers at treatment schools credited trainings with providing them with the knowledge they needed to implement the NNP. 'The training was fruitful as new methods of teaching were attained', a standard 3 teacher said.

Teachers' responses in KIIs conducted at treatment schools primarily focused on the most recent of several rounds of training they had received, which focused on topics such as data

site items included if the lesson plan was informed by the learner workbook; if teacher-led activities prepared learners to do to the page in the workbook; if the teacher set tasks from the workbook that learners must work on independently; the pr

handling and measurement. The training was well-received, according to teachers, and they credited it with providing with them with new materials, strategies, and other methodologies to use in their lessons. For one standard 4 teacher, the recent training on measurement marked the first NNP training she had attended, 'so everything was very important', she said. 'It was an eye-opener. It's where I learnt how to teach the new mathematics'. A standard 3 and 4 teacher described how trainings had taught her how to interact differently with learners in the classroom. '[I learned about] active engagement of learners in teaching and learning', she said. 'Learners are now able to find solutions on their own, they sometimes give solutions that teacher never expected. It's like learners are now the ones teaching the teachers'.

Trainers corroborated teachers' positive views about the effectiveness of trainings during an earlier research period—ongoing data collection point B. Eight of the 10 trainers who participated in KIIs said that training prepared the majority of teachers and trainers well to deliver NNP content, based on the feedback they have received from teachers and subsequent classroom observations that they have conducted. One trainer said that based on classroom observations that 'there are more positive aspects than challenges. … The reflection strategies are being followed by most teachers. Learners like the use of workbooks, and the community likes it too because it involves learners throughout the lessons'.

Overall, all 10 trainers described how the NNP had changed teachers' views about teaching mathematics and learners' capabilities. 'In the past, teachers used to memorize what to do, but now they are thinking critically, and this has helped them to prepare lessons well', one trainer said. 'In the past, examples were already given, but now the teachers have to formulate their own.'

EQ 6. ARE THE PILOT COACHING SESSIONS AND TEACHER LEARNING CIRCLES (TLC) WORKING AS INTENDED? EQ 7.A. UNDER WHAT CONDITIONS ARE THE COACHING SESSIONS EFFECTIVE?

Teachers at treatment schools did not speak in great detail about coaching sessions or TLCs during the KIIs conducted at endline. While most agreed that these support activities were at least minimally helpful, they largely used coaching and TLCs as a forum for lobbying for additional resources.

More comprehensive data related to coaching and TLCs was collected during ongoing data collection point B. Enumerators conducted KIIs with 10 trainers and 10 coaches and observed TLCs, and teachers at treatment schools completed an anonymous questionnaire about TLCs.

All coaches interviewed credited the coaching model with helping teachers improve their practices in the classroom by helping them get up to speed on the new content and prepare more effectively to deliver it. Most coaches said that teachers needed help in two areas—how to get and use resources in the classroom and how to manage large classes. Seven coaches reported each of these as areas in which teachers needed more support. Several coaches described how these two areas of need were connected, as teachers had trouble

getting enough resources for such large classes. The two primary challenges related to their own roles that coaches mentioned were the difficulty to find time to observe fellow teachers and issues with the classroom observation tool.

Some of the key findings from TLCs observations and teachers' perceptions of TLCs included:

- Teachers at treatment schools were most satisfied with how TLCs have helped them understand how to explain or do the activities in the learner workbook, the central activity in the lesson routine practice session, as well as how the TLCs have taught them to better explain math concepts to their pupils.
- Most teachers at treatment schools said that they always find the numeracy TLCs interesting and very useful (71.9 percent) and learnt something new about teaching math in every TLC (71.6 percent).
- The duration of TLCs varied, but they were generally shorter than recommended.
- Head teachers and teacher advisors played a limited role in the TLCs.

EQ 7. IN WHAT WAYS ARE THE TEACHER GUIDES SUPPORTING TEACHERS TO UNDERSTAND THE METHODOLOGY/APPROACH BEING IMPLEMENTED? [46]

EQ 8. IN WHAT WAYS DO THE TEACHER GUIDES SUPPORT TEACHERS IN IMPLEMENTING THE INTENDED METHODOLOGY APPROACH AND USING LEARNER MATERIALS EFFECTIVELY AND WITH FIDELITY?

Overall, a greater proportion of teachers at treatment schools at endline than baseline strongly agreed with statements about the NNP teacher guide, though the percentage of teachers who disagreed or strongly disagreed remained the same (see Annex V for detailed tables). For instance, the proportion of teachers who strongly agreed that the guide is 'easy to use' increased from 19.1 percent at baseline to 29.9 percent at endline, while the percentage who strongly disagreed or disagreed was similar—15.5 percent at baseline and 15.7 percent at endline. In addition, nearly half of the teachers at endline—47.8 percent—strongly agreed that the teacher guide 'provides sufficient guidance' on implementing NNP lessons, a statistically significant increase from 28.4 percent at baseline. Again, the proportion of teachers who disagreed or strongly disagreed was relatively unchanged at both time points—9.9 percent at baseline and 8.9 percent at endline.

Still, some questions remain for teachers at treatment schools about NNP implementation, though they reported having fewer at endline than baseline. At endline, 44.6 percent of teachers strongly agreed or agreed with the statement, 'I have many questions that the teacher guide does not address', compared with 52.4 percent at endline, as shown in Table

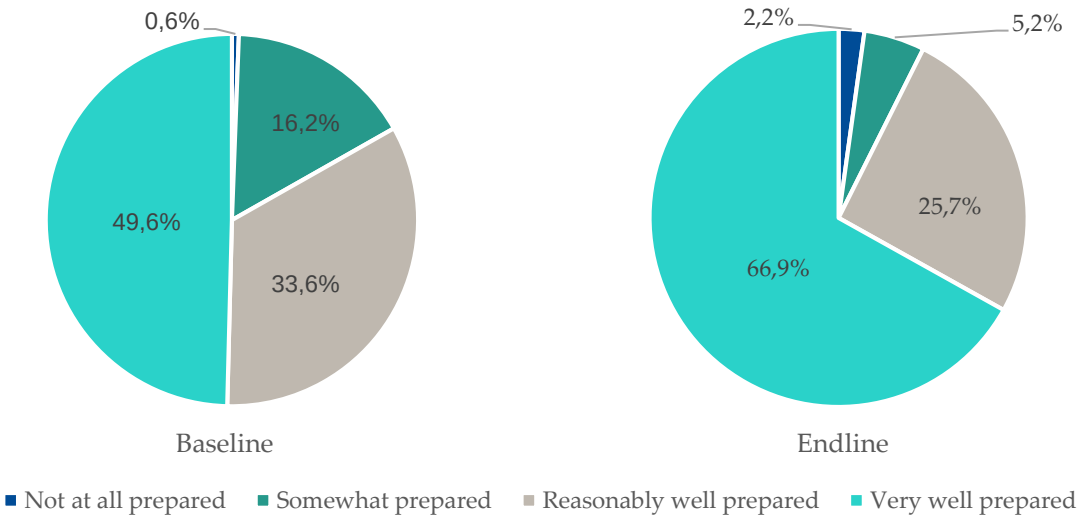
oportion of the lesson during which learners worked independently; if t

36. Notably, the proportion of teachers who strongly disagreed with the statement increased from 2.4 percent at baseline to 12.7 percent at endline.

As part of their preparation, teachers at treatment schools said they utilized other resources along with the teacher guide. A vast majority of teachers at endline—91.8 percent—reported reviewing or referencing the learner workbooks daily when developing their lessons, a slight increase from the proportion who said they did so at baseline—87.7 percent. During classroom observations, nearly all teachers had a lesson plan—95.6 percent (see Annex V for detailed tables).

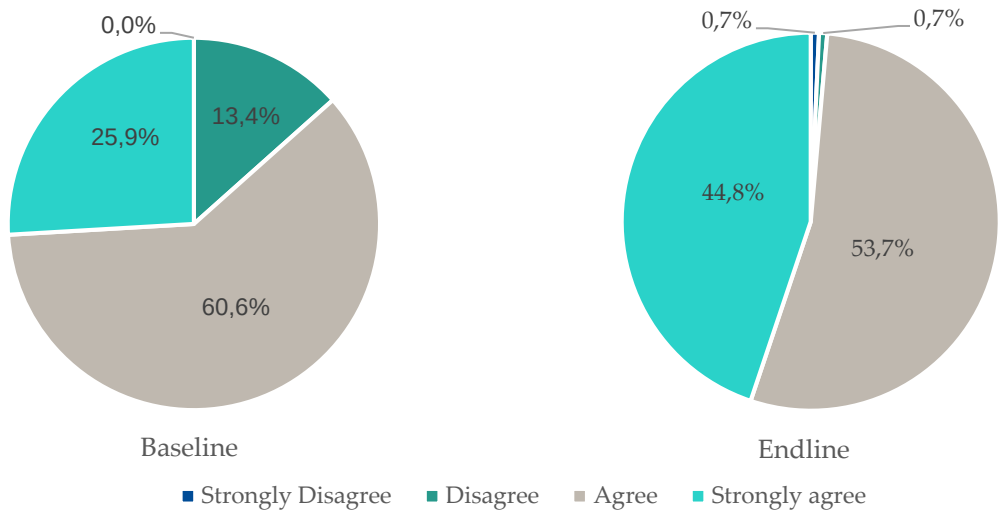
Overall, most teachers at treatment schools said they felt prepared to implement the NNP (Figure 11). At endline, 66.9 percent of teachers said they felt 'very well prepared' to implement the new program, which was a statistically significant increase from the proportion of teacher who reported so at baseline (49.6 percent). The percentage of teachers who reported being 'not at all prepared' or 'somewhat prepared' decreased from 16.8 percent at baseline to 7.4 percent at endline.

Figure 11: Teachers in treatment group reporting that they feel prepared to implement NNP



Nearly all teachers at treatment schools agreed that they valued the NNP materials. At endline, 44.8 percent of teachers strongly agreed with the statement that 'the teacher guide supports me to implement the NNP effectively', as shown in Figure 12, which was a statistically significant increase from baseline (25.9 percent). Only 1.4 percent strongly disagreed or disagreed with the statement at endline, compared with 13.4 percent who disagreed at baseline.

Figure 12: Teachers in treatment group reporting that they value NNP materials



EQ 9. HOW ARE THE TRAINING VIDEOS BEING USED? EQ 9.A. ARE THE TRAINING VIDEOS PERCEIVED AS A USEFUL TRAINING RESOURCE? EQ 9.B. HOW COULD THEY BE IMPROVED AND MADE MORE USEFUL?

Teachers in treatment schools provided both quantitative and qualitative data at endline that revealed their nuanced views about training videos. Nearly all teachers viewed the videos as an asset. The vast majority—94.5 percent—agreed or strongly agreed with the statement, 'NNP training videos are very helpful in showing me what to do', with 36.2 percent strongly agreeing with the statement. Teachers also believed that the videos corresponded to what trainers have imparted to them. Only 15.8 percent of teachers agreed or strongly agreed that the videos 'contradict the information I have received from the NNP coaches or trainers' (see Annex V for detailed tables).

Despite teachers at treatment schools reporting their overall satisfaction with the videos, they also shared that they could be improved and made more useful. At endline, 70.0 percent of teachers agreed or strongly agreed that the videos 'are too short to be helpful', while more than half of teachers at endline (51.0 percent) disagreed or strongly disagreed that the videos 'cover all the issues I need support with'.

Qualitative data from teacher KIIs at treatment schools matched what they shared on the teacher questionnaires. Although teachers said they valued the videos, they suggested certain improvements that they and other respondents such as coaches and trainers already mentioned at baseline and other data collection periods. Many teachers explained how the videos were not representative of their classrooms, with the videos featuring small classes with many high-performing learners. 'Videos should be captured with large classes and with learners who are from the rural setting so as to give a true picture of what will be happening on the ground', a standard 1 and 2 teacher said.

Some teachers at treatment schools also said they have issues accessing the videos due to issues with their devices. The videos can take up a lot of space on smartphones due to their large file size, while some teachers have difficulty viewing the videos on their older devices. 'Some of the phones used to watch the videos do not have a clear focus', said a standard 2 teacher, 'hence, we are unable to view the demonstrations clearly'.

CONCLUSIONS AND RECOMMENDATIONS

CONCLUSIONS

Several statistically significant differences emerged between learners' performance on the EGMA at treatment schools compared with comparison schools over the course of the pilot evaluation. The only statistically significant difference in gains of overall EGMA scores from baseline to endline between school groups was found in standard 3, with the average gain for learners in treatment schools nearly four points greater than the average gain for their peers in comparison schools. In other words, the gains that standard 3 learners in treatment schools achieved would have taken more than 40 percent longer for their counterparts in control schools to attain. As for individual subtask results, most notably, the gains for standard 1 learners at treatment schools from baseline to endline were statistically significantly greater than the gains for their counterparts at comparison schools on three subtasks—number identification, pattern completion, and problems. When considering gains from baseline to endline, it is important to note that the short time frame between the baseline in January and February of 2022 and the endline in August of 2022—nearly seven months that amount to no more than two-thirds of a typical school year in Malawi—likely resulted in limited variation in learner performance between comparison and treatment schools.

As for NNP materials, the statistically significant increases in how teachers and learners in treatment schools positively view them illustrates how their use has become ingrained in classrooms and how teachers and learners feel more comfortable using them. For instance, the proportion of teachers who strongly agreed with the statement that the 'teacher guide supports me to implement the NNP effectively' statistically significantly increased from 25.9 percent at baseline to 44.8 percent at endline, and the proportion of learners who stated they had difficulty with NNP workbooks declined from baseline to endline in all four standards. These finding may be related to several factors, including that classes have covered more material at endline or that teachers' and learners' familiarity with materials has made them more confident in their engagement with materials.

Despite fewer learners at treatment schools stating they have difficulty with materials, language of instruction remains an issue for some learners. Although the proportion of learners at treatment schools who reported the workbook was too difficult to use declined for all four standards from baseline to endline, the proportion of learners who said they did not understand the language in the workbooks remained similar between the two time points in standards 2 and 3. Learners corroborated this finding in FGDs, explaining how teachers mix English and Chichewa or another language like Chitumbuka to help those in their classrooms who struggle with English. Learners' struggles with English as the language of instruction in mathematics are not unexpected because all other subjects do not use English materials or English as the language of instruction.

Teachers in treatment schools, according to classroom observation data, are displaying higher-quality instruction than their peers in comparison schools in certain domains. Teachers in treatment schools, for instance, more effectively discussed mathematical methods and procedures and explained why they worked than teachers at comparison schools. This demonstration of higher-quality teaching in treatment schools than comparison schools supports the responses from teachers in KIIs about how the NNP has improved their analytical skills in mathematics.

Classroom observation data from treatment schools reveals that one of the main objectives to introducing the NNP—for learners to not only know mathematics, but to make sense of it and reason with it—is likely taking hold in classrooms, notably during lessons' reflection period. During reflection activities, while statistically significantly more teachers in comparison schools than treatment schools told learners what they expected them to notice, statistically significantly more teachers in treatment schools than comparison ones asked learners to explain how their observations helped them complete tasks during lessons. These findings indicate that reflection activities may be more profound in treatment schools, with active listening and learning from others more encouraged, and corroborates the findings in the quality of instruction data that teachers in treatment schools more effectively invite learner responses and respond to them than their peers in comparison schools.

Highlighting the effect of NNP's training and support activities, 87.5 percent of observed teachers in treatment schools at endline implemented the NNP methodology with fidelity. This finding corroborates teachers' self-reported data, namely 66.9 percent of teachers in treatment schools reporting that they felt 'very well prepared' to implement the NNP.

Despite expressing how videos are beneficial, teachers at treatment schools echoed the same issues that others had voiced during earlier data collection periods. Teachers' suggested improvement for videos have been well-documented over the course of the pilot evaluation. First, the videos do not mirror many teachers' classrooms because they feature high-achieving learners in small-sized classrooms. Second, some teachers have difficulty accessing the videos due to their large file size or not having a smartphone or computer to view them.

RECOMMENDATIONS

Further investigate the specific mathematics skills on which learners in treatment schools statistically significantly outperformed their peers in comparison schools. For example, standard 1 learners in treatment schools scored statistically significantly higher on two subtasks assessing more foundational skills —number identification and number discrimination. It may be interesting to research if exposure to numeracy in pre-primary may have affected results or to follow this cohort of learners to standard 2 next school year to see if the statistically significant differences persist.

Better understand the reasons why, at endline, fewer teachers and learners believed the materials were less difficult for learners. More research is likely needed to understand why more teachers and learners believed at endline that the workbook was not difficult for learners. Data collected to answer this question could inform the scale-up of the NNP to schools across Malawi to mitigate the proportion of learners and teachers who may have initial difficulty with materials.

Research the difficulties that some learners have with mathematics being taught in English. More data is needed to determine the extent to which learners have issues with the language of instruction, including possible effects on EGMA performance, as well as any effects on specific groups of learners, including those in rural areas or those who speak certain local languages in Malawi.

Share relevant details of the pilot research when the NNP is scaled. For instance, the number of learners who said the workbook was too difficult decreased from baseline to endline, which is relevant information to relay to teachers and others so that they do get discouraged if some learners initially struggle with the new materials.

Determine if any additions to content training videos are feasible and if any devices could be provided to teachers to improve their access to them. At endline, teachers continued to suggest that videos could be improved if they portrayed the realities of classrooms more accurately, with larger class sizes and learners of varied abilities. If it is not feasible for NNP to produce additional videos to address these concerns, perhaps it could include disclaimers to teachers on the videos' limitations to ensure teachers that the project is aware of their concerns. In addition, some teachers reported having difficulty accessing video content due to their smartphones being obsolete or the cumbersome video file sizes. To ensure all teachers can access training content, the possibility of providing devices to schools for this direct purpose should be explored.

ANNEXES

ANNEX I: EGMA CUT SCORES[47]

Overall, although performance improved for learners from baseline to endline, most learners did not meet or exceed expectations for mathematics skills, as the score distributions for the addition and subtraction level 1 subtasks illustrate in all four standards.[48] Performance was classified using three categories for all subtasks except for pattern completion and problems—no performance (or zero correct answers on the subtask); partially meeting expectations; and meeting or exceeding expectations. Using the Global Proficiency Framework (GPF), a cut score for each individual subtask was set for minimum proficiency in each of the four standards.

In standard 1, only about one in three learners in both treatment and comparison schools met or exceeded expectations in addition, as displayed in Table 16, and only about one in four learners did so in subtraction, as detailed in Table 17.

Table 16: Standard 1 cut scores for addition level 1 at endline, based on GPF

Category	No Performance	Partially Meets Expectations	Meets or Exceeds Expectations
Questions	n/a	1–3, 5	4, 6–20
Performance Descriptor	Unable to add any one-digit numbers	Add numbers to 5	Add numbers greater than 5
Range of Scores, Per Category	0	1–4	5–20
Cut Score, Per Category	0	1	5
% of Learners in Category in Treatment Schools	25%	42%	34%
% of Learners in Category in Comparison Schools	39%	28%	34%

Table 17: Standard 1 cut scores for subtraction level 1, based on GPF

Category	No Performance	Partially Meets Expectations	Meets or Exceeds Expectations
Questions	n/a	1–3, 5	4, 6–20
Performance Descriptor	Unable to subtract any one-digit numbers	Subtract numbers to 5	Subtract numbers between 6–20
Range of Scores, Per Category	0	1–4	5–20
Cut Score, Per Category	0	1	5
% of Learners in Category in Treatment Schools	45%	27%	28%

he teacher provided feedback to learners during independent work; if the teacher led reflection activities; what happened during reflection

; the number of learners the teacher engaged in discussion during reflection; and what the teacher did

Category	No Performance	Partially Meets Expectations	Meets or Exceeds Expectations
% of Learners in Category in Comparison Schools	45%	31%	24%

In standard 2, the majority of learners in both treatment and comparison schools—about three in four—partially met expectations in addition, as displayed in Table 18, while less than one in five met or exceeded expectations in both school groups. Performance on subtraction was stronger, with approximately one in four learners in both school groups meeting or exceeding expectations, as shown in Table 19.

Table 18: Standard 2 cut scores for addition level 1, based on GPF

Category	No Performance	Partially Meets Expectations	Meets or Exceeds Expectations
Questions	n/a	1–8, 10–14	9, 15–20
Performance Descriptor	Unable to add any one-digit numbers	Add numbers to 10	Add numbers greater than 11
Range of Scores, Per Category	0	1–13	14–20
Cut Score, Per Category	0	1	10
% of Learners in Category in Treatment Schools	6%	77%	17%
% of Learners in Category in Comparison Schools	7%	76%	18%

Table 19: Standard 2 cut scores for subtraction level 1, based on GPF

Category	No Performance	Partially Meets Expectations	Meets or Exceeds Expectations
Questions	n/a	1–8, 10–14	4, 6–20
Performance Descriptor	Unable to subtract any one-digit numbers	Subtract numbers to 10	Subtract numbers between 11–20
Range of Scores, Per Category	0	1–13	14–20
Cut Score, Per Category	0	1	10
% of Learners in Category in Treatment Schools	18%	57%	24%
% of Learners in Category in Comparison Schools	19%	56%	25%

In standard 3, the majority of learners partially met expectations. Approximately three in four learners in both treatment and comparison schools did so in both addition and subtraction, as detailed in Table 20. More learners met or exceeded expectations in addition—21.0 percent in treatment schools and 23.0 percent in comparison schools—than subtraction—about one in 10 learners in both school groups.

Table 20: Standard 3 cut scores for addition and subtraction level 1, based on GPF

Category	No Performance	Partially Meets Expectations	Meets or Exceeds Expectations
Questions	n/a	1–8, 10–14	9, 15–20
Performance Descriptor	Unable to add/subtract any one-digit numbers	Add/subtract numbers to 10	Add/subtract numbers greater than 11
Range of Scores, Per Category	0	1–13	14–20
Cut Score, Per Category	0	1	14
% of Learners in Category for Addition Level 1 at Treatment Schools	3%	76%	21%
% of Learners in Category for Addition Level 1 at Comparison Schools	2%	75%	23%
% of Learners in Category for Subtraction Level 1 at Treatment Schools	12%	78%	10%
% of Learners in Category for Subtraction Level 1 at Comparison Schools	16%	75%	9%

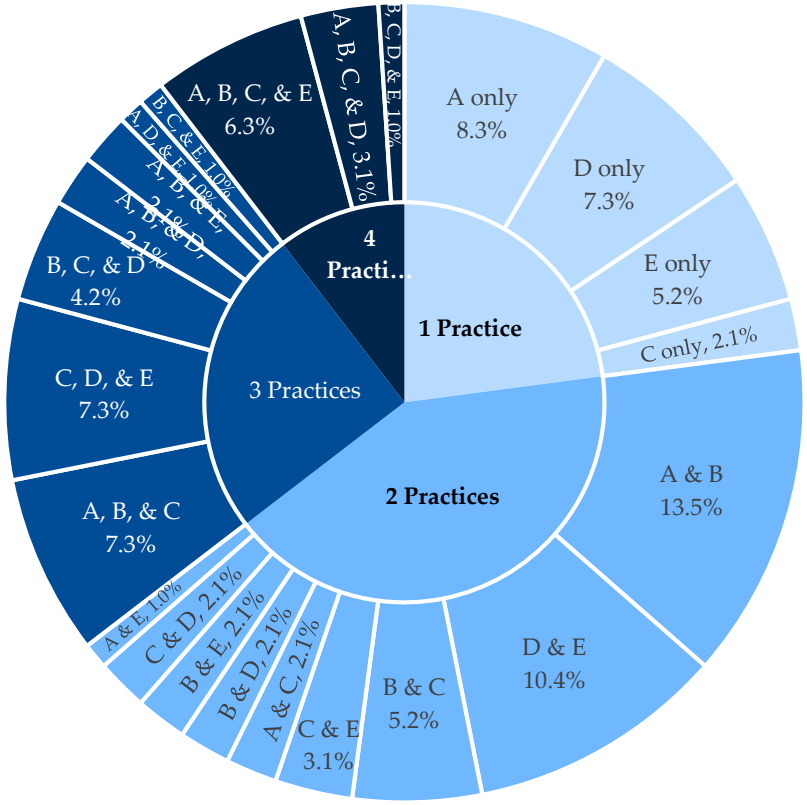
In standard 4, more learners met or exceeded expectations in both addition and subtraction, as shown in Table 21. Nearly half of learners did so in addition in both treatment and comparison schools, while about one in three learners did so in subtraction in both school groups.

Table 21: Standard 4 cut scores for addition and subtraction level 1, based on GPF

Category	No Performance	Partially Meets Expectations	Meets or Exceeds Expectations
Questions	n/a	1–8, 10–14	9, 15–20
Performance Descriptor	Unable to add/subtract any one-digit numbers	Add/subtract numbers to 10	Add/subtract numbers greater than 11
Range of Scores, Per Category	0	1–13	14–20
Cut Score, Per Category	0	1	14
% of Learners in Category for Addition Level 1 at Treatment Schools	1%	50%	49%
% of Learners in Category for Addition Level 1 at Comparison Schools	0%	51%	49%
% of Learners in Category for Subtraction Level 1 at Treatment Schools	4%	64%	33%
% of Learners in Category for Subtraction Level 1 at Comparison Schools	7%	60%	33%

ANNEX II: OBSERVED TEACHER PRACTICES

Figure 13: Frequency of teachers' reflection practices with learners, control group



Practice A: Teacher tells learners what she expects them to notice.

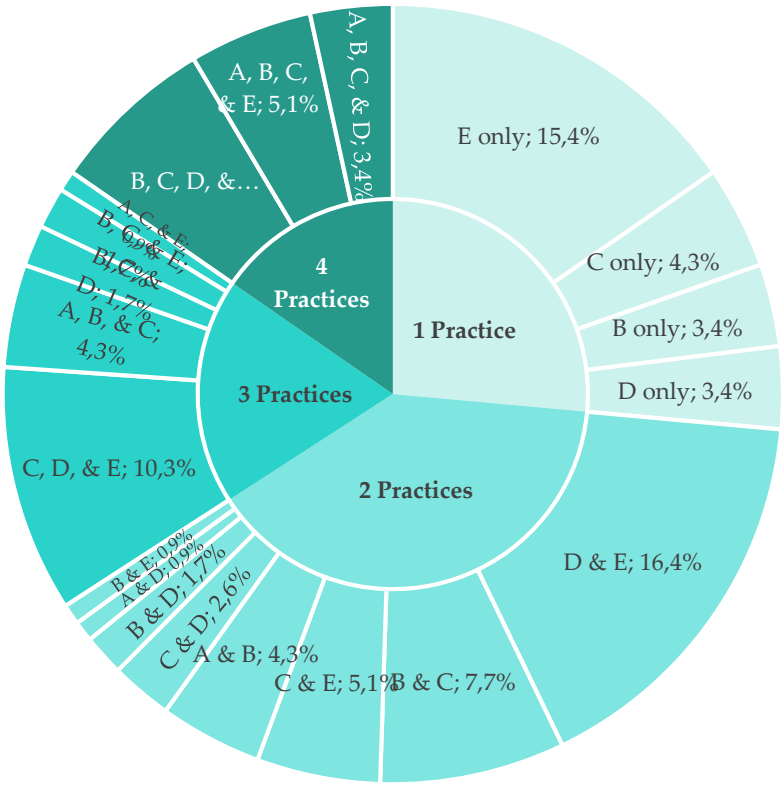
Practice B: Teacher asks learners about what they noticed.

Practice C: Teacher asks another learner to repeat in their own words certain learner's response.

Practice D: Teacher asks learners about what they noticed and provides prompts if they do not respond.

Practice E: Teacher builds on learner responses by asking them to explain how their observation helped them to complete the task.

Figure 14: Frequency of teachers' reflection practices with learners, treatment group



Practice A: Teacher tells learners what she expects them to notice.

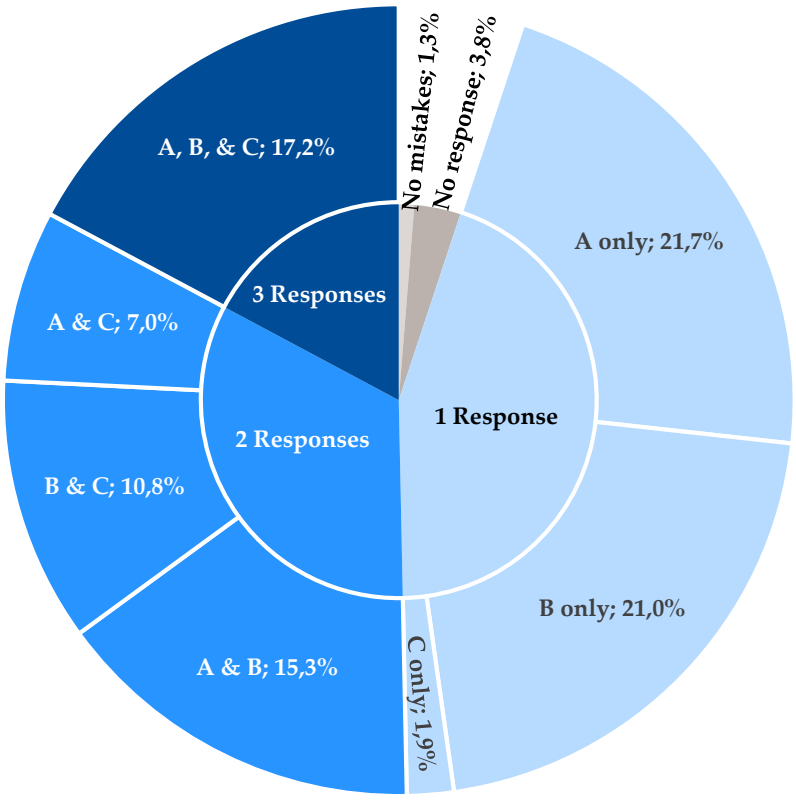
Practice B: Teacher asks learners about what they noticed.

Practice C: Teacher asks another learner to repeat in their own words certain learner's response.

Practice D: Teacher asks learners about what they noticed and provides prompts if they do not respond.

Practice E: Teacher builds on learner responses by ask them to explain how their observation helped them to complete the task.

Figure 15: Frequency of teachers' response to learner's mistakes, control group

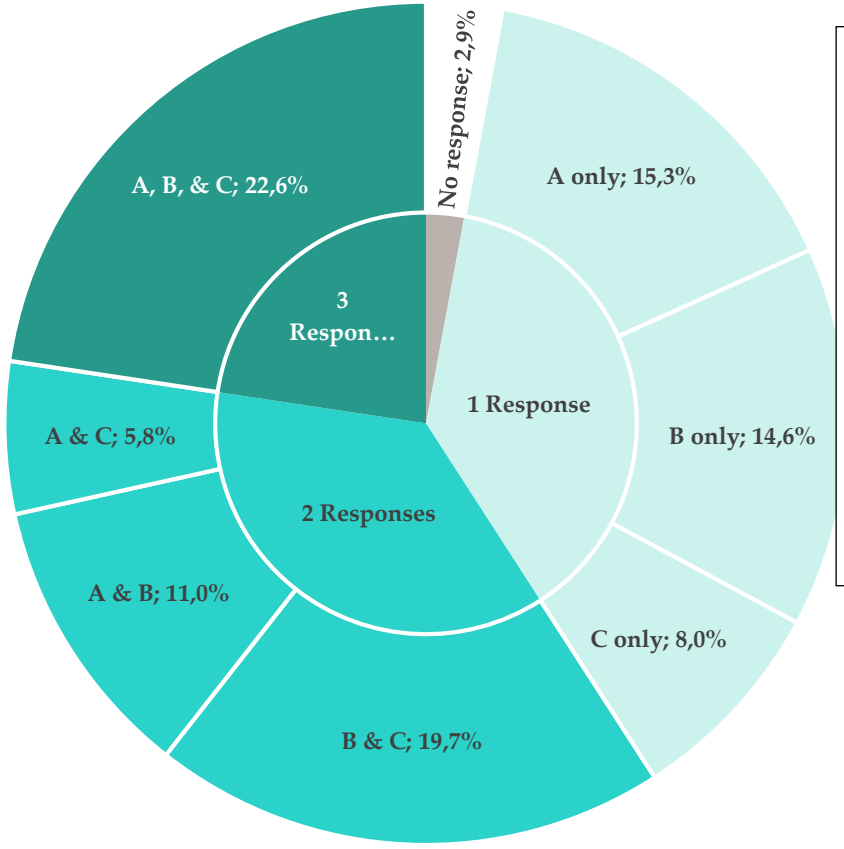


Response A:
Teacher provides comments that are simple, evaluative statements (e.g., "that is incorrect," "try again").

Response B:
Teacher provides learners with specific comments / prompts that help clarify learners' misunderstanding and/or correct their mistakes.

Response C:
Teacher re-explains what is expected of the learner using an alternate explanation or strategy.

Figure 16: Frequency of teachers' response to learner's mistakes, treatment group



Response A:
Teacher provides comments that are simple, evaluative statements (e.g., "that is incorrect," "try again").

Response B:
Teacher provides learners with specific comments / prompts that help clarify learners' misunderstanding and/or correct their mistakes.

Response C:
Teacher re-explains what is expected of the learner using an alternate explanation or strategy.

ANNEX III: RELIABILITY MEASURES

The Cronbach alpha—an estimate of reliability of a subtask's scores—was calculated for each EGMA subtask to assess its psychometric qualities. Cronbach alpha scores were computed separately for subtasks.

Table 22: Cronbach Alpha Values by Subtask

Subtasks	Standard 1	Standard 2	Standard 3	Standard 4
Number identification	0.7888	0.8013	0.6937	0.6706
Number discrimination	0.8234	0.8838	0.8282	0.8039
Pattern recognition	0.5928	0.5627	0.4886	0.4966
Addition Level 1	0.8576	0.8878	0.9101	0.9132
Subtraction Level 1	0.8732	0.9050	0.9092	0.9119
Addition Level 2	n/a	n/a	0.6268	0.5934
Subtraction Level 2	n/a	n/a	0.5840	0.5185
Problems	0.6234	0.6699	0.6388	0.4822

ANNEX IV: QUALITY OF INSTRUCTION RUBRIC

To measure quality of instruction, enumerators rated teachers in five different categories on a scale of zero to three—artefacts/manipulatives, writing, methods/procedures, connections, and justification of learner responses. The scores for each standard by category are presented in this annex, along with a breakdown of how each category was scored. The rubric was developed by Aarnout Brombacker, Fraser Gobede, Justina Longwe, and Mercy Kazima based on a 2018 article published by Hamsa Venkat and Mike Askew.[49]

Table 23: Quality of mathematics instruction rubric score for standard 1

Category	Comparison	Treatment
Artefacts/manipulatives	2.05	2.06
Writing	2.15	2.37
Methods/procedures	1.34	1.85
Connections	1.60	2.03
Justification of learner response	2.05	2.26
Total	9.19	10.57

Table 24: Quality of mathematics instruction rubric score for standard 2

Category	Comparison	Treatment
Artefacts/manipulatives	1.97	2.21
Writing	2.38	2.21
Methods/procedures	1.59	1.85
Connections	1.85	2.03
Justification of learner response	2.21	2.42
Total	10.00	10.72

Table 25: Quality of mathematics instruction rubric score for standard 3

Category	Comparison	Treatment
Artefacts/manipulatives	2.00	2.18
Writing	2.54	2.40
Methods/procedures	1.34	1.85
Connections	1.85	2.03
Justification of learner response	2.13	2.56
Total	9.86	11.02

Table 26: Quality of mathematics instruction rubric score for standard 4

Category	Comparison	Treatment
Artefacts/manipulatives	1.84	2.47
Writing	2.51	2.82
Methods/procedures	1.59	1.92
Connections	1.92	2.12
Justification of learner response	2.29	2.50
Total	10.15	11.83

[49] Venkat, H. & Askew, M. (2018). Mediating primary mathematics: theory, concepts and a framework for studying practice. Educational Studies in Mathematics, 97, 71–92.

The five categories in the quality of instruction rubric are presented below with the criteria for each rating on a scale from 0 to 3.

Table 27: Artefacts/manipulatives

Score	Criteria
0	No manipulatives are used, and the lesson would have benefited from using manipulative(s)
1	The teacher uses manipulatives; however, the link to the mathematical task(s) of the lesson is unclear
2	The teacher uses manipulatives, and they help to clarify the mathematics task/concept
3	Either the teacher uses manipulatives, and they help to clarify the mathematics task/concept, and the learners can use them independently to complete the same/similar mathematics task(s) or the teacher did not use manipulatives because learners are able to complete the mathematics task(s) confidently without manipulatives

Table 28: Writing

Score	Criteria
0	No writing by the teacher on the board or on a chart
1	There is writing on the board or on a chart; however, it does not support concept development (e.g. date, register, exercise to be completed, etc.)
2	There is writing on the board or on a chart, and it supports concept development; however, it includes mathematical errors that go unnoticed
3	There is writing on the board or on a chart, and it supports concept development; the writing may include mathematical errors which are noticed and addressed

Table 29: Methods/procedures

Score	Criteria
0	No discussion (telling) of methods or procedure for mathematical task
1	A single mathematical method/procedure is provided, and the method only applies to a specific problem/task
2	A mathematical method/procedure is provided together with an explanation of why/how the method/procedure works
3	Alternative mathematical methods/procedures, including learner productions, for the same mathematical task are discussed, including explanations of why/how they work as well as the advantage of each

Table 30: Connections

Score	Criteria
0	Mathematical examples/tasks are dealt with thorough guessing/chorusing
1	Mathematical examples/tasks are treated in isolation
2	Mathematical examples/tasks are treated in relation to similar examples/tasks

3	There is a discussion of the connections between different representations of the mathematics examples/tasks (e.g. similar previous examples/tasks; the manipulatives and/or writing used in the lesson)

Table 31: Justification of learner response

Score	Criteria
0	No mathematical responses are invited from learners
1	Learners' mathematical responses are invited, but not evaluated
2	Learners' mathematical responses are invited and evaluated in terms of yes/no, correct/incorrect, etc.
3	Learners' mathematical responses are invited and evaluated in terms of why/how they are correct/incorrect

ANNEX V: QUALITATIVE TABLES

Table 32: Perceptions regarding learners' engagement following NNP in treatment group

	Answer Option	Baseline	Endline
Learner Responses			
What do you do if you do not understand your teacher in a mathematics lesson?	Nothing	16.6%	11.2%
	I try to understand without asking for help	9.1%	7.2%
	I ask a friend/classmate to explain again	31.8%	30.8%
	I ask the teacher to explain again	57.2%	75.5%

Table 33: Teachers' responses to learners' engagement with workbooks in treatment group

	Answer Option	Baseline	Endline
Teacher Responses			
Are there learners who do not engage with their mathematics workbooks during the mathematics lesson?	No	41.5%	59.7%
	Yes	58.5%	40.3%
If yes, what portion of your learners do not engage with the workbook?	Few (less than 25 percent)	85.1%	85.2%
	Some (25 to 49 percent)	9.4%	3.7%
	A lot (50 to 74 percent)	4.5%	1.8%
	Most (more than 75 percent)	1.0%	9.2%
Do your learners often have questions about what they are expected to do in the workbooks?	No	26.0%	27.8%
	Yes	74.0%	72.2%
Are the learner workbooks too easy, too difficult, or adequate for learners?	Too easy	6.2%	9.0%
	Adequate	55.4%	76.7%
	Too difficult	38.4%	14.2%
Learners enjoy working with the workbooks.	Strongly disagree	0.4%	0%
	Disagree	10.7%	2.3%
	Agree	66.4%	47.8%
	Strongly agree	22.5%	50.0%

Table 34: Learners' responses to engagement with workbooks in treatment group

	Answer Option	Baseline	Endline
Learner Responses			
This [hold up/show] workbook is easy to use.	Not at all	12.0%	5.2%
	A little bit	32.6%	31.1%

	Answer Option	Baseline	Endline
	Completely	55.4%	63.6%
This [hold up/show] workbook is fun to work in.	Not at all	5.9%	2.2%
	A little bit	21.3%	19.9%
	Completely	72.7%	77.8%
This [hold up/show] workbook is the best book I have owned.	Not at all	6.1%	2.2%
	A little bit	24.1%	25.2%
	Completely	69.8%	72.6%
I like working in this [hold up/show] workbook.	Not at all	5.4%	2.2%
	A little bit	22.5%	24.3%
	Completely	72.1%	73.5%
I understand the language used in this [hold up/show] workbook.	Not at all	18.8%	17.7%
	A little bit	38.8%	46.6%
	Completely	42.3%	35.7%

Table 35: Teachers' interaction with learners during lessons

	Answer Option	Baseline		Endline	
		Comparison	Treatment	Comparison	Treatment
Teacher Responses					
On how many days each week do you make learners work by themselves, either alone or in a group?	No days	0.0%	2.7%	11.46%	0%
	1 day	0.0%	2.2%	0%	0.73%
	2 days	1.0%	5.5%	3.1%	3.6%
	3 days	15.6%	5.9%	6.3%	4.4%
	4 days	3.8%	1.5%	2.1%	0.73%
	5 days	79.7%	82.2%	77.1%	90.5%
Learner Questionnaire					
Does your teacher treat boys and girls in the same way in mathematics class?	No	29.2%	19.3%	7.9%	16.4%
	Yes	70.8%	80.7%	92.1%	83.6%
Teacher ignores girls.	Selected	33.0%	27.7%	43.8%	63.5%
Teacher ignores boys.	Selected	33.8%	22.0%	36.5%	66.1%
Teacher scolds girls.	Selected	3.9%	2.1%	2.7%	8.6%
Teacher scolds boys.	Selected	13.3%	8.5%	2.1%	4.8%
Teacher asks harder questions to girls.	Selected	1.8%	3.6%	6.2%	1.4%
Teacher asks harder questions to boys.	Selected	3.3%	1.3%	2.5%	1.2%
	No	31.0%	18.0%	20.2%	18.0%

	Answer Option	Baseline		Endline	
		Comparison	Treatment	Comparison	Treatment
Does your teacher treat learners with functional difficulties the same as other children?	Yes	69.0%	82.0%	79.8%	81.9%
Does your teacher provide extra support to learners who struggle with mathematics	No	29.2%	18.0%	27.6%	29.1%
	Yes	70.8%	82.0%	72.4%	70.9%

Table 36: Teachers' responses regarding the teacher guide in treatment group

	Answer Option	Baseline	Endline
Teacher Responses			
The NNP teacher guide is easy to use. The NNP teacher guide is easy to use.	Strongly disagree	1.2%	1.5%
	Disagree	14.3%	14.2%
	Agree	65.5%	54.4%
	Strongly agree	19.1%	29.9%
The NNP teacher guide is easy to use.	Strongly disagree	1.1%	0.0%
	Disagree	6.5%	4.5%
	Agree	66.3%	58.2%
	Strongly agree	26.1%	37.3%
I have many questions that the teacher guide does NOT address.	Strongly disagree	2.4%	12.7%
	Disagree	45.2%	42.5%
	Agree	47.2%	38.8%
	Strongly agree	5.2%	6.0%
The teacher guide provides sufficient guidance on how to implement the three parts of the lesson routine (teacher-led activity, independent learner activity, and reflection).	Strongly disagree	0.0%	2.2%
	Disagree	9.9%	6.7%
	Agree	61.7%	43.3%
	Strongly agree	28.4%	47.8%

Table 37: Teachers' responses regarding referencing learner workbooks in treatment group

	Answer Option	Baseline	Endline
Teacher Responses			
How often do you review or reference the learner workbooks when preparing your mathematics lessons?	Never	0.9%	0%
	Once a week	7.1%	2.2%
	Every other day	4.4%	5.9%
	Every day	87.7%	91.8%
The teacher guide is a useful material.	Not Selected	11.0%	13.0%

	Answer Option	Baseline	Endline
	Selected	89.0%	86.9%
The teacher guide is the most useful material.	Not Selected	65.8%	57.9%
	Selected	34.2%	42.1%
Classroom Observation			
Teacher has a lesson plan.	No	2.3%	4.3%
	Yes	97.7%	95.6%

Table 38: Teachers' responses regarding video content in treatment group

	Answer Option	Baseline	Endline
Teacher Responses			
The NNP training videos are very helpful in showing me what to do.	Strongly disagree	0.0%	0.7%
	Disagree	12.3%	4.5%
	Agree	61.5%	58.3%
	Strongly agree	26.2%	36.2%
The NNP training videos cover all the issues I need support with.	Strongly disagree	2.2%	5.2%
	Disagree	57.7%	45.8%
	Agree	34.8%	36.8%
	Strongly agree	5.2%	12.1%
The NNP training videos are too short to be helpful	Strongly disagree	1.5%	1.5%
	Disagree	36.1%	28.6%
	Agree	50.7%	54.9%
	Strongly agree	11.8%	15.1%
The NNP training videos contradict the information I have received from the NNP coaches or trainers.	Strongly disagree	16.9%	18.0%
	Disagree	67.5%	66.2%
	Agree	14.5%	10.5%
	Strongly agree	1.1%	5.3%

School-to-School International

1625 Palmetto Avenue, Suite A

Pacifica, CA 94044

STS-International.org