

Malawi National Numeracy Programme Evaluation

Endline Report

Principal Investigators: ***Lontum Nchadze****

Alperen Açıkol*
Evaewero French*
Esme Kadzamira[‡]
Todd Pugatch^
Lizzie Chiwaula[‡]
Elizabeth Meke[‡]

Additional contributors: ***Chikondi Sepula[‡]***

March 14, 2025

*** Oregon State University**

[‡] Center for Educational Research and Training, (CERT), University of Malawi

^ University at Buffalo

We thank our partners at Cambridge Education; Prevail Fund; the UK Foreign, Commonwealth, and Development Office (FCDO), and the Malawi Ministry of Education. Particular thanks are due to Nancy Chidzankufa, Rory Fenton, Ravinder Gera, Charlie Gordon, Stuart Johnson, Leanne Ketterlin-Gellar, Kyla Longman, Jennifer Swift-Morgan, Vanessa Tamo, Joshua Wakabi, and Arianna Zanolini.

Contents

1. Executive Summary	1
1.1. Key Findings	1
1.2. Recommendations	2
1.3. Conclusion	3
2. Introduction	4
2.1. Background on numeracy education and the situation in Malawi	4
2.2. Overview of the National Numeracy Programme	5
2.3. Timeline and next steps in the intervention	8
2.3.1. The first phase - Inception phase (till end 2021)	8
2.3.2. The second Phase - Pilot phase (2022 school year)	8
2.3.3. The third phase - Extended pilot phase (2023/24 school year)	8
2.3.4. Next steps of the intervention (2024/2025)	8
2.4. Research questions and objectives of the evaluation	9
2.5. Data collection	10
2.5.1. Field report	10
3. Methodology	12
3.1. Research Design	12
3.2. Sampling	12

3.3.	Data Collection	14
3.4.	Estimation Strategy	15
3.4.1.	Attrition	16
3.4.2.	Adjustments for Multiple Hypothesis Tests.....	17
3.5.	Outcomes	18
3.5.1.	Primary Outcomes.....	18
3.5.2.	Secondary Outcomes.....	22
3.6.	Attrition	23
	Table 1: Attrition by Treatment Status and Baseline Characteristics	23
4.	Results	25
4.1.	Numeracy under the status quo.....	25
4.2.	Numeracy: main results	27
4.2.1.	Numeracy: Results by Overall EGMA Score.....	27
	Table 2: Main Effects on numeracy, full sample	27
	Figure 1: EGMA Overall Score	30
	Figure 2: EGMA Overall Score Standard 1	31
	Figure 3: EGMA Overall Score Standard 2	32
	Figure 4: EGMA Overall Score Standard 3	33
	Figure 5: EGMA Overall Score Standard 4	34
	Table 3: Main Effects on numeracy (By Grade)	35

4.2.2.	Numeracy: Results by Sub Scores.....	36
	Table 4: Main Effects by Sub Scores	36
	Table 5: Main Effects by Sub Scores (Standard 1)	36
	Table 6: Main Effects by Sub Scores (Standard 2)	37
	Table 7: Main Effects by Sub Scores (Standard 3)	38
	Table 8: Main Effects by Sub Scores (Standard 4)	39
4.2.3.	Numeracy: Proficiency Analysis.....	40
	Table 9: Mean Proficiency Rates by Treatment Status, full sample	40
	Table 10: EGMA Proficiency Rates by Standards	41
	Table 11: Effects on Zero Scores and Proficiency	43
4.2.4.	Numeracy: Transforming Numeracy Gains into Schooling Impact (EYOS and LAYS).....	44
	Table 12: NNP Learning Gains	45
	Figure 6: Equivalent Years of Schooling	46
4.3.	Numeracy: heterogeneous treatment effects.....	47
	Table 13: Main Effects on numeracy by Gender	47
	Table 14: Main Effects on Numeracy by Baseline EGMA Tercile	48
	Table 15: Results by Baseline and Endline Attendance Level.....	49
	Table 16: Results by Student-Teacher Ratio Category	50
	Table 17: Main Effects on numeracy by Poverty Tercile	51
	Table 18: Results by School Feeding Status	51

4.4.	Programme participation and teaching quality.....	52
	Table 19: Treatment Effects on participation in teacher training	52
	Table 20: Classroom Observation Student Attendance Rate	53
	Table 21: Compliance with NNP Classroom Practices: Treatment schools only	54
	Table 22: Workbook Progress: Treatment Schools Only	55
	Table 23: Treatment Effects on Basic Implementation	55
	Table 24: Teaching Quality.....	56
4.5.	Attitudes towards mathematics	56
	Table 25: Treatment effects on student attitudes towards the learning of mathematics	57
	Table 26: Treatment effects on student attitudes (standard 1) towards the learning of mathematics	57
	Table 27: Treatment effects on student attitudes (standard 2) towards the learning of mathematics	58
	Table 28: Treatment effects on student attitudes (standard 3) towards the learning of mathematics	58
	Table 29: Treatment effects on student attitudes (standard 4) towards the learning of mathematics	58
	Table 30: Treatment Effects on Head Teacher Attitudes towards Mathematics	59
	Table 31: Treatment Effects on teacher attitudes towards the learning of mathematics.....	61
5.	Discussion	61
6.	Recommendations and Conclusion	63
7.	Conflict of Interest Statement	66
8.	References	66
9.	Appendix.....	69

9.1. Examples of mathematics curriculum materials	69
9.1.1. Control group mathematics curriculum	69
9.1.2. National Numeracy Programme curriculum	73
9.2. EGMA Proficiency Benchmarks	76
Appendix 9.3: A Note on Score Construction.....	77

List of Figures

1. Figure 1: EGMA Overall Score
2. Figure 2: EGMA Overall Score Standard 1
3. Figure 3: EGMA Overall Score Standard 2
4. Figure 4: EGMA Overall Score Standard 3
5. Figure 5: EGMA Overall Score Standard 4
6. Figure 6: Equivalent Years of Schooling
7. Figure 7: Cover page of Mathematics Learner textbook for standard 2 (Chichewa language)
8. Figure 8: Standard 2, Page 24: Mathematics Exercise on Addition of figures up to 20
9. Figure 9: Standard 2 Mathematics exercise on subtraction for figures up to 20
10. Figure 10: Standard 2 Mathematics Exercise on Addition of numbers up to 20
11. Figure 11: Standard 2 Mathematics Learner workbook for term 2: cover page (English language)
12. Figure 12: Standard 2 Mathematics Learner workbook for term 2, Page 2: Number Operations and Relationships
13. Figure 13: Standard 2 Mathematics Learner workbook for term 2, Page 4: Number Operations and Relationships

List of Tables

1. **Table 1:** Attrition by Treatment Status and Baseline Characteristics
2. **Table 2:** Main Effects on numeracy, full sample
3. **Table 3:** Main Effects on numeracy (By Grade)
4. **Table 4:** Main Effects by Sub Scores
5. **Table 5:** Main Effects by Sub Scores (Standard 1)
6. **Table 6:** Main Effects by Sub Scores (Standard 2)
7. **Table 7:** Main Effects by Sub Scores (Standard 3)
8. **Table 8:** Main Effects by Sub Scores (Standard 4)
9. **Table 9:** Mean Proficiency Rates by Treatment Status, full sample
10. **Table 10:** EGMA Proficiency Rates by Standards
11. **Table 11:** Effects on Zero Scores and Proficiency
12. **Table 12:** NNP Learning Gains
13. **Table 13:** Main Effects on numeracy by Gender
14. **Table 14:** Main Effects on Numeracy by Baseline EGMA Tercile
15. **Table 15:** Results by Baseline and Endline Attendance Level
16. **Table 16:** Results by Student-Teacher Ratio Category
17. **Table 17:** Main Effects on numeracy by Poverty Tercile
18. **Table 18:** Results by School Feeding Status

19. **Table 19:** Treatment Effects on take-up of teacher training
20. **Table 20:** Classroom Observation Student Attendance Rate
21. **Table 21:** Compliance with NNP Classroom Practices: Treatment schools only
22. **Table 22:** Workbook Progress: Treatment Schools Only
23. **Table 23:** Treatment Effects on Basic Implementation
24. **Table 24:** Assessing Teaching Quality
25. **Table 25:** Treatment effects on student attitudes towards the learning of mathematics
26. **Table 26:** Treatment effects on student attitudes (standard 1) towards the learning of mathematics
27. **Table 27:** Treatment effects on student attitudes (standard 2) towards the learning of mathematics
28. **Table 28:** Treatment effects on student attitudes (standard 3) towards the learning of mathematics
29. **Table 29:** Treatment effects on student attitudes (standard 4) towards the learning of mathematics
30. **Table 30:** Treatment Effects on Head Teacher Attitudes towards Mathematics
31. **Table 31:** Treatment Effects on teacher attitudes towards the learning of mathematics
32. **Table 32:** EGMA Benchmarks for this study

1. Executive Summary

The National Numeracy Programme (NNP) is an ambitious initiative launched by the Ministry of Education in Malawi, with technical support from Cambridge Education and funding from the UK Foreign, Commonwealth, and Development Office (FCDO). Its primary goal is to improve Mathematics learning outcomes and teaching quality in Standards 1 to 4, equipping children with foundational numeracy skills essential for academic and personal development. The programme incorporates innovative teaching materials, professional development for teachers, and student-focused pedagogical approaches to foster problem-solving and reasoning skills.

This report presents findings from a cluster randomized controlled trial (RCT) evaluating the NNP's extended pilot phase across 150 schools in three districts. The evaluation's objectives were to assess the programme's impact on numeracy outcomes and teaching quality using a comprehensive set of instruments, including the Early Grade Mathematics Assessments (EGMA), classroom observations, and teacher, headteacher, student, and parent surveys. The evaluation coincides with a dramatic scale-up of the programme to more than 1,000 schools, en route to reaching full national scale for the 2024-2025 school year. Evaluation results are intended to inform this effort.

1.1. Key Findings

The evaluation demonstrates significant improvements in both Mathematics learning outcomes and teaching quality. Students in treatment schools achieved a 0.19 standard deviation (SD) improvement in overall numeracy, equivalent to 0.56 Malawi Equivalent Years of Schooling (EYOS) and 0.25 Learning Adjusted Years of Schooling (LAYS). These gains reflect meaningful progress toward narrowing the learning gap between Malawian students and their regional peers.

The programme also significantly improved teaching quality, with a 0.12 increase in a composite index of instructional practices measured on a 0-1 scale, or an 18 percent increase over the control group mean. Programme participation was high among teachers, as was with compliance with NNP teaching practices in the classroom. The programme meaningfully changed the mathematics curriculum in treatment schools. Additionally, the programme significantly improved head teachers' attitudes toward mathematics

teaching and learning. However, some aspects of the curriculum were implemented with less fidelity. For instance, audits of student programme workbooks revealed a high proportion of empty pages.

The analysis revealed heterogeneity in programme effects. Numeracy gains were stronger for students in Standards 3 and 4. Students in the bottom third of the performance distribution and from the poorest households benefited less. We fail to find evidence that high pupil-teacher ratios and low student attendance explain these findings, though data limitations prevent us from ruling out these explanations. The findings suggest that the programme disproportionately helped students with stronger baseline skills, emphasizing the need for strategies to support the most disadvantaged learners.

1.2. Recommendations

Given the demonstrated success of the intervention in improving learning outcomes and teaching quality, we recommend continuing and strengthening the nation-wide scale-up. The following recommendations aim to ensure the scale-up is effectively implemented and its impact maximized:

1. **Enhance Implementation Fidelity:** Although implementation fidelity was high, gaps in compliance, such as the appropriate use of manipulatives (69.9%) and learner workbook coverage, should be addressed. Strengthened teacher training and continuous monitoring are key to improving these areas.
2. **Support Marginalized Groups:** Strategies should focus on students in the lowest performance and poverty tiers. Complementary interventions, such as school feeding, could help mitigate barriers to learning.
3. **Leverage Scaling Experience:** The programme's progression from pilot to extended pilot phases across 1,200 schools offers valuable lessons for the upcoming national scale-up. This experience can help mitigate common challenges associated with scaling educational interventions.
4. **Continue to evaluate and iterate.** Scaling and sustaining a successful programme is challenging. We recommend continued evaluation, iteration, and testing (including impact evaluation, A/B testing of programme elements, and process evaluation) as the programme evolves.

We are agnostic on how each of these recommendations may be carried out. We encourage stakeholders to probe these and other relevant issues further to determine the best courses of action for the programme's future.

1.3. Conclusion

The National Numeracy Programme has made significant strides in improving Mathematics learning outcomes and teaching quality in Malawi. These results highlight the programme's potential to address foundational learning gaps and align Malawian students' numeracy skills with regional and global standards. However, the challenges identified, particularly in supporting the most disadvantaged learners and maintaining fidelity at scale, must be addressed to maximize its impact. By building on the Ministry of Education's scaling experience and fostering continued collaboration among stakeholders, the NNP can serve as a model for transformative education reform in Malawi.

2. Introduction

2.1. Background on numeracy education and the situation in Malawi

Foundational literacy and numeracy skills are essential building blocks for education worldwide, and Malawi is no exception. However, the global context reveals a sobering reality: despite substantial efforts to expand access to education, the proficiency levels of children in these fundamental domains remain distressingly deficient. This is glaringly evident in low- and middle-income countries, where less than half of all children are proficient readers by the age of 10 (UNESCO, 2017). The state of numeracy skills is even more disheartening, with a mere 18% of children in half of the countries recently analysed by UNICEF possessing foundational numeracy skills (UNICEF, 2022). Within this disconcerting global landscape, the situation in Malawi is particularly challenging. Existing evidence underscores the unsettling disparities in numeracy competencies between Malawian children and their peers in the region (Brombacher, 2011, 2019, UNICEF, 2021), with Malawian children found to consistently under-perform their regional counterparts in large-scale international assessments like the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SEACMEQ). This alarming performance deficit in numeracy assessments not only undermines the immediate educational prospects of Malawi's youth but also jeopardizes the nation's long-term development aspirations.

It is against this background that the National Numeracy Programme (NNP) was established to improve mathematics outcomes for Malawian students in Standards 1 to 4, ensuring that both girls and boys build a strong foundation in basic skills for future academic success and personal development. Implemented by the Ministry of Education with technical support from Cambridge Education and funding from the UK government, the programme has evolved through multiple phases. The pilot phase in 2022 introduced new teaching materials and training methods across 200 schools. A non-experimental evaluation of this phase demonstrated a statistically significant improvement in learning outcomes for Standards 3 and 4.

The extended pilot phase, which focused on refining its approach for long-term sustainability began in 2023 and concluded at the end of the school year, in 2024. Oregon State University (OSU), in collaboration with CERT, conducted an impact evaluation of the extended pilot through a cluster randomized controlled trial (RCT) to measure its effectiveness. The evaluation was conducted in two phases: baseline, completed in September 2023, and endline, completed in July 2024. This report presents the findings from the endline survey, which included mathematics assessments of learners in Standards 1 to 4, along with insights from teachers, headteachers and observations of mathematics lessons in Standards 1 to 4.

2.2. Overview of the National Numeracy Programme

Mathematics education in Malawi begins at the preschool level and continues through primary, secondary, and university levels. It is a compulsory subject in both primary and secondary education. Since 2014, the language of instruction across all levels has officially been English, as stipulated in the Education Act of 2013. However, in practice, Chichewa (the national language) is still predominantly used as the medium of instruction in Standards 1–4, primarily because textbooks and teacher guides for these grades remain in Chichewa, which was the previous language of instruction. For the National Numeracy Programme (NNP), however, learner workbooks and teacher guides are provided in English.

Mathematics has been taught as a subject in primary schools since the introduction of formal education in Malawi. However, over the years, with curriculum changes, the subject was renamed from Arithmetic to Mathematics to align with the curriculum changes. Malawi Institute of Education, (1991 cited in Kazima et al, 2022) explains that with Arithmetic, the subject put emphasis on learners' acquisition of number knowledge and computational skills with speed and accuracy but after the 1991 curriculum review, the subject focus was to develop learners' mathematical capabilities that would help them solve problems in their everyday lives, hence, the name Mathematics.

Over the years, various interventions have been implemented in Malawi to enhance the teaching and learning of mathematics. Kazima et al (2022) highlighted a number of these interventions as follows:

- Numeracy Boost Intervention
 - a. Implemented by Save the Children in 2012 in selected rural schools within one district.
 - b. Key components included:
 - i. Establishing numeracy camps in rural communities.
 - ii. Building the capacity of camp facilitators and teacher camp supervisors.
 - iii. Conducting learner assessments to monitor progress.
- Unlocking Talent Project
 - a. Funded by the Norwegian Embassy and implemented in collaboration with the Voluntary Services Overseas (VSO), the Ministry of Education, One Billion, DFID, UNICEF, and the Scottish Government.
 - b. Launched in 2018, this project focused on tablet-based learning to develop basic numeracy skills.
 - c. It was implemented in over 100 schools nationwide.

- Improving the Quality and Capacity of Mathematics Teacher Education in Malawi
 - a. A collaborative initiative between the University of Malawi and the University of Stavanger, Norway, implemented from 2014 to 2018.
 - b. The project aimed to enhance the capacity of mathematics teacher education across all public primary teacher education colleges.
- Strengthening Numeracy in the Early Years of Primary Education
 - a. Also implemented by the University of Malawi in collaboration with the University of Stavanger, Norway, this initiative focused on the professional development of teachers for Standards 1–4.
 - b. Activities included:
 - i. Conducting workshops to improve teaching methods.
 - ii. Facilitating lesson studies to promote collaborative learning among teachers.
- JICA Numeracy Project
 - a. Launched in 2016, this initiative aimed to help learners transition from unit counting to more advanced concepts like composition and decomposition.
 - b. The project targeted Standards 1–3 teachers in two pilot schools, focusing on building foundational numeracy skills.

The NNP sought to usher in a transformative vision for mathematics education in Malawi that transcends the traditional paradigm of rote learning; instead, it endeavoured to help children understand the mathematics they learn, develop problem-solving and reasoning skills, and apply the mathematics they learn in real life. This holistic approach holds the promise of nurturing a generation of Malawian learners who are not just literate in mathematics but possess the capacity to wield it as a formidable tool for life's multifaceted challenges.

The programme used different approaches to achieve its objectives. These included:

- Revision of the Mathematics curriculum for standards 1 to 4
- Training teachers on new teaching strategies through workshops, teacher learning circles, coaching, and instructional videos among others
- Development and distribution of teaching and learning materials aligned to the revised curriculum

- Piloting and refining the new teaching and learning materials.¹

Most importantly, the NNP curriculum content differed from the previous approach in that it focused on problem-solving, the application of mathematics skills in real life and being able to apply mathematics learnt to solve unfamiliar problems, and the use of different strategies or procedures to solve mathematics accompanied by the reasoning behind the strategies/procedures used. In addition, there are specific child-led activities and teacher-led activities for each topic covered with the aim of making the teaching and learning of mathematics very engaging. Learners are also introduced to new strategies and topics including the use of number chains, pyramids, flow diagrams, number lines, bubbles and tables in solving mathematical problems in early grades (standards 1 to 4) for the first time in Malawi and new topics such as data handling among others (NNP facilitator Manual, 2023/24).

Furthermore, there is a reflection session in each Mathematics lesson, where learners are asked to clarify how they came up with solutions to the mathematics problems or are asked to respond to reflection questions based on the work done to enforce reasoning skills. Reflection sessions also help teachers to assess whether learners understand the concepts and offer the required support to those struggling.

Learner workbooks are key to the implementation of this programme and each learner receives his/her workbook to use during Mathematics lessons. The workbooks are specific to each term's work. Hence, learners receive a workbook at the beginning of every term. Notably, the Mathematics learner textbooks for the old approach are in the local language (Chichewa) while the NNP workbooks for learners are written in English.

Annex 9.1 demonstrates the differences between the traditional mathematics curriculum and the National Numeracy Programme curriculum. Figure 8 to Figure 10 show pages from the textbook of the traditional curriculum. Each page lists a series of arithmetic operations to solve, without context. Figure 12 to Figure 13 show pages from the NNP workbook. Most of the NNP exercises provide visual representations of the problem. In some cases, the images are drawn from the local context, such as counting groups of cassava. Word problems require learners to apply arithmetic to hypothetical situations. Finally, there is a reflection question for learners to explain how they got the answer ("how did you get this?").

¹<https://nnpmalawi.org/>

2.3. Timeline and next steps in the intervention

The programme has undergone several phases as follows:

2.3.1. The first phase - Inception phase (till end 2021)

This involved the development of the vision for the teaching and learning of Mathematics in Malawi; the development of materials that included learner workbooks, teachers' guides, a revised curriculum for standards 1 to 4 and teacher training methodology for piloting.

2.3.2. The second Phase - Pilot phase (2022 school year)

The pilot phase involved the piloting of both the materials, teaching and learning approaches as well as teacher training approaches in 200 schools across the country and refinement of materials based on the evaluation findings of the pilot implementation (School-To-School, 2022).

2.3.3. The third phase - Extended pilot phase (2023/24 school year)

The extended pilot phase ran during the 2023/24 school year in over 1,000 schools. During this phase, updated materials and teacher training materials were piloted. More specifically, it included the following:

- improving and adapting the programme to be more sustainable at scale. The key changes included the introduction of large print workbooks for learners with low vision, and the alignment of teacher guides to the learner workbooks.
- emphasized school-based training and zone-level training.
- strengthened learner workbooks and teacher guides for standards 1 – 4 and significantly revised standards 1 and 2 workbooks following evidence and concerns raised by stakeholders and external reviewers.
- incorporated Teacher Learning Circles (TLCs) in Teacher Continuous Professional Development (TCPD) to foster effective classroom instruction and peer reflection aligned with the NNP curriculum.

2.3.4. Next steps of the intervention (2024/2025)

In the 2024/25 academic year, the intervention was scaled up to encompass Standards 1 and 2 in all public schools in Malawi, with Standards 3 and 4 scheduled for inclusion in the 2025/26 academic year. In October 2024, the initiative was renamed the

National Mathematics Curriculum Reform Programme, marking its transition to a government-led reform implemented in every primary school in Malawi.

Implementation of the intervention is accompanied by in-service training and ongoing school and classroom support to all teachers in standards 1 to 4. There is also ongoing monitoring of the programme across the country.

2.4. Research questions and objectives of the evaluation

In light of the profound importance of the NNP, a rigorous impact evaluation of the extended pilot was employed to assess the programme's effectiveness in advancing its key objectives.

The evaluation's main objective was to assess the causal impact of the National Numeracy Programme on Mathematics learning outcomes and mathematics teaching quality in Malawi. Specifically, the evaluation aimed to respond to the following research questions:

1. What impact did the NNP have on the mathematics learning outcomes of Standard 1 to 4 learners in the expanded pilot schools?
2. What impact did the NNP have on the quality of mathematics instruction provided by Standard 1 to 4 teachers in the expanded pilot schools?

To address these questions, the evaluation adopted a cluster randomized controlled trial (cRCT) design. This design entailed the random allocation of 75 treatment schools and 75 control schools across three districts in Malawi: Mzimba South in the northern region, Mchinji in the Central region, and Chikwawa in the southern region.

The main purpose of conducting the impact evaluation using a cluster randomized controlled trial was to provide the Ministry of Education with robust evidence of the causal impact of the NNP interventions on mathematics learning outcomes and the quality of mathematics instruction provided by the teachers in standards 1 to 4. This was achieved by comparing changes in mathematics learning outcomes between the control and treatment schools at two points in time: baseline and endline.

2.5. Data collection

2.5.1. Field report²

Preparation for Training and Fieldwork

To prepare for fieldwork, the research team obtained the Learner Identification Number (LIN) from the Ministry of Education for students sampled during the baseline study. The LIN was used to identify and verify students during the endline follow-up accurately. Demographic details, including date of birth, age, sex, class, and parent/guardian contact information, collected at baseline, were pre-filled in the data collection instruments. The baseline tools—such as the Early Grade Mathematics Assessment (EGMA), student questionnaire, and head teacher questionnaire—were reviewed, with some questions modified, removed, or added as needed. Additionally, new instruments, including a teacher questionnaire and classroom observation tool, were developed. Details on the survey instruments appear in Section 3.3.

Training of Enumerators and Observers

63 enumerators and 12 observers participated in a five-day training in Zomba from June 17-21, 2024. Since the fieldwork required 60 enumerators, three additional enumerators were trained as reserves to replace any who would be unable to continue during data collection. The classroom observation tool, which involved making professional judgments on the quality of teaching and learning, led to the recruitment of 12 observers from the two Colleges of Education under the Ministry of Education to conduct the classroom observations.

On the first day, the observers trained alongside the enumerators before being divided into separate groups on the second day. On the first day of training, participants were introduced to the NNP evaluation survey and its tools, followed by a thorough review of research ethics, safeguarding, and child protection guidelines to be adhered to during the survey. After this session, all participants signed the training, data collection, and confidentiality agreements. Later, they were guided through the data collection applications, Tangerine and CSPRO, beginning with tablet basics for those with limited or no prior experience.

Training of enumerators

² This section reports on fieldwork for the endline. For details on baseline fieldwork, see the baseline report.

Enumerators received detailed training on administering and scoring the EGMA tool on the second day, including practical sessions and an Assessor Accuracy Measure (AAM) quiz to gauge preparedness. Selection for the main survey was based on quiz results, EGMA theory tests, and practice interviews. Day three focused on training for the student and head teacher questionnaires, as well as survey logistics. On day four, participants visited six schools in Zomba Rural for practice, followed by a debriefing session the next day. The training and field practice helped refine the instruments and prepared enumerators for data collection.

Training of Observers

Observer training was facilitated by a team of three officials who have been involved in the development of NNP materials and/or were NNP master trainers and were assisted by CERT staff. On day two, observers were introduced to NNP lessons, teaching materials, and the classroom observation tool, with a lesson demonstration by the master trainer. Practice sessions followed using tablets. Day three involved field practice in local schools, followed by a debriefing. On the final day, observers participated in another lesson demonstration and were introduced to the revised scoring rubric and teacher questionnaire, followed by practical interviews and logistics training.

Data Collection

Fieldwork was carried out by a team of 72 members comprising 48 enumerators, 12 supervisors, and 12 observers. Half of the team were female. They were organised into twelve teams of 6 members each (4 teams per district) and collected the data from 24th June to 10th July 2024. Each team consisted of 1 Supervisor, 4 enumerators and 1 observer. On average each team visited 12.5 schools.

During the first week, 3 PIs, each covering a district, supported the teams. Apart from overall supervision, the PIs worked with the supervisors to establish an efficient way of tracking students who were absent on the day of data collection. The daily tracking form was updated to provide detailed information³ about absent students, which helped to identify students who could be easily followed up for callbacks.

³ The six categories were: 1. Absent, 2. Dropped out, 3. Transferred within district, 4. Transferred to another district, 5. Migrated to neighbouring country, 6. Whereabouts not traced. Students in categories 1 to 3 were followed up in their villages.

During field visits, enumerators conducted EGMA assessments and administered student and head teacher questionnaires. Observers conducted lesson observations and teacher interviews. Supervisors managed logistics, including contacting schools and education officials and securing permission to follow up with students at home when necessary.

Overall, the endline survey was successful as we managed to trace and assess the vast majority of students assessed at baseline. During the baseline survey, 9,400 students were assessed and interviewed. At the endline survey, 8,320 students were traced and successfully assessed using the EGMA instrument. This represents a slightly over 88% response rate. We managed to interview 149 out of the 150 head teachers targeted, 542 out of 600 teachers in standards 1- 4, and observed 290 out of the 3000 possible classroom observations.

3. Methodology

3.1. Research Design

This evaluation employs a cluster randomized controlled trial (RCT), with schools as the unit of treatment in a matched pair design. We (the researchers) formed pairs from an evaluation sample of 150 schools, matched by district and baseline school characteristics. We randomly assigned schools in each pair to treatment or control. There were 75 treatment schools and 75 control schools across three districts in Malawi: Mzimba South in the North region, Mchinji in the Central region, and Chikwawa in the South region. Treatment schools received the NNP curriculum, while control schools received the status quo mathematics curriculum. The intervention targeted students enrolled in Standards (grades) 1-4.

3.2. Sampling

The sampling strategy for this evaluation involved a multi-phase process that involved district, school, and student/teacher levels. In consultation with CERT, we selected three districts not exposed to the NNP extended pilot: Mzimba South in the North region, Mchinji in the Central region, and Chikwawa in the South region. Among the districts not previously exposed to NNP, these three districts were chosen to be representative of each region. Within each selected district, we defined eligible schools as schools outside zones in the NNP pilot and outside the sampling frame of potential confounding activities. Specifically, we excluded schools in the sampling frame of Building Education Foundations through Innovation & Technology (BEFIT), a concurrent education technology

intervention. We also excluded schools in the Malawi Longitudinal School Survey (MLSS) scheduled for 2023 data collection due to calendar overlap with our baseline survey. We then drew a sample of 298 eligible schools (100 schools in each of Mzimba South and Mchinji, plus all 98 eligible schools in Chikwawa), stratified by zone.

From this sampling frame of 298 schools, we grouped schools within each district into sets of four based on baseline characteristics. We defined groups using the first principal component of the following school attributes from the 2021 Education Management Information System (EMIS):

- Age of the school
- Enrolment
- Pupil-teacher ratio (PTR)
- enrollment percentage
- Female teacher percentage
- Remoteness categories (from Asim et al., 2019)
- Share of schools with Early Childhood Development (ECD) exposure
- Share of repeaters in S1-S4
- Pass rate in the Primary School Leaving Certificate Examination (PSLCE)
- Management index, calculated as the proportion of the following reported as active at the school: Parent-Teacher Association (PTA); School Management Committee (SMC); Community members; and School Improvement plans.

We used these criteria to group schools into sets of four. Within each set, we randomly assigned two schools to the treatment group and two to the control group, following the recommendation by Athey and Imbens (2017). We randomly selected one school from each treatment-control pair for inclusion in the sample, and the other served as a replacement, in case of need. Within the selected schools, eligibility extended to all students in standards 1 to 4, the parents or guardians of these students, teachers responsible for these standards, and the school head teachers.

Schools assigned to treatment were intended to receive the full programme for the first time as part of this study. Treatment schools would not have received the programme during the evaluation period otherwise. Schools assigned to control were intended to not participate in the programme during the evaluation period, nor had they participated previously. The implementing partners were charged with carrying out these assignments for the study.

The research design and sample size were informed by power calculations conducted before baseline data collection. We later updated the power calculations based on actual sample sizes and intra-cluster correlations of numeracy scores from the baseline data. The sample has 80% power for a minimum detectable effect size of 0.15 standard deviations (sd) in the full sample, or 0.16 sd when disaggregated by gender. Detailed power calculations appear in the baseline report.

3.3. Data Collection

We collected two rounds of primary data: baseline (October 2023) and endline (June-July 2024), corresponding to the beginning and end of the academic year. However, the intervention schedule required teachers in treatment schools to receive training and begin the curriculum in September 2023, before the baseline survey could reach the field. Treatment schools were exposed to the NNP curriculum for 3-5 weeks before baseline data were collected.⁴ Nonetheless, the timing of the baseline does not invalidate the analysis, which is based on experimental treatment assignment and endline data, not the presence of a unblemished baseline.⁵ We address implications of this timing for the evaluation methodology in Section 3.4.

For the endline data collection, we used the following instruments:

- Student learning assessments: Early Grade Mathematics Assessment (EGMA).
- Student questionnaire: A concise demographic questionnaire was administered concurrently with EGMA.
- Head teacher survey: Focused on school characteristics to complement administrative data.
- Classroom observation: Designed to observe the quality of teaching and learning and the implementation of the curriculum.

⁴ Schools opened the week of September 11, 2023. Training for teachers in treated schools took place during that week. NNP materials also reached schools by the end of the first week (prior to school opening in Chikwawa and Mchinji, and during the first week for Mzimba South). Baseline data collection occurred from October 2-18, 2023.

⁵ A potential concern is if the school composition changed in response to treatment assignment before baseline data were collected. For instance, students or teachers might transfer to NNP schools to take advantage of the treatment. We consider these scenarios unlikely. Schools learned their treatment assignments shortly before the beginning of the academic year, leaving little time for such transfers. We heard no reports of this activity from implementers or enumerators in the field.

The complete set of instruments appears in the Appendix. CERT trained data collectors in the data collection techniques to ensure data quality and minimize errors, as well as ethical considerations. CERT collected raw data and securely stored it in a centralized data repository, with access restricted to authorized personnel only. CERT also removed personally identifiable information (PII) from all data shared with researchers not covered by ethical approval.

3.4. Estimation Strategy

Our analysis follows a plan submitted to the AEA RCT Registry (AEARCTR-0011964), with all analysis outside this plan labeled as exploratory. To analyze endline results, we compare outcomes between treatment and control schools using the following equation, estimated by ordinary least squares (OLS):

$$y_{isg} = \alpha + \beta NNP_{sg} + \gamma_g + \varepsilon_{isg} \quad (1)$$

where i indexes students, s indexes schools, and g indexes sampling strata (i.e., the matched group of four schools used in the random assignment); y is the end-line outcome; NNP is an indicator variable for random assignment to the NNP programme; γ is a stratum (i.e., matched pair) fixed effect; and ε is the error term. This equation compares outcomes (such as EGMA scores) between schools assigned to the NNP and the control group, adjusting for the matched-pair study design. Our coefficient of interest is β , which measures the effect (intent to treat, or ITT) of assignment to NNP on the outcome. We cluster standard errors by strata, following recommendations for inference with small strata (de Chaisemartin and Ramirez-Cuellar, 2024).

Baseline data were collected after the intervention had begun, and therefore could represent nascent treatment effects. We therefore omit the baseline outcome from equation (1). This specification mitigates any bias due to baseline timing, as it does not control for variables which may have been affected by treatment status. However, omitting the baseline outcome may lead to less precise estimates, so we later include it as a robustness check.

For student learning outcomes, we estimate equation (1) for all students in the school and separately by grade. The effect of the programme may differ among students, hence we estimate treatment effect heterogeneity using observable characteristics. Using the overall EGMA score as the dependent variable, we estimate equation (1) for the following subsamples of students:

- Boys
- Girls
- Baseline EGMA tercile (bottom/middle/top third of distribution)
- Baseline poverty tercile (bottom/middle/top third of Poverty Likelihood Index)

For student learning outcomes, for robustness we also estimate a version of equation (1) with baseline outcomes. We also estimate a version of equation (1) including covariates for which we found statistically significant differences between treatment groups at baseline.

In our analysis, we do not impute missing values for any dependent variables. For covariates (e.g., baseline outcomes), we replace missing values with the control group mean and include a dummy for missing values in the regression (Haushofer and Shapiro 2015). Also, we omit analyzing any binary outcome (outcome with only two values, e.g., yes or no) for which the proportion in the control group exceeds 90 percent. In other words, if more than 90 percent of the control group said “yes” to an outcome we planned to analyze, we refrain from including this outcome, because there is little scope for an increase in the treatment group. This rule applies to stand-alone outcomes and components of an outcome index. For instance, if a binary outcome in an index has a control proportion greater than 90 percent, we omit that outcome from the index.

3.4.1. Attrition

We analyze attrition using the regressions:

$$A_{isg} = \alpha + \delta NNP_{sg} + \gamma_g + \varepsilon_{isg} \quad (2)$$

$$A_{isg} = \alpha + \delta_1 NNP_{sg} + \delta_2 (NNP_{sg} \times X_{0isg}) + \theta X_{0isg} + \gamma_g + \varepsilon_{isg} \quad (3)$$

where A is an indicator for attrition, X is a vector of baseline characteristics, and all else is as in equation (1). In equation (2), the coefficient δ measures any differential attrition rate between treatment and control. In equation (3), δ_2 measures any differences in observable characteristics between students who attrit in the treatment and control groups.⁶

The baseline student characteristics included in X are:

- Indicator for female
- Indicators for standard (grade level)
- Poverty Likelihood Index (PPI)⁷

3.4.2. Adjustments for Multiple Hypothesis Tests

We report statistical significance using conventional p-values. In the academic journal article we are preparing for this evaluation, we will also adjust inference for multiple hypothesis tests within each outcome family by reporting sharpened q-values (Benjamini, Krieger, and Yekutieli 2006). When the treatment effect for an outcome index is statistically significant after applying multiple hypothesis test adjustments, we estimate treatment effects separately for each component of the index.

⁶ According to our analysis plan, if i) attrition exceeded 20 percent, or ii) we rejected either of the null hypotheses $H_0: \delta = 0$ or $H_0: \delta_1 = \delta_2 = 0$ at the 5% level, then we would calculate Lee bounds on treatment effects for our main outcomes. Neither scenario occurred in practice.

⁷ The Poverty Likelihood Index (PPI), developed by Innovations for Poverty Action (IPA), calculates the likelihood of a household falling below a specified poverty threshold, which, in our study, corresponds to Malawi's national poverty line. To determine the PPI, we followed these steps: First, we incorporated 10 specific questions about household assets and characteristics into the baseline parent questionnaire. Each response to these questions is assigned a predetermined score derived from statistical analysis, which establishes associations between certain characteristics, asset ownership, and the likelihood of living below the poverty line. Next, we utilize IPA's scorecards to correlate these responses with their corresponding scores, aggregating them to generate PPI scores for each household. Finally, we match each PPI score with its corresponding category using a lookup table. With a statistically representative sample, the average PPI provides an estimation of the poverty rate within the population.

3.5. Outcomes

3.5.1. Primary Outcomes

The primary outcome for this research is student numeracy. We assess numeracy using the Early Grade Mathematics Assessment (EGMA). EGMA consists of eight sections: number identification, number discrimination, addition (two levels), subtraction (two levels), missing numbers, and word problems. The assessment is administered orally, allowing students with no or limited literacy to participate.

EGMA was developed by RTI International and adapted for this evaluation by CERT. We administered the same version of EGMA as at baseline, and the same version across all standards. We administered the test in Chichewa and Tumbuka, the two main local languages used in Malawian primary schools. Some sections are timed, allowing scores to be expressed in terms of accuracy (number of correct items) or fluency (correct items per minute). All other sections are untimed, with scores measured in number of correct items.

We construct two separate composite scores based on response processing, following the formulation outlined in Piper et al. (2016). The first, Procedural Fluency, aggregates all timed assessments and is measured as the number of correct items per minute. Procedural Fluency includes the tasks of number identification, addition level 1, and subtraction level 1. The second, Conceptual Understanding, encompasses untimed tasks and is measured as the percentage of correct responses. Conceptual Understanding includes the tasks of number discrimination, missing numbers, addition level 2, subtraction level 2, and word problems. Each composite score is standardized using the control group's mean and standard deviation within each grade level.

The overall EGMA score is computed as the mean of these two standardized composite scores. This approach implicitly assigns equal weight to fluency and conceptual understanding in assessing numeracy skills. However, it also places disproportionate emphasis on the subscores that make up the fluency score, as its composite score is derived from three subscores, whereas conceptual understanding is based on five.

For robustness, we construct and report results using an alternative scoring method that assigns equal weight to each subscore. A detailed discussion of these methodological choices and the psychometric properties of the scores is provided in Appendix 9.3.

Because the overall EGMA score is our primary outcome, we do not adjust for multiple hypothesis tests for this outcome. If we find significant treatment effects on this outcome, we will also estimate treatment effects on each subscore to identify potential drivers of the main treatment effect.

In addition to overall EGMA scores, we also measure numeracy using proficiency benchmarks. No universal set of EGMA proficiency benchmarks exists, nor are we aware of a numeracy benchmarking exercise for Malawi. We therefore define proficiency levels ourselves for purposes of this study, building on EGMA benchmarks established for other Sub-Saharan African countries (RTI International 2014a, 2014b, 2015). See Appendix 9.2 for details. In constructing these benchmarks, we prioritised skills that align with the NNP's goals of enhancing problem-solving and sense making. We also included addition and subtraction (level 1), as a basic operation on which we would expect proficiency in these standards. Therefore, for each skill (addition/subtraction level 1, missing numbers, and word problems), we estimate equation (1) using as dependent variables an indicator for a zero score and an indicator for proficiency. We calculate treatment effects for the full sample and separately by standard (grade level). For proficiency benchmarks, we will adjust inference for multiple hypothesis tests.

The table below lists primary and secondary outcomes included in our analysis plan.

Table: Outcomes included in analysis plan

Outcome family	Outcome	Unit of analysis	Description	Disaggregate by standard (grade level)?
Numeracy	Overall*	Student	EGMA, average z-score of timed and untimed composite scores	yes
Numeracy	Addition/subtraction Level 1	Student	EGMA, indicator for scoring zero	yes
Numeracy	Addition/subtraction Level 1	Student	EGMA, indicator for proficiency	yes
Numeracy	Missing numbers	Student	EGMA, indicator for scoring zero	yes
Numeracy	Missing numbers	Student	EGMA, indicator for proficiency	yes
Numeracy	Word Problems	Student	EGMA, indicator for scoring zero	Yes

Numeracy	Word Problems	Student	EGMA, indicator for proficiency	Yes
Take-up	Teacher training	Teacher	Mean of indicators for: - received in-person training in mathematics instruction at least once in past academic year - participated in mathematics teacher learning circles at least three times	No
Take-up	Basic implementation	Head teacher	Mean of indicators for: - learners received new mathematics workbooks this academic year - workbooks made available every term - enough workbooks for all learners - Strongly agree: Teachers had the materials they needed to teach mathematics effectively. - Strongly agree: Teachers had ongoing training and support to improve their mathematics teaching skills. - Strongly agree: Teachers were able to implement the mathematics curriculum well.	No
Take-up	Curriculum implementation	Classroom observation	Mean of: - Proportion of students present with workbook - Indicator for lesson plan informed by learner workbook or teacher guide - Indicator for lesson used workbook - Indicator for teacher using manipulatives that are appropriate for the page in the workbook - Indicator for teacher lead reflection on the tasks/activities with the learners - Indicator for teacher set tasks from the workbook that the learners must work on independently	No

Take-up	Curriculum implementation	Classroom observation	Index of the following: • Lesson page reached in learner workbook (units: proportion of completed workbook) • Ability to apply skills and strategies learned to solve problems in learner workbook (Units: 0 = “Progress is not satisfactory”, 0.5 = “Progress is satisfactory”, 1 = “Progress is effective”.) • Workbook pages with all exercises attempted (units: proportion of maximum value in sample)	No
Teaching quality	Teaching quality	Classroom observation	Mean of (scoring: 0 = not effective, 0.33 = satisfactory, 0.66 = partially effective, 1 = effective): • Teacher explicitly articulates the objectives of the lesson • Teacher helps learners to apply the skills and strategies they have learned in order to solve the problems in their learner workbook • Teacher relates classroom activities to the objectives • Teacher draws attention to the skills that link activities on a page and reminds learners of learning from previous lessons • Teacher asks learners to explain their answers and uses this to recognise when an incorrect answer is due to a simple calculation mistake or a misconception • Teacher asks questions that call for critical thinking • Teacher's explanation of content is clear • Children have the opportunity to practice at least twice in class • Teacher is always assessing the progress of learners and adjusts the teaching if he/she finds that the learners are not making progress • Teacher has a lesson plan and/or scheme of work that supports children's learning in the lesson • Quality of lesson plan implementation	No
Teacher attitudes	Teacher attitudes	teacher	Mean of indicators for: • Strongly agree: Mathematics education should place core emphasis on problem solving and sense making rather than procedures • Strongly agree: All students can learn and excel in mathematics regardless of their background	No
Teacher attitudes	Head teacher attitudes	Head teacher	Mean of indicators for:	no

			- Strongly agree: In my school this academic year, mathematics education placed core emphasis on problem solving and sense making rather than procedures - Strongly agree: In my school this academic year, teachers incorporated real-world problem-solving and practical applications into their mathematics lessons in standards 1 to 4 - Strongly agree: Learners made good progress in their understanding of mathematics. - Strongly agree: Learners enjoyed mathematics more than they did at the beginning of the academic year.	
Attitudes	Student Attitudes	Student	Mean of indicators for: - How much did you enjoy your mathematics class this school year? - Is Mathematics a useful subject? - Is Mathematics important for your future? - Is Mathematics hard? - Do you like explaining to your friends and family how you solved a math problem?	Yes

This table summarizes primary and secondary outcomes for analysis.

* Inference will not be adjusted for multiple hypothesis tests. All other outcomes will be adjusted for multiple hypotheses, within outcome family.

We also include some exploratory analysis not included in the analysis plan.

3.5.2. Secondary Outcomes

Secondary outcomes refer to outcomes hypothesized to be affected by the programme along the theory of change from treatment assignment to improved student numeracy. Our secondary outcomes include:

- Intervention take-up
- Curriculum implementation
- Teaching quality

- Student engagement
- Teacher attitudes

We measure each secondary outcome using an index. The components of each index appear in Table 1.

3.6. Attrition

Before presenting the main outcome results, we analyze sample attrition. Attrition refers to the case in which a student assessed at baseline is missing from the endline sample. Attrition presents a challenge to the evaluation because we do not know the endline numeracy of students who attrit. Attrition from the endline was 13 percent, meaning we succeeded in assessing almost nine of every ten students from the baseline.

Despite this relatively low overall attrition rate, attrition could still pose a challenge to results if attrition is asymmetric between the treatment and control groups. For instance, if the programme changed the quantity or composition of students participating in the endline relative to the control group, the endline sample may not be representative of the original baseline sample. We test for these differential patterns of attrition in Table 1. In column (1), we find no evidence that attrition differed significantly between the treatment and control groups. In column (2), we test if the programme changed the composition of the sample by interacting the treatment indicator with several student characteristics, including poverty likelihood, standard (grade level) dummies, and a dummy indicating if the student is a boy. None of these interaction terms is statistically significant, indicating no systematic differences in sample composition between treatment and control groups. We conclude that differential patterns of attrition are not a threat to the internal validity of the study.

Table 1: Attrition by Treatment Status and Baseline Characteristics

	<u>Outcome: attrition from endline</u>	
	(1)	(2)
Treatment	-0.02 (0.01)	-0.02 (0.02)

Treatment * Poverty Likelihood		0.00 (0.00)
Treatment * Grade (Standard 2)		-0.01 (0.03)
Treatment * Grade (Standard 3)		0.02 (0.02)
Treatment * Grade (Standard 4)		0.00 (0.02)
Treatment * Boy		-0.01 (0.02)
N	9,272	9,272
Mean Attrition	0.13	0.13
Control group attrition	0.13	0.13
<i>H</i> ₀ : all interactions = 0 (p-value)		0.39

This table reports the differences in observable characteristics (proportions) between students who attrit in the treatment and control groups. Standard errors (clustered by strata) in parenthesis. Final row reports p-value on joint hypothesis test that all interaction coefficients are zero. Significance: ***=.01, **=.05, *=.1.

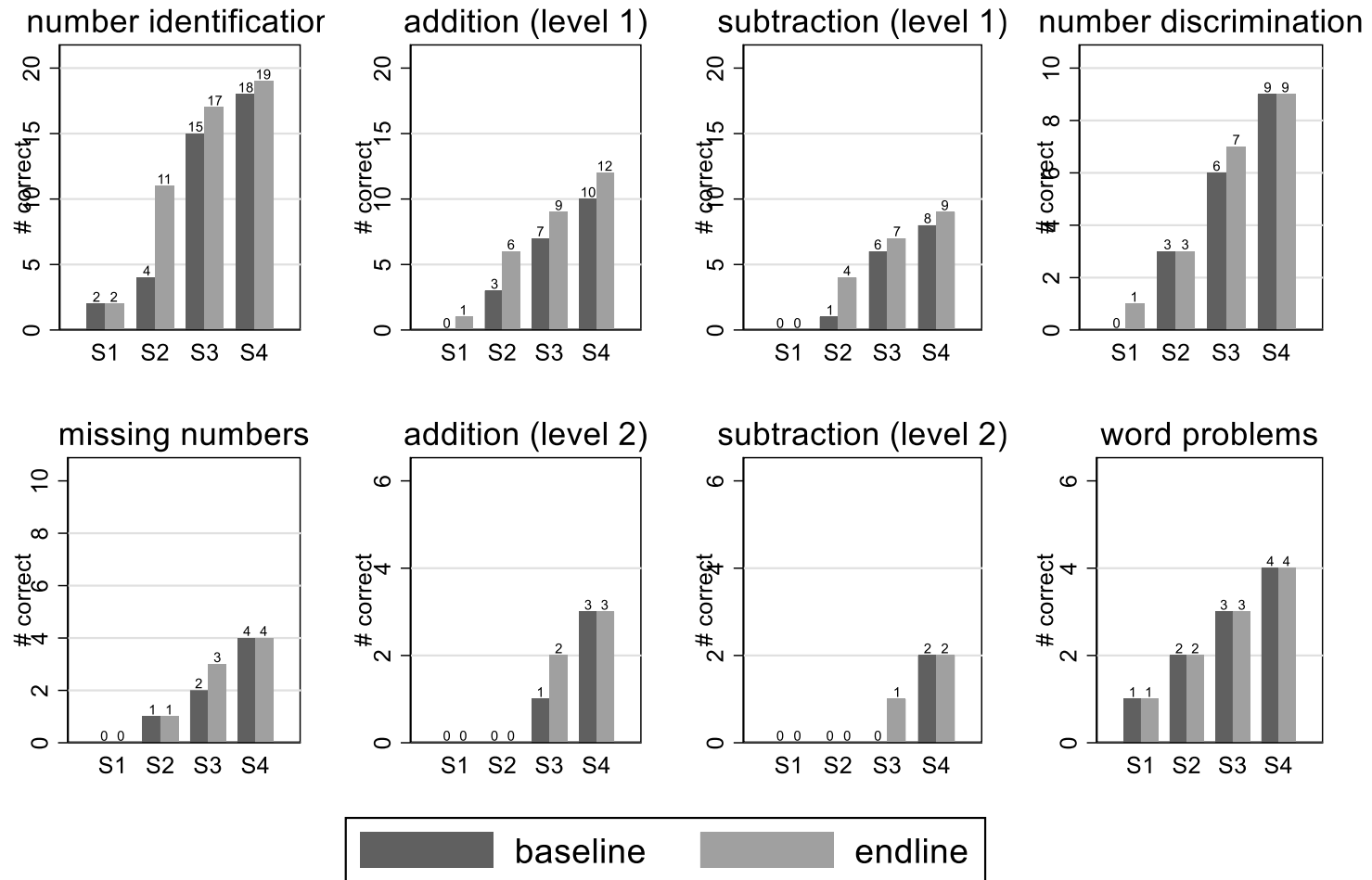
4. Results

4.1. Numeracy under the status quo

Before presenting results of the impact evaluation, we describe the numeracy of students within our study population under the status quo mathematics curriculum. The figure below shows median number of items correct on each EGMA task among students in the control group, disaggregated by standard. We limit the sample to students observed at both baseline and endline, to avoid any changes in student composition across assessments. The vertical axis of each graph corresponds to the number of items on the indicated task. By limiting the sample to the control group, we measure numeracy of the typical student under the status quo curriculum. Changes between baseline and endline represent growth in those skills over the academic year.

Figure : Median correct items by EGMA task (control group)

Median correct items by EGMA task (control group)



We draw two conclusions from the figure. First, student numeracy is relatively low. By the endline, the median S1 student can answer just one addition level 1 and zero subtraction level 1 item (most of these items are single-digit operations). Second, learning

within the academic year is modest, particularly for higher-order skills. Across most standards, lower-order skills such as number identification and addition/subtraction level 1 show improvements between baseline and endline. Progress in higher-order skills, like addition/subtraction level 2 or word problems, is scant or absent. These patterns are consistent with the Malawi numeracy scoping study (Brombacher 2019), which motivated the design of the NNP.

4.2. Numeracy: main results

4.2.1. Numeracy: Results by Overall EGMA Score.

The NNP had a compelling effect on the overall EGMA score for the full sample. In the model without baseline controls (second column of table 2), our preferred specification, the programme effect was 0.19 SD ($p < 0.01$). Does this difference reflect the nascent treatment effects observed at baseline, or subsequent effects of the programme on numeracy? To answer this question, we add the baseline EGMA score as a control (Column 3 of table 2). The effect remains robust, with a slight reduction in magnitude, after controlling for baseline EGMA score (0.17 SD, $p < 0.01$), and controlling for other covariates that were not balanced at baseline (third column of table 2).

Analyzing composite scores, we observed a significant positive impact from the intervention on students' numerical abilities. Both timed procedural fluency and untimed conceptual understanding showed improvements. The effect size for conceptual understanding (0.23, $P < 0.01$) was notably larger and more statistically significant than for procedural fluency (0.16, $P < 0.05$). Furthermore, when a composite score was created by equally weighting all subcomponents of the numeracy assessment, we found a substantial and significant effect on overall numeracy (0.20, $P < 0.01$). These results indicate a meaningful improvement in students' numeracy skills due to the intervention.

Table 2: Main Effects on numeracy, full sample

	<u>Overall EGMA score</u>			<u>Procedural fluency</u>	<u>Conceptual understanding</u>	<u>Overall score, reweighted</u>
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.19*** (0.06)	0.17*** (0.05)	0.17*** (0.05)	0.16** (0.06)	0.23*** (0.06)	0.20*** (0.06)

Baseline Score		0.49*** (0.03)	0.49*** (0.03)			
N	8,239	8,239	7,381	8,239	8,241	8,239
Control mean	0.00	0.00	0.00	0.00	0.00	0.00
Control for covariates imbalanced at baseline			X			

Table reports the main effects (standard deviation units) of the NNP on overall EGMA score (z) for the full sample. Regressions including baseline EGMA score also include dummy for missing baseline EGMA score. Controls include student age, pre-school attendance, poverty likelihood, and household size. EGMA consists of eight sections: number identification, number discrimination, addition (two levels), subtraction (two levels), missing numbers, and word problems. We calculate the overall EGMA score as the mean of the two normalized composite scores. We calculate composite scores separately for timed (correct items per minute) and untimed sections, following recommended practice based on the psychometric properties of each composite score (Geller et al. 2018). Then, we normalize by the control group mean and standard deviation within each standard (grade). The outcomes in columns 4 and 5 are EGMA scores for procedural fluency (timed composite) and conceptual understanding (untimed composite) scores. The outcome in Column 6 shows the overall EGMA score, reweighted by the number of tasks in each composite score from Columns 4-5 (37.5% for procedural fluency and 62.5% for conceptual understanding). Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1.

	<u>Numeracy (overall EGMA score, full sample)</u>			<u>Numeracy (EGMA Composite score, full sample)</u>		
	Main Specification	Main outcome (Controlling for baseline EGMA score)	Main outcome (Controlling for baseline EGMA score and other variables significant at baseline)	Timed composite score (Numerical fluency)	Untimed composite score (Conceptual understanding)	Alternative specification of main outcome (With equal weight for each EGMA subscore)
Treatment	0.19*** (0.06)	0.17*** (0.05)	0.17*** (0.05)	0.162** (0.063)	0.226*** (0.058)	0.202*** (0.058)
Baseline Score		0.49*** (0.03)	0.49*** (0.03)			
N	8,239	8,239	7,381	8239	8241	8239
Control mean	0.00	0.00	0.00	-0.000	0.000	0.000

Table reports the main effects (standard deviation units) of the NNP on overall EGMA score (z) for the full sample. Regressions including baseline EGMA score also include dummy for missing baseline EGMA score. Controls include student age, pre-school attendance, poverty likelihood, and household size. EGMA consists of eight sections: number identification, number discrimination, addition (two levels), subtraction (two levels), missing numbers, and word problems. We calculate the overall EGMA score as the mean of the two normalized composite scores. We calculate composite scores separately for timed (correct items per minute) and untimed sections, following recommended practice based on the psychometric properties of each composite score (Geller et al. 2018). Then, we normalize by the control group mean and standard deviation within each standard (grade). Columns 4,5, and 6 show the effect on numeracy for the EGMA untimed composite, timed composite, and composite reweighted scores. Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1.

Figure 1 shows kernel density estimates of the endline EGMA score distribution, separately for treatment and control. The horizontal axis is the overall EGMA score, while the height of the vertical axis represents the density of students with that score. A density farther to the right indicates higher scores. Figure 1 illustrates the shift in the distribution of overall EGMA scores, with the treatment group consistently outperforming the control group across the range of scores. The density plots show a clear rightward shift for the treatment group, particularly around the mean and higher end of the distribution. Figures 2-5 show the distribution of scores separately for each grade. The programme appears to have shifted the distribution most dramatically for Standards 3-4.

Figure 1: EGMA Overall Score

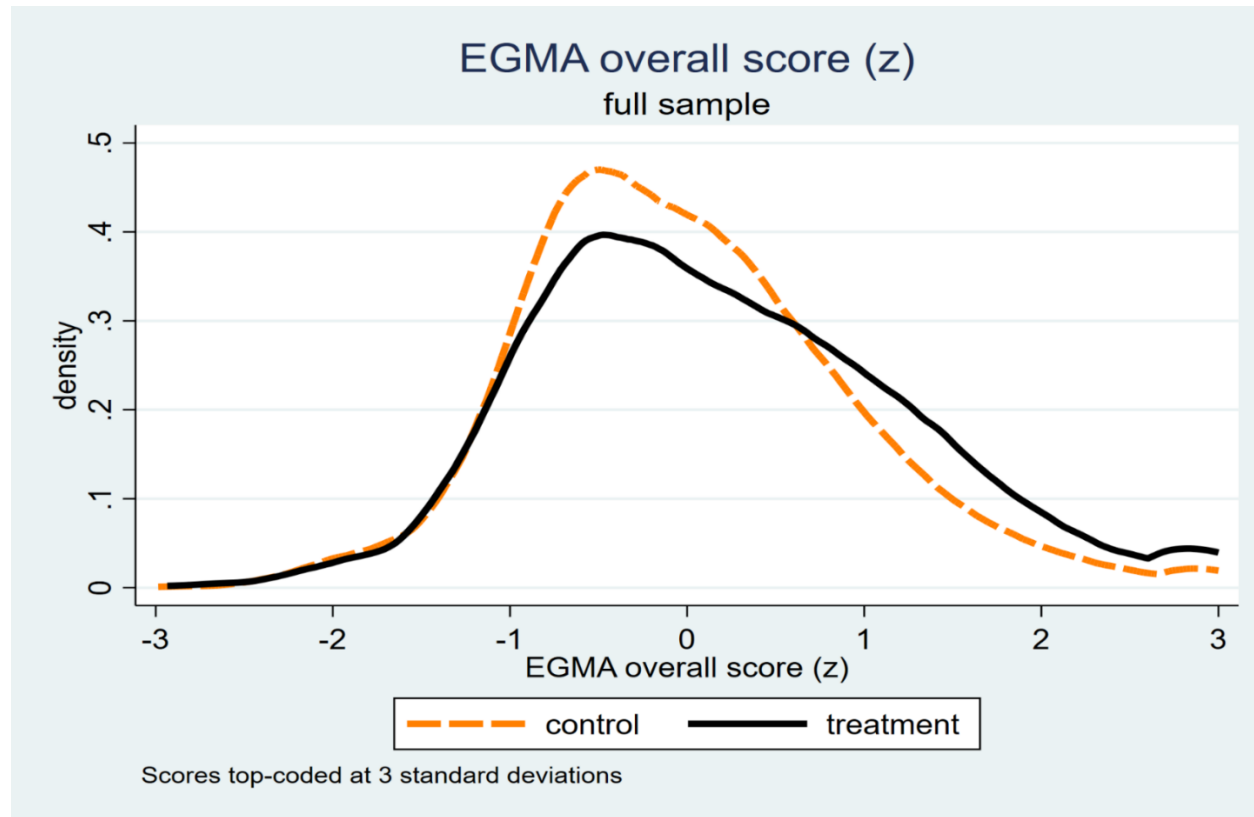


Figure shows kernel density estimates of endline EGMA overall z-score in full sample, by treatment status. Scores top-coded at ± 3 standard deviations for purposes of visual display.

Figure 2: EGMA Overall Score Standard 1

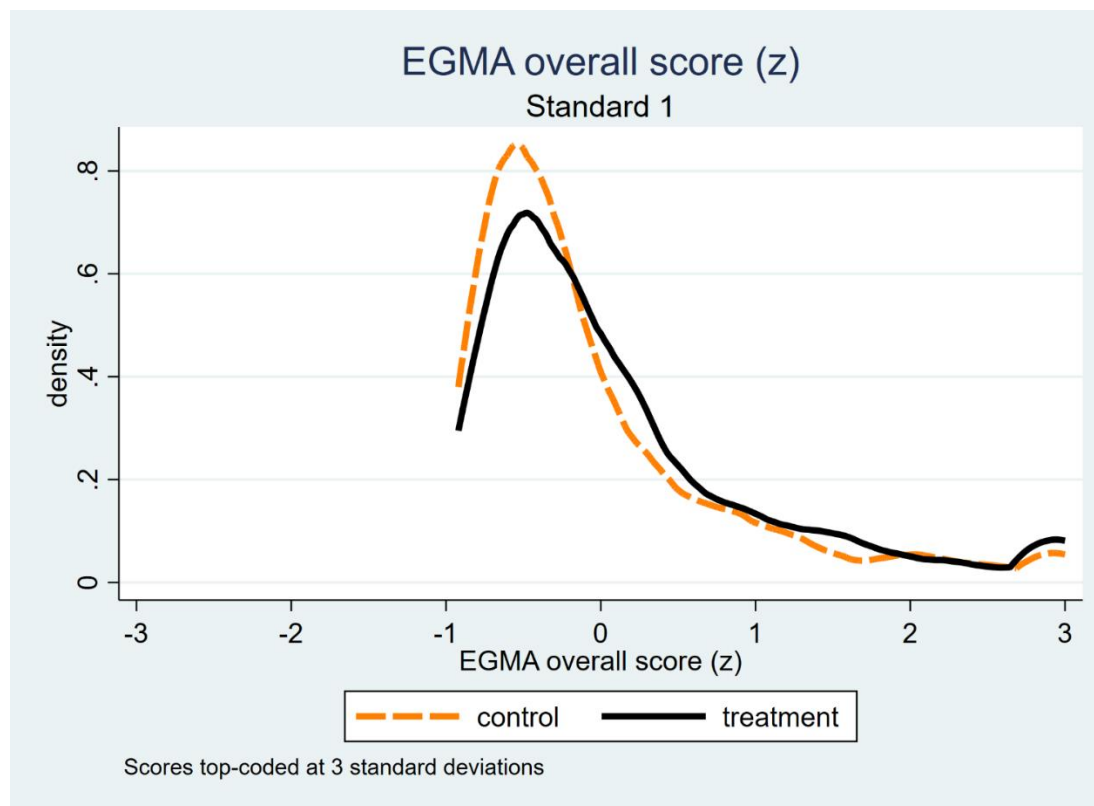


Figure shows kernel density estimates of endline EGMA overall z-score for Standard 1 students, by treatment status. Scores top-coded at ± 3 standard deviations for purposes of visual display.

Figure 3: EGMA Overall Score Standard 2

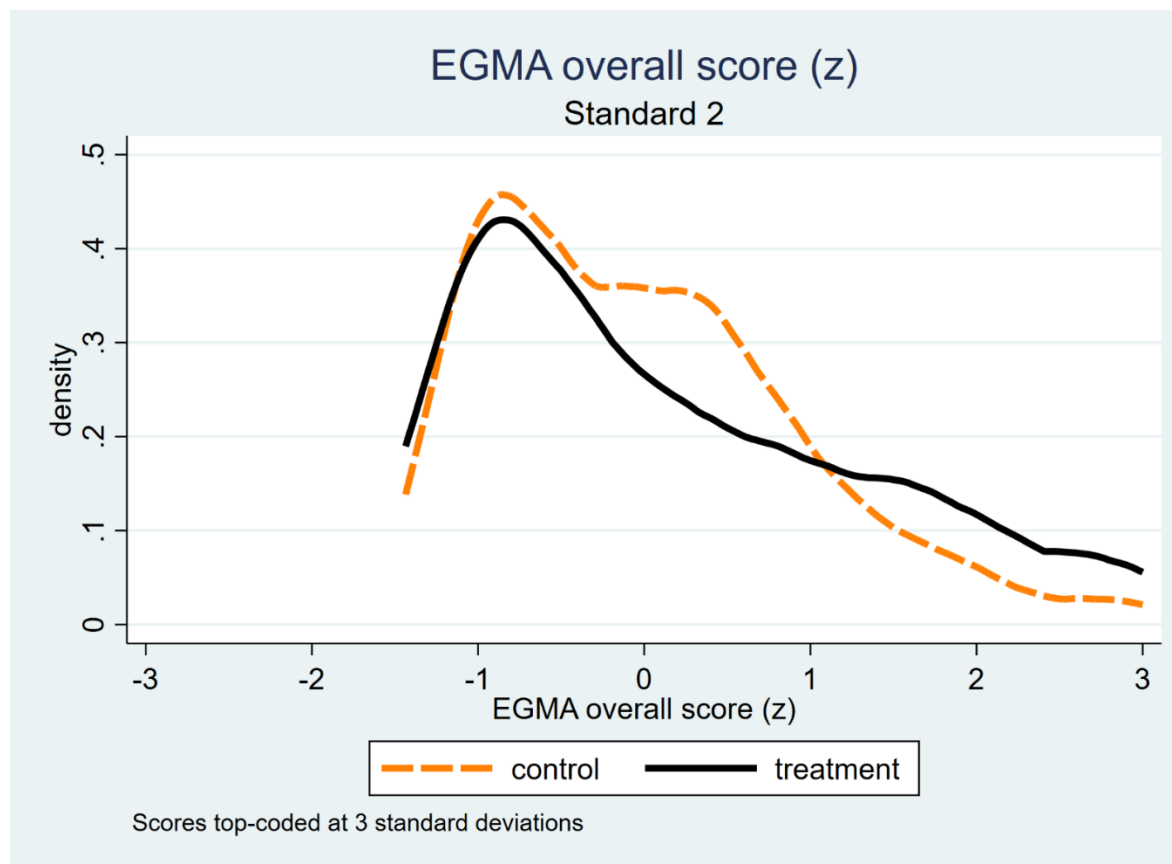


Figure shows kernel density estimates of endline EGMA overall z-score for Standard 2 students, by treatment status. Scores top-coded at ± 3 standard deviations for purposes of visual display.

Figure 4: EGMA Overall Score Standard 3

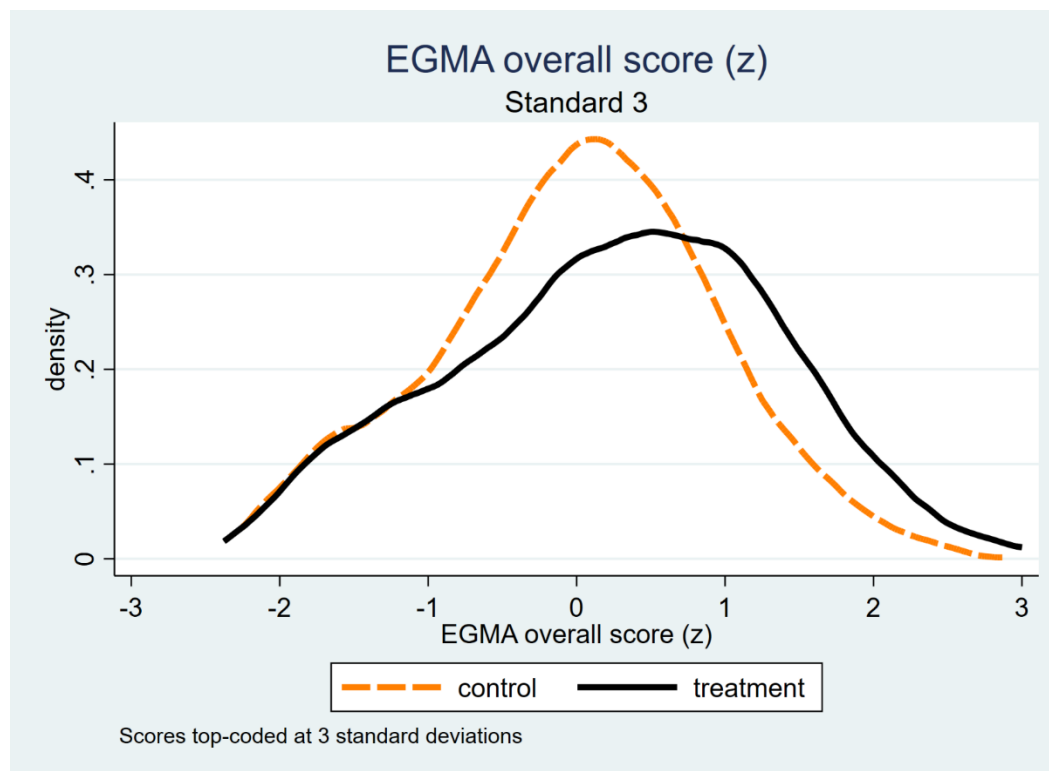


Figure shows kernel density estimates of endline EGMA overall z-score for Standard 3 students, by treatment status. Scores top-coded at ± 3 standard deviations for purposes of visual display.

Figure 5: EGMA Overall Score Standard 4

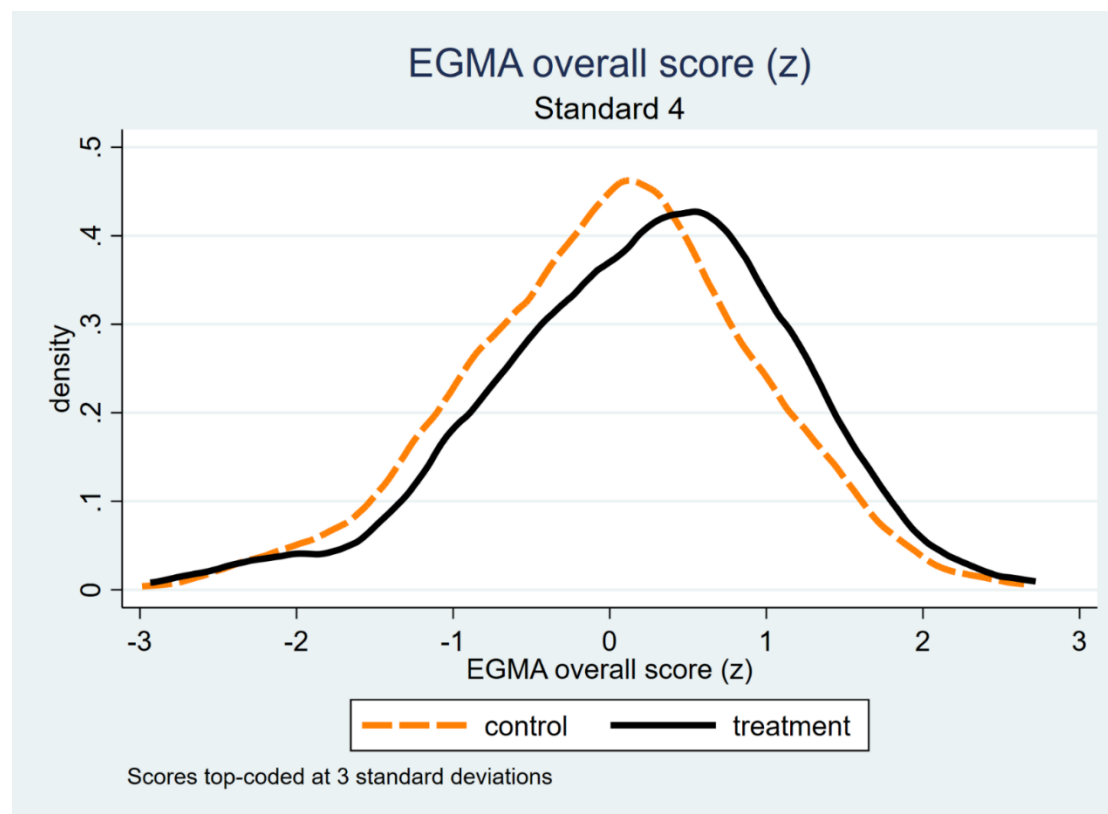


Figure shows kernel density estimates of endline EGMA overall z-score for Standard 4 students, by treatment status. Scores top-coded at ± 3 standard deviations for purposes of visual display.

The programme's impact varied across grades, as shown in Table 3. In Standard 1, The programme had a strong effect of 0.17 SD ($p < 0.05$) in the uncontrolled model, but the effect size decreased slightly to 0.15 SD ($p < 0.01$) after controlling for baseline scores. Figure 2 supports these results, showing a noticeable improvement in the density of scores for the treatment group. In Standard 2, The programme effect was small (0.17 and 0.14 SD), $p < 0.1$ in both models. For Standard 3, NNP produced compelling gains of 0.24

SD ($p < 0.01$) in the uncontrolled model, which remained unchanged after including baseline controls. The effect of NNP was also compelling in standard 4 (0.21 SD, $p < 0.01$) in the uncontrolled model and remained robust when baseline scores were accounted for, though with a much smaller magnitude (0.14 SD). Overall, in most cases, controlling for baseline EGMA scores resulted in small reductions in effect size but no change in significance, suggesting that the programme's impact is robust and not dependent on pre-existing student performance. We also find the strongest gains in standards 3 and 4. Point estimates are smaller for standards 1-2, but confidence intervals overlap our larger estimates for standards 3-4, indicating no statistically significant differences. Effect sizes range from 0.17 to 0.24 standard deviations.

Table 3: Main Effects on numeracy (By Grade)

	Standard 1		Standard 2		Standard 3		Standard 4	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment Group	0.17** (0.07)	0.15*** (0.07)	0.17* (0.09)	0.14* (0.09)	0.24*** (0.07)	0.26*** (0.06)	0.21*** (0.06)	0.14*** (0.05)
Baseline Overall Score		0.30*** (0.11)		0.55*** (0.04)		0.63*** (0.03)		0.59*** (0.03)
N	2,026	2,026	2,091	2,091	2,061	2,061	2,061	2,061
Control mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

This table reports the main effects (SD units) of the NNP on the overall EGMA score (z) by grades. Regressions including baseline EGMA score also include dummy for missing baseline EGMA score. We calculate the overall EGMA score as the mean of the two normalized composite scores. We calculate composite scores separately for timed (correct items per minute) and untimed sections, following recommended practice based on the psychometric properties of each composite score (Geller et al. 2018). Then, we normalize by the control group mean and standard deviation within each standard (grade). Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1.

4.2.2. Numeracy: Results by Sub Scores

The treatment effects of the NNP on EGMA subscores reveal important insights into the specific numeracy skills most influenced by the NNP intervention. Across the eight subtests, certain tasks showed particularly strong gains, indicating where the programme was most impactful.

Among the strongest effects was the improvement in Number Identification, where students in the treatment group consistently outperformed their control peers. In the overall sample (Table 4), the treatment effect was a compelling increase of 2.29 correct items per minute ($p < 0.01$). This pattern persisted across all grades, with similarly strong effects, particularly in Standards 3 and 4 (Tables 7 and 8), where treatment school students achieved gains exceeding 3 correct items per minute.

Table 4: Main Effects by Sub Scores

	Number Identification	Addition Level 1	Subtraction Level 1	Number Discrimination	Missing Numbers	Addition Level 2	Subtraction Level 2	Word Problems
Treatment	2.29*** (0.50)	0.02 (0.21)	-0.33 (0.21)	0.54*** (0.12)	0.54*** (0.09)	0.21*** (0.06)	0.10 (0.06)	0.12* (0.07)
N	8,241	8,241	8,239	8,241	8,241	8,241	8,241	8,241
Control Mean	16.10	7.17	5.85	4.92	2.28	1.46	1.13	2.53
Units	# correct per minute	# correct per minute	# correct per minute	# correct (of 10)	# correct (of 10)	# correct (of 5)	# correct (of 5)	# correct (of 6)

This table reports the main effects of the NNP by sub scores. Standard errors in parenthesis. Significance: ***=.01, **=.05, *=.1. Number identification, Addition level 1 and subtraction level 1 are timed subtests and the unit of measurement is number of correct items per minute. The remaining subtests: number discrimination (10 items); missing numbers (10 items); addition level 2 (5 items); subtraction level 2 (5 items) and word problems (6 items) are untimed and the unit of measurement is the number of correct items.

Table 5: Main Effects by Sub Scores (Standard 1)

	Number Identification	Addition Level 1	Subtraction Level 1	Number Discrimination	Missing Numbers	Addition Level 2	Subtraction Level 2	Word Problems
Treatment	1.53***	-0.09	-0.32*	0.39***	0.37***	0.05	0.03	-0.02

	(0.29)	(0.20)	(0.17)	(0.11)	(0.06)	(0.04)	(0.03)	(0.09)
N	2,026	2,026	2,026	2,026	2,026	2,026	2,026	2,026
Control Mean	3.75	2.23	1.73	1.09	0.48	0.13	0.09	1.24
Units	# correct per minute	# correct per minute	# correct per minute	# correct (of 10)	# correct (of 10)	# correct (of 5)	# correct (of 5)	# correct (of 6)

This table reports the main effects of the NNP by sub scores for standard 1. Standard errors in parenthesis. Significance: ***=.01, **=.05, *=.1. Number identification, Addition level 1 and subtraction level 1 are timed subtests and the unit of measurement is number of correct items per minute. The remaining subtests: number discrimination (10 items); missing numbers (10 items); addition level 2 (5 items); subtraction level 2 (5 items) and word problems (6 items) are untimed and the unit of measurement is the number of correct items.

Table 6: Main Effects by Sub Scores (Standard 2)

	Number Identification	Addition Level 1	Subtraction Level 1	Number Discrimination	Missing Numbers	Addition Level 2	Subtraction Level 2	Word Problems
Treatment	1.79** (0.69)	-0.14 (0.35)	-0.55 (0.35)	0.71*** (0.21)	0.42*** (0.12)	0.27** (0.11)	0.18* (0.09)	0.12 (0.10)
N	2,092	2,092	2,092	2,092	2,092	2,092	2,092	2,092
Control Mean	11.09	5.49	4.50	3.49	1.54	0.83	0.63	2.12
Units	# correct per minute	# correct per minute	# correct per minute	# correct (of 10)	# correct (of 10)	# correct (of 5)	# correct (of 5)	# correct (of 6)

This table reports the main effects of the NNP by sub scores for standard 2. Standard errors in parenthesis. Significance: ***=.01, **=.05, *=.1. Number identification, Addition level 1 and subtraction level 1 are timed subtests and the unit of measurement is number of correct items per minute. The remaining subtests: number discrimination (10 items); missing numbers (10 items); addition level 2 (5 items); subtraction level 2 (5 items) and word problems (6 items) are untimed and the unit of measurement is the number of correct items.

Other compelling improvements were observed in Number Discrimination and Missing Numbers, untimed subtests measuring conceptual understanding of numbers. These tasks showed gains of 0.54 correct responses ($p < 0.01$) each in the overall sample (Table

4), with strong and consistent effects across grades. For example, in Standard 3 (Table 7), the treatment effect on Missing Numbers reached 0.88 correct responses ($p < 0.01$), underscoring the programme's impact on students' ability to identify numerical patterns.

Table 7: Main Effects by Sub Scores (Standard 3)

	Number Identification	Addition Level 1	Subtraction Level 1	Number Discrimination	Missing Numbers	Addition Level 2	Subtraction Level 2	Word Problems
Treatment	3.13*** (0.70)	0.36 (0.26)	-0.16 (0.30)	0.66*** (0.18)	0.88*** (0.15)	0.24** (0.11)	0.09 (0.09)	0.24** (0.11)
N	2,062	2,062	2,062	2,062	2,062	2,062	2,062	2,062
Control Mean	20.59	8.97	7.43	6.75	2.93	2.05	1.59	3.03
Units	# correct per minute	# correct per minute	# correct per minute	# correct (of 10)	# correct (of 10)	# correct (of 5)	# correct (of 5)	# correct (of 6)

This table reports the main effects of the NNP by sub scores for standard 3. Standard errors in parenthesis. Significance: ***=.01, **=.05, *=.1. Number identification, Addition level 1 and subtraction level 1 are timed subtests and the unit of measurement is number of correct items per minute. The remaining subtests: number discrimination (10 items); missing numbers (10 items); addition level 2 (5 items); subtraction level 2 (5 items) and word problems (6 items) are untimed and the unit of measurement is the number of correct items.

Results for some higher-order skills, particularly in Addition Level 2 and Word Problems, were also notable. In the overall sample (Table 4), Addition Level 2 showed a strong effect of 0.21 correct responses ($p < 0.01$), while Word Problems demonstrated a smaller improvement of 0.12 correct responses ($p < 0.1$). These gains reflect the programme's influence on more complex problem-solving skills, though the magnitude of effects was smaller compared to the conceptual number tasks described above. Notably, the magnitude and significance of the treatment effects on these tasks improved with grade, reaching 0.30 correct items ($P < 0.01$) for addition level 2 and 0.19 correct items ($P < 0.05$) for word problems in Standard 4 (Table 8). This suggests cumulative benefits of the programme for these tasks as students progress.

Table 8: Main Effects by Sub Scores (Standard 4)

	Number Identification	Addition Level 1	Subtraction Level 1	Number Discrimination	Missing Numbers	Addition Level 2	Subtraction Level 2	Word Problems
Treatment	3.08*** (0.71)	0.08 (0.26)	-0.18 (0.29)	0.51*** (0.12)	0.55*** (0.16)	0.30*** (0.11)	0.10 (0.10)	0.19** (0.09)
N	2,061	2,061	2,061	2,061	2,061	2,061	2,061	2,061
Control Mean	28.69	11.86	9.64	8.28	4.15	2.81	2.18	3.69
Units	# correct per minute	# correct per minute	# correct per minute	# correct (of 10)	# correct (of 10)	# correct (of 5)	# correct (of 5)	# correct (of 6)

This table reports the main effects of the NNP by sub scores for standard 4. Standard errors in parenthesis. Significance: ***=.01, **=.05, *=.1. Number identification, Addition level 1 and subtraction level 1 are timed subtests and the unit of measurement is number of correct items per minute. The remaining subtests: number discrimination (10 items); missing numbers (10 items); addition level 2 (5 items); subtraction level 2 (5 items) and word problems (6 items) are untimed and the unit of measurement is the number of correct items.

In contrast, the programme had no impact on Addition Level 1 and Subtraction Level 1, both timed tasks, and for subtraction level 2. For the overall sample and for all grades, there were no treatment effects on these subscores at standard significance levels.

Overall, the NNP intervention had its strongest and most consistent effects on Number Identification, Number Discrimination, and Missing Numbers. While improvements in more advanced problem-solving tasks like Addition Level 2 and Word Problems were evident, these effects were generally smaller and more grade-specific, with stronger effects for higher grades.

We exercise caution in interpreting the differential effects of the programme across various EGMA subscores for several reasons. While improvements in higher-order tasks, such as addition level 2 and word problems, were smaller in magnitude, they became more pronounced in higher grades, suggesting these skills require a longer period to fully develop. Importantly, all subtests demonstrated positive point estimates of treatment effects, even in cases where statistical significance was not achieved in the overall sample. This consistent pattern underscores the broad-based effectiveness of the intervention, albeit with variability in the magnitude and timing of its impacts.

4.2.3. Numeracy: Proficiency Analysis.

In our proficiency analysis, we focused on evaluating performance at the extremes of the ability distribution (proficient and scoring zero) for several EGMA subtests (see appendix 9.2 for details): addition and subtraction level 1, missing numbers, and word problems. We selected these subtests for proficiency analysis due to the range of skills represented, consistency with proficiency benchmarks established for other Sub-Saharan African countries, and for the alignment of word problems with skills targeted by the NNP. For clarity of exposition and consistency with proficiency benchmarks established else, we do not define proficiency for all subtests. The previous section reports scores from all subtests for the interested reader.

For addition and subtraction level 1, we combined the number of correct items across the two tasks (20 items per task) and defined a 0 score as scoring zero on this combined 40-item measure. Proficiency for addition and subtraction was defined as achieving a score of 32 out of 40 or higher, equivalent to 80%. For missing numbers and word problems, proficiency was defined as achieving 7 out of 10 and 4 out of 6 correct responses respectively, corresponding to 70% accuracy.

Table 9, which reports mean proficiency rates by treatment status for the overall sample, reveals a consistent pattern of higher proficiency rates in the treatment group, except for addition/subtraction level 1. For missing numbers, the proportion of proficient students was 11% in the treatment group compared to 7% in the control group, a meaningful difference. Similarly, for word problems, proficiency in the treatment group reached 34%, compared to 29% in the control group. Zero scores were also lower in the treatment group for missing numbers (21% against 26% of students in the control group), but slightly higher in the treatment group for addition/subtraction level 1 and word problems.

Table 9: Mean Proficiency Rates by Treatment Status, full sample

	<u>Proficient</u>		<u>Zero Scores</u>	
	Treatment	Control	Treatment	Control
Addition/Subtraction	0.03 (0.00)	0.03 (0.00)	0.15 (0.00)	0.13 (0.01)
Missing Numbers	0.11	0.07	0.21	0.26

	(0.00)	(0.00)	(0.01)	(0.01)
Word Problems	0.34	0.29	0.15	0.14
	(0.01)	(0.01)	(0.01)	(0.01)
N	4,105	4136	4,105	4,136

This table reports the overall mean proficiency score (proportions) for addition/subtraction, missing numbers, and word problems by treatment status. We define proficiency benchmark for addition and subtraction level 1 at 80% and Missing numbers and word problems at 70%. Standard errors in parenthesis.

Table 10 provides proficiency and zero score rates disaggregated by grade, showing some notable patterns. Proficiency gains were most pronounced in Standards 3 and 4, where the treatment group consistently outperformed the control group in both missing numbers and word problems. For example, in Standard 3, the proficiency rate for word problems was 47% for the treatment group and 37 for the control group, while for standard 4 it was 63% versus 57%. For Missing numbers, the proportions are 24% versus 20% for standard 4 and 16% versus 8% in standard 3. Additionally, the proportion of zero scores for missing numbers was visibly lower in the treatment group, particularly in Standard 1.

Table 10: EGMA Proficiency Rates by Standards

Standard 1				
	<u>Proficient</u>		<u>Zero Scores</u>	
	Control	Treatment	Control	Treatment
Addition/Subtraction	0	0.00	0.35	0.38
	-	(0.00)	(0.02)	(0.02)
Missing Numbers	0.00	0.00	0.65	0.48
	(0.00)	(0.00)	(0.2)	(0.02)
Word Problems	0.04	0.04	0.34	0.35
	(0.01)	(0.01)	(0.01)	(0.01)
N	2,026	2,026	2,026	2,026
Standard 2				
	<u>Proficient</u>		<u>Zero Scores</u>	

	Control	Treatment	Control	Treatment
Addition/Subtraction	0.06 (0.00)	0.01 (0.00)	0.13 (0.01)	0.19 (0.01)
Missing Numbers	0.02 (0.00)	0.04 (0.01)	0.26 (0.01)	0.28 (0.01)
Word Problems	0.18 (0.01)	0.22 (0.01)	0.13 (0.01)	0.16 (0.01)
N	2,092	2,092	2,092	2,092
Standard 3				
	<u>Proficient</u>		<u>Zero Scores</u>	
	Control	Treatment	Control	Treatment
Addition/Subtraction	0.02 (0.00)	0.04 (0.00)	0.02 (0.00)	0.03 (0.01)
Missing Numbers	0.08 (0.01)	0.16 (0.01)	0.1 (0.01)	0.06 (0.01)
Word Problems	0.37 (0.02)	0.47 (0.02)	0.04 (0.01)	0.06 (0.01)
N	2,062	2,062	2,062	2,062
Standard 4				
	<u>Proficient</u>		<u>Zero Scores</u>	
	Control	Treatment	Control	Treatment
Addition/Subtraction	0.08 (0.01)	0.08 (0.01)	0.00 (0.00)	0.01 (0.00)
Missing Numbers	0.20 (0.01)	0.24 (0.01)	0.04 (0.01)	0.05 (0.01)
Word Problems	0.57 (0.02)	0.63 (0.02)	0.03 (0.01)	0.03 (0.00)
N	2,061	2,061	2,061	2,061

This table reports the overall mean proficiency rate (proportions) for addition/subtraction, missing numbers, and word problems by treatment status. We define proficiency benchmark for addition and subtraction level 1 at 80% and Missing numbers and word problems at 70%. Standard errors in parenthesis.

Table 11: Effects on Zero Scores and Proficiency

Sample	Zero Score					Proficiency Score				
	Overall	Standard 1	Standard 2	Standard 3	Standard 4	Overall	Standard 1	Standard 2	Standard 3	Standard 4
Panel A: Outcome: Addition/subtraction (level 1)										
Treatment	0.027*** (0.01)	0.03 (0.03)	0.05*** (0.02)	0.01 (0.01)	-0.00 (0.00)	0.004 (0.006)	0.001 (0.001)	0.006 (0.008)	0.014 (0.010)	-0.004 (0.019)
N	8,241	2,026	2,092	2,062	2,061	8,241	2,026	2,092	2,062	2,061
Control group mean	0.13	0.35	0.13	0.02	0.00	0.29	0.04	0.18	0.37	0.57
Panel B: Outcome: Missing numbers										
Treatment Group	-0.05*** (0.01)	-0.18*** (0.03)	0.01 (0.02)	0.01 (0.01)	-0.00 (0.00)	0.036*** (0.009)	-0.001 (0.002)	0.028*** (0.011)	0.078*** (0.020)	0.041 (0.025)
N	8,241	2,026	2,092	2,062	2,061	8241	2026	2092	2062	2061
Control group mean	0.26	0.65	0.26	0.10	0.04	0.073	0.002	0.015	0.076	0.197
Panel C: Outcome: Word problems										
Treatment Status	0.01 (0.01)	0.01 (0.03)	0.03 (0.02)	0.01 (0.01)	-0.01 (0.01)	0.05*** (0.02)	0.00 (0.01)	0.04** (0.02)	0.09*** (0.03)	0.05*** (0.03)
N	8,241	2,026	2,092	2,062	2,061	8,241	2,026	2,092	2,062	2,061
Control group mean	0.14	0.34	0.13	0.04	0.03	0.29	0.04	0.18	0.37	0.57

In terms of proficiency, Table 11 shows that the treatment had a compelling impact on word problem proficiency, with an overall increase of 5 percentage points ($p < 0.01$). These gains were particularly notable in Standard 3, where proficiency rose by 9 percentage points ($p < 0.01$), and Standard 4, where the treatment group also experienced a 5 percentage point increase ($p < 0.01$).

Outcomes with minimal variation in the control group, such as proficiency for addition/subtraction and missing numbers, were excluded from further analysis. Overall, the effect of NNP on zero scores was mixed and driven by particular grades. The intervention increased zero scores in addition/subtraction level 1 (driven by standard 2), reduced zero scores in missing numbers (driven by standard 1) and had no effect on zero scores for word problems. The programme increased the proportion of students achieving proficiency in word problems, with stronger effects for students in higher grades.

4.2.4. Numeracy: Transforming Numeracy Gains into Schooling Impact (EYOS and LAYS)

To make the results of the NNP impact evaluation more relatable, we translated learning gains into measures that reflect their real-world implications: Malawi Equivalent Years of Schooling (EYOS) and Learning Adjusted Years of Schooling (LAYS). These measures help communicate how much additional schooling the NNP's learning gains represent, both within the local context and compared to a high-performing global benchmark.

EYOS reflects how much extra learning students achieved, expressed as the equivalent of years spent in a typical Malawian school. For example, a gain of 0.5 EYOS means students learned an additional half a year of typical Malawian schooling as a result of the programme. LAYS builds on this by comparing the gains to a global benchmark for high-quality education, reflecting how far students' learning has progressed toward international standards.

Table 12 presents the learning gains described in Table 2 – reproduced in the row named “NNP Learning Gain (SD Units)” - in terms of EYOS and LAYS. In terms of EYOS, this gain translates to approximately 0.56 years of additional schooling for the average student in the NNP programme. By grade, the largest EYOS gains were observed in Standard 1 (1.61 EYOS), largely the result of very low status quo learning in standard 1. Gains in other grades ranged from 0.33 EYOS in Standard 2 to 0.62 EYOS in Standard 3, reflecting consistent improvements across the board.

Table 12: NNP Learning Gains

Gain	Overall	Standard 1	Standard 2	Standard 3	Standard 4
Learning in the status quo (SD Units)	0.35*** (0.04)	0.10* (0.06)	0.50*** (0.06)	0.41*** (0.05)	0.38*** (0.05)
NNP Learning Gain (SD Units)	0.19*** (0.06)	0.17** (0.07)	0.17* (0.09)	0.24*** (0.07)	0.21*** (0.06)
Equivalent Years of Schooling (EYOS) gain	0.56	1.61	0.33	0.62	0.53
Learning Adjusted Years of Schooling (LAYS) gain	0.25	0.20	0.21	0.32	0.26

The table reports NNP learning impacts in terms of Malawi Equivalent Years of Schooling (EYOS) and Learning Adjusted Years of Schooling (LAYS). NB:

1. To calculate status quo learning, we first standardize the endline scores using the baseline distribution of the control group. Then, we subtract the baseline EGMA scores from the re-standardized endline scores for students in the control group.
2. Equivalent years of Schooling is obtained by expressing NNP learning gain relative to business as usual learning, in this case learning in control schools:

$$E = \beta^{EGMA} \div D^{EGMA} \text{ (See Evans and Yuan, 2019)}$$

Where β is the learning gain produced by the intervention, D is the status quo learning rate per academic year in Malawi (approximated using learning rate in control schools), and EGMA is the test used to measure learning.

3. Learning Adjusted Years of schooling (LAYS) is obtained by adjusted the equivalent years of schooling above relative to learning in a high-performance benchmark country as follows:

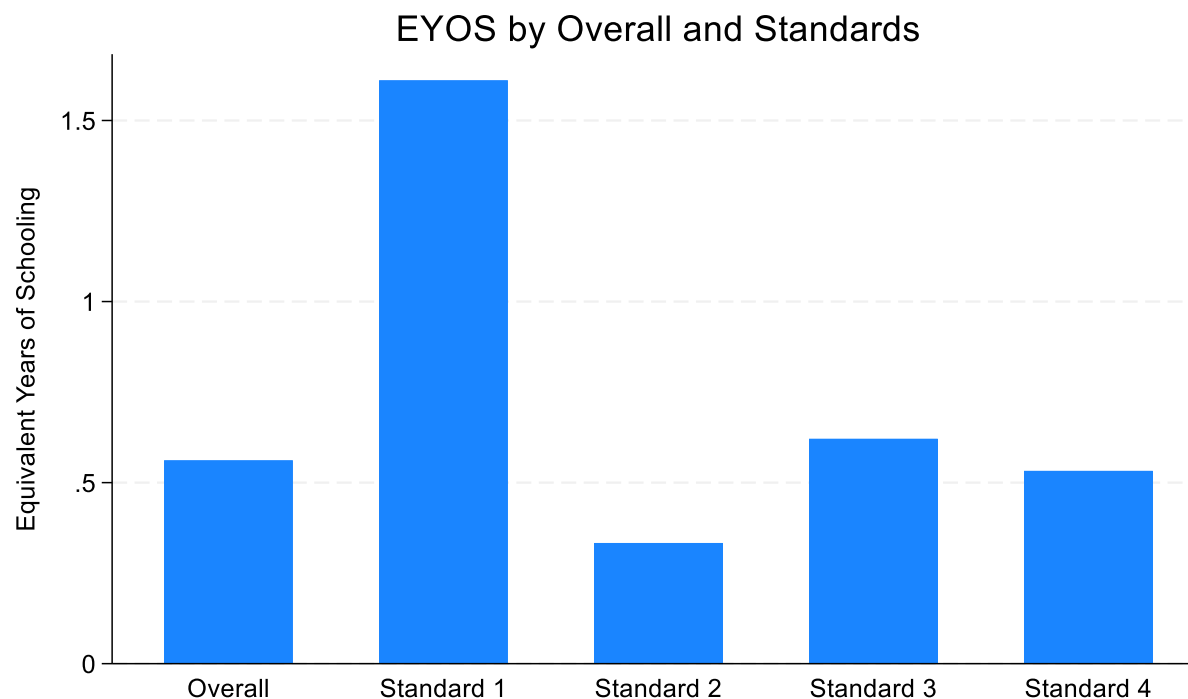
$$LAYS_i = \beta \div D_{h,n}$$

Where β is as above and $D_{h,n}$ is the high-quality benchmark learning rate. Here, we employ an artificial benchmark learning rate of 0.8 SD, following recommendations by Angrist et al (2023)

When adjusted to reflect global benchmarks, these results translate into LAYS of approximately 0.25 years overall, with individual grades achieving between 0.20 LAYS in Standard 1 and 0.32 LAYS in Standard 3. While smaller in magnitude than EYOS, these results highlight the programme's contribution to narrowing the gap between Malawian students' learning and that of students in high-performing education systems.

Figure 6 visually represents these results, showing both EYOS and LAYS gains for the full sample and by grade. These findings demonstrate the NNP’s ability to provide measurable learning gains that go beyond statistical significance, translating improvements in numeracy into tangible schooling outcomes that are meaningful to educators, policymakers, and the general public.

Figure 6: Equivalent Years of Schooling



This figure represents the Equivalent Years of Schooling (EYOS) for the overall sample and by standards. EYOS is obtained by expressing NNP learning gain relative to business as usual learning, in this case learning in control schools.

4.3. Numeracy: heterogeneous treatment effects

This section reports whether the numeracy results vary according to student characteristics.

Table 13 reports results by student gender. The programme improved numeracy of both girls and boys. We find a treatment effect of 0.23 standard deviations for girls, significant at 1%. We find a treatment effect of 0.15 standard deviations for boys, significant at 10%. Although the point estimate of the treatment effect is larger for girls than boys, this may reflect statistical noise rather than a truly larger effect for girls. To test whether the larger point estimate for girls is statistically distinguishable from the effect for boys, in column (3) we interact treatment with the indicator for girls. The interaction term is positive but not statistically distinguishable from zero, so we cannot conclude that the programme had different effects for girls and boys.

Table 13: Main Effects on numeracy by Gender

	Outcome: overall EGMA score (z)		
	(1) Girl	(2) Boy	(3) Interacted model
Treatment	0.232*** (0.062)	0.152* (0.078)	0.155** (0.077)
Treatment * Girl			0.075 (0.074)
N	4,178	4,061	8,239
Control Mean	-0.078	0.080	0.000

This table reports the main effects of the NNP on the overall EGMA score (z) by gender. We calculate the overall EGMA score as the mean of the two normalized composite scores. We calculate composite scores separately for timed (correct items per minute) and untimed sections, following recommended practice based on the psychometric properties of each composite score (Geller et al. 2018). Then, we normalize by the control group mean and standard deviation within each standard (grade). Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1.

Did the effects of the NNP depend on a student's beginning numeracy level? To test this, we split the sample into three groups, based on terciles of the baseline EGMA distribution. Because EGMA scores have already been normalized within each standard, these groupings account for different numeracy levels by standard. The bottom tercile ranges from the lowest performers to the moderately

low performers, with a maximum score of -0.43 standard deviations below the control group mean. The middle tercile ranges from moderately low to moderately high performers, from -0.43 to 0.28 standard deviations relative to the control group mean. The top tercile ranges from moderately high performers to the top performers. The score ranges within each tercile are reported at the bottom of Table 14.

After grouping students in this way, we estimate treatment effects separately within each tercile. As shown in Table 14, the middle and top terciles benefitted from the NNP. For the bottom tercile, we fail to find evidence that the NNP improved numeracy. These results suggest the programme disproportionately benefitted students with better initial numeracy skills.

Table 14: Main Effects on Numeracy by Baseline EGMA Tercile

Sample: baseline EGMA tercile	Outcome: overall EGMA score (z)		
	Bottom Tercile	Middle Tercile	Top Tercile
Treatment	0.05 (0.05)	0.21*** (0.06)	0.29*** (0.08)
N	2,738	2,747	2,754
Control Mean	-0.48	0.01	0.49
Baseline EGMA range within tercile (Standard deviation units)			
	[-2.84, -0.43]	[-0.43, 0.28]	[0.28, 10.10]

This table reports the main effects of the NNP on the overall EGMA score (z) by the baseline EGMA tercile. We calculate the overall EGMA score as the mean of the two normalized composite scores. We calculate composite scores separately for timed (correct items per minute) and untimed sections, following recommended practice based on the psychometric properties of each composite score (Geller et al. 2018). Then, we normalize by the control group mean and standard deviation within each standard (grade). Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1.

In addition to initial skills, the benefits of the programme may also depend on intensity of exposure. Programme implementers expressed concern that high levels of student absenteeism might dilute the programme's benefits. Unfortunately, we lack data on student attendance at school during the evaluation period. To test indirectly if programme exposure explains the programme's impact, we use two proxies for student attendance. First, we use an item from the baseline parent survey which asked how many days their child was present in the previous week. Second, we use the attendance rate reported in the endline classroom observation. For both proxies, we split the sample into two groups based on the control group median: "low attendance" for below the median,

“high attendance” for at or above the median. The baseline control group median was five days present in school. The endline control group median attendance rate in observed classrooms was 66.5%.

The baseline attendance proxy appears to split the sample as intended, with the low attendance group faring worse on the endline EGMA than the high attendance group (control group means reported at bottom of Table 15). The NNP increased numeracy within both groups. Surprisingly, the point estimate for the low attendance group is larger than for the high attendance group, but the size of the standard errors suggests they are statistically indistinguishable.

The endline attendance proxy is more difficult to interpret, because students within the sample are those who showed up for the endline assessment. This may explain the nearly identical mean EGMA scores at endline between the low and high attendance groups using the endline proxy. We fail to find significant impacts of the NNP for either the low or high attendance group when using the endline proxy, although both point estimates are positive.

It remains possible that the programme’s effects were mediated by student attendance. We are unable to find evidence for this hypothesis using available data, however.

Table 15: Results by Baseline and Endline Attendance Level

	Outcome: overall EGMA score (z)			
	Baseline		Endline	
	(1) Low Attendance	(2) High Attendance	(3) Low Attendance	(4) High Attendance
Treatment	0.230*** (0.086)	0.178*** (0.062)	0.135 (0.106)	0.205 (0.130)
Mean (Control)	-0.057	0.023	0.007	-0.007
N	1,212	6,229	4,336	3,845

This table reports the main effects by the baseline and endline attendance level. We calculate the overall EGMA score as the mean of the two normalized composite scores. We calculate composite scores separately for timed (correct items per minute) and untimed sections, following recommended practice based on the psychometric properties of each composite score (Geller et al. 2018). Then, we normalize by the control group mean and standard deviation within each standard (grade). Standard errors in parentheses. Low Attendance refers to attendance that is less than the control group median and high attendance refers to attendance that is equal to or greater than the control group median. For baseline attendance, we use number of days the student was present in school in the previous week as reported by parents, with a control group median of 5 days. For endline attendance, we use classroom attendance rate as recorded during classroom observations at endline, with a control group median of 66.5%. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Another concern expressed by programme implementers was dilution of treatment effects due to large class sizes. Teachers might have been stretched too thin to pay attention to all students and their needs. To test this hypothesis, we split the sample according to the control group median student-teacher ratio from the endline classroom observation, which was 74. Students in high student-teacher ratio classrooms performed worse on the endline EGMA, as expected (control group means reported in Table 16). However, treatment effect estimates were nearly identical between the two groups, 0.181 and 0.188 for the low and high student-teacher ratio groups, respectively. We therefore fail to find evidence that high student-teacher ratios are associated with smaller treatment effects.

Table 16: Results by Student-Teacher Ratio Category

	Outcome: overall EGMA score (z)	
	(1)	(2)
	Low Student-Teacher Ratio	High Student-Teacher Ratio
Treatment	0.181* (0.099)	0.188** (0.076)
Mean (Control)	0.077	-0.059
N	3,661	3,640

This table reports the main effects by student-teacher ratio. We calculate the overall EGMA score as the mean of the two normalized composite scores. We calculate composite scores separately for timed (correct items per minute) and untimed sections, following recommended practice based on the psychometric properties of each composite score (Geller et al. 2018). Then, we normalize by the control group mean and standard deviation within each standard (grade). Standard errors in parentheses. Low Student-Teacher Ratio Group refers to student-teacher ratio that is less than the control group median and High Student-Teacher Ratio Group refers to student-teacher ratio that is equal to or greater than the control group median. We compute student-teacher ratios using class enrolment figures as reported by head teachers and number of teachers assigned to the entire class as reported by teachers in the teacher survey. The control group median used for the cutoff is a student-teacher ratio of 74. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

A student's household circumstances might also influence their ability to benefit from the programme. To test this, we calculate each student's Poverty Likelihood Index (PPI), which aggregates scores from the baseline parent survey on 10 household asset and characteristic indicators based on a predetermined statistical model to estimate the probability of a household falling below Malawi's national poverty line. We then split the sample into terciles based on the sample distribution of PPI. We find the middle and least poor terciles saw increased numeracy from the NNP (Table 17). For the poorest tercile, the treatment effect estimate is not

statistically distinguishable from zero. These results suggest the programme disproportionately benefitted students from relatively more advantaged households.

Table 17: Main Effects on numeracy by Poverty Tercile

	Outcome: overall EGMA score (z)		
	Least poor Tercile	Middle Tercile	Poorest Tercile
Treatment	0.23*** (0.07)	0.29*** (0.07)	0.05 (0.08)
N	2,814	2,932	2,493
Control Mean	0.08	-.08	0.00
Poverty likelihood range	[0.5, 52.6]	[55, 74.2]	[76, 93.6]

This table reports the main effects of the NNP on overall EGMA score (z) by the baseline poverty tercile. “Bottom Tercile” means least poor, and “Top Tercile” means poorest. Poverty likelihood is calculated using the Poverty Likelihood Index (PPI), which aggregates scores from 10 household asset and characteristic indicators based on a predetermined statistical model to estimate the probability of a household falling below Malawi’s national poverty line. We calculate the overall EGMA score as the mean of the two normalized composite scores. We calculate composite scores separately for timed (correct items per minute) and untimed sections, following recommended practice based on the psychometric properties of each composite score (Geller et al. 2018). Then, we normalize by the control group mean and standard deviation within each standard (grade). Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1.

Finally, we explore whether programme impacts differ between schools with and without school feeding programs. We find positive point estimates regardless of school feeding programme availability (Table 18). However, treatment effects are larger (0.25 sd) and statistically significant only for schools with a feeding programme. This is consistent with an enabling effect of school feeding, perhaps through encouraging attendance or improving student readiness to learn. However, we cannot conclude the school feeding programme has a causal effect, as the school feeding programme is not randomly assigned.

Table 18: Results by School Feeding Status

	Outcome: overall EGMA score (z)	
	(1)	(2)

	<u>School feeding programme?</u>	
	No	yes
Treatment School	0.080 (0.086)	0.251** (0.119)
Mean (Control)	0.099	-0.142
N	4,293	3,901

This table reports the main effects of the NNP on overall EGMA score (z) by school feeding programme status. We calculate the overall EGMA score as the mean of the two normalized composite scores. We calculate composite scores separately for timed (correct items per minute) and untimed sections, following recommended practice based on the psychometric properties of each composite score (Geller et al. 2018). Then, we normalize by the control group mean and standard deviation within each standard (grade). Standard errors (clustered by strata) in parenthesis. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

4.4. Programme participation and teaching quality

In this section we report on the quality of teaching and the implementation of the curriculum.

The intervention had a large effect on participation in teacher training, with a treatment effect of 0.68, significant at 1% (Table 19). Compared to the control group mean of 0.05 meaning only 5% of the teachers in the control group participated in teacher training, this result indicates a strong uptake of training activities, such as in-person training sessions and participation in teacher learning circles, in the treatment group, suggests that the intervention successfully motivated or enabled teachers to engage in professional development opportunities focused on improving mathematics instruction. Results also demonstrate minimal contamination by the control group in NNP training activities.

Table 19: Treatment Effects on participation in teacher training

	Outcome: Take up of teacher training
Treatment Group	0.68*** (0.03)
N	542
Control group mean	0.05

This table reports treatment effects on the take-up of teacher training. Results from teacher surveys. The take-up of teacher training is measured as the mean of dummy variables for responses to the following questions: "Did you receive in-person training in mathematics instruction at least once in the past academic year?" (1 = "Yes"; 0 otherwise); and "Did you participate in mathematics teacher learning circles at least three times this year?" (1 = "Yes"; 0 otherwise). Standard errors in parenthesis. Significance: ***=.01, **=.05, *=.1.

The participation in teacher training is measured as the mean of dummy variables for responses to the following questions:

1. "Did you receive in-person training in mathematics instruction at least once in the past academic year?" (1 = "Yes"; 0 otherwise)
2. "Did you participate in mathematics teacher learning circles at least three times this year?" (1 = "Yes"; 0 otherwise).

Table 20 shows student attendance rates during classroom observations, broken down by gender and treatment status. Overall, attendance is slightly higher in control schools (66%) than in treatment schools (64%), with an average of 65%. Girls consistently have a higher attendance rate (68%) in both control and treatment schools, while boys show a lower overall rate (64%), with minor differences between control (65%) and treatment schools (63%). The small and consistent standard errors indicate precise estimates, highlighting minimal differences between groups and slightly higher attendance for girls compared to boys.

Table 20: Classroom Observation Student Attendance Rate

	Overall	Control	Treatment
Girl	0.68 (0.01)	0.68 (0.02)	0.68 (0.01)
Boy	0.64 (0.01)	0.65 (0.02)	0.63 (0.02)
Overall attendance	0.65 (0.01)	0.66 (0.01)	0.64 (0.01)
N	287	143	144

This table reports student attendance rate as recorded during classroom observation for the control and treatment schools and by gender. Attendance rate for each category is calculated as the proportion of enrolled students in that category who are present at the time of observation. Standard errors in parenthesis.

In assessing the implementation fidelity of the curriculum in treatment schools, we find high overall compliance with NNP practices, especially in integrating the workbooks into lesson plans and facilitating independent learning.

Table 21: Compliance with NNP Classroom Practices: Treatment schools only

Variables	Yes
Was the lesson plan informed by the schemes of work?	86.2%
Was the lesson plan informed by the learner workbook and/or the teacher guide?	92.5%
Did the teacher use learner workbooks during the lesson?	94.5%
Did the teacher effectively use manipulatives that are appropriate for the page	69.9%
In the teacher-led activity, the activity then leads to the independent work of the learner	95.9%
Did the teacher set tasks from the workbook that the learners must work on independently?	95.2%
Did the teacher lead reflection on the tasks/activities with the learners?	72.6%
N	145

This table reports the proportion of observed treatment school classrooms where observed practices are consistent with NNP guidelines.

In addition, we observed the use of the NNP workbooks for ten randomly selected students per observed class within treatment schools, as a measure of the extent of the curriculum implementation. Table 22 summarizes progress in the use of NNP workbooks. On average, 51% of pages had all exercises attempted, suggesting only a slight majority of workbook content was being actively utilized. Moreover, an average of 28% of the workbook pages were completely empty, indicating that some topics may not have been covered. Furthermore, in the average workbook, 13% of pages had only one exercise attempted, reflecting partial engagement with some workbook tasks.

Table 22: Workbook Progress: Treatment Schools Only

Variable	Mean
proportion of completely empty pages in the workbook	0.28
proportion of workbook pages with all exercises attempted	0.51
proportion of workbook pages with only one attempted exercise	0.13
N	145

This table reports progress on the use of NNP workbooks as noted during classroom observations in treatment schools. Data is from a random sample of 10 workbooks per observed class.

Comparisons with the control reveal that the NNP effectively changed the mathematics curriculum in the treatment schools. We created a curricular implementation index consisting of the mean of six indicator variables, such as receiving new mathematics workbooks this academic year and having enough workbooks for all learners. The index ranges from 0-1, with 1 representing full implementation. The treatment group shows an increase of 0.60 in the curriculum implementation index compared to the control group, or five times the control mean, significant at 1%.

Table 23: Treatment Effects on Basic Implementation

Outcome: curriculum implementation index	
Treatment Group	0.60*** (0.04)
N	150
Control group mean	0.12

This table reports the effects of the treatment on basic implementation of the curriculum. Results from head teacher surveys. Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1. NB: The basic implementation index ranges from 0-1, calculated as the mean of the following indicator variables: (I) learners received new mathematics workbooks this academic year. (II) workbooks made available every term. (III) enough workbooks for all learners (IV) Strongly agree: Teachers had the materials they needed to teach mathematics effectively. (V) Strongly agree: Teachers had ongoing training and support to improve their mathematics teaching skills. (VI) Strongly agree: Teachers were able to implement the mathematics curriculum well.

To evaluate the impact of the NNP on teaching quality, we developed a teaching quality index ranging from 0 to 1 reflecting various dimensions of effective teaching, including articulating lesson objectives, linking activities to objectives, encouraging critical thinking, assessing learner progress, and facilitating learner practice, as measured in classroom observations. The index ranges from 0-1, with 0 indicating ratings of “not effective” on all index items and 1 indicating “effective” on all index items. As shown in Table 24, the NNP had a significant positive impact on teaching quality in treatment schools. Specifically, the treatment group showed an increase of 0.12 in the teaching quality index compared to the control group, which is statistically significant at 1%. This demonstrates that the intervention effectively enhanced key aspects of teaching practice.

Table 24: Teaching Quality

Outcome: teaching quality index (0-1)	
Treatment Group	0.12*** (0.03)
N	290
Control group mean	0.66

This table reports the effects of treatment on teaching quality. Results from classroom observations. The unit of observation is a single classroom observation. Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1. NB: The teaching quality index ranges from 0-1, calculated as the Mean of scoring: 0 = not effective, 0.33 = satisfactory, 0.66 = partially effective, 1 = effective in the following: (I)The teacher explicitly articulates the objectives of the lesson. (II) The teacher helps learners to apply the skills and strategies they have learned in order to solve the problems in their learner workbook. (III) The teacher relates classroom activities to the objectives. (IV) the teacher draws attention to the skills that link activities on a page and reminds learners of learning from previous lessons. (V) the teacher asks learners to explain their answers and uses this to recognize when an incorrect answer is due to a simple calculation mistake or a misconception. (VI) The teacher asks questions that call for critical thinking. (VII) The teacher's explanation of content is clear. (VIII) the teacher is always assessing the progress of learners and adjusts the teaching if he/she finds that the learners are not making progress. (IX) The children have opportunity to practice at least twice during the class

4.5. Attitudes towards mathematics

In this section we report on how the programme influenced attitudes towards mathematics among students, head teachers, and teachers.

Tables 27 to 31 present the effects of the treatment on students' attitudes towards learning mathematics in the full sample and across different grade levels. The student attitudes are measured as the mean of dummy variables for responses to the following questions:

1. "How much did you enjoy your mathematics class this school year?" (1 = "Enjoyed it a lot", 0 otherwise)
2. "Is Mathematics hard?" (1 = "No"; 0 otherwise)
3. "Do you like explaining to your friends and family how you solved a math problem?" (1 = "Yes", 0 otherwise).

Overall, the treatment effects on student attitudes towards learning mathematics are small in magnitude. None of the point estimates is statistically distinguishable from zero.

Table 25: Treatment effects on student attitudes towards the learning of mathematics

Outcome: Student Attitude	
Treatment Group	-0.01 (0.01)
N	8,227

This table reports treatment effects on the student's attitude towards the learning of mathematics. The student attitudes are the mean of the following indicator variables: 1) "Enjoyed it a lot" in response to, "How much did you enjoy your mathematics class this school year?"; 2) "No" in response to, "Is Mathematics hard?"; 3) "Yes" in response to, "Do you like explaining to your friends and family how you solved a math problem?" Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1.

Table 26: Treatment effects on student attitudes (standard 1) towards the learning of mathematics

Outcome: Student Attitude	
Treatment Group	-0.02 (0.02)
N	2,034
Control group mean	0.61

This table reports treatment effects on the student's attitude in standard 1 towards the learning of mathematics. The student attitudes are the mean of the following indicator variables: 1) "Enjoyed it a lot" in response to, "How much did you enjoy your mathematics class this school year?"; 2)

"No" in response to, "Is Mathematics hard?"; 3) "Yes" in response to, "Do you like explaining to your friends and family how you solved a math problem?" Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1.

Table 27: Treatment effects on student attitudes (standard 2) towards the learning of mathematics

Outcome: Student Attitude	
Treatment Group	-0.03 (0.02)
N	2,034
Control group mean	0.72

This table reports treatment effects on the student's attitude in standard 2 towards the learning of mathematics. The student attitudes are the mean of the following indicator variables: 1) "Enjoyed it a lot" in response to, "How much did you enjoy your mathematics class this school year?"; 2) "No" in response to, "Is Mathematics hard?"; 3) "Yes" in response to, "Do you like explaining to your friends and family how you solved a math problem?" Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1.

Table 28: Treatment effects on student attitudes (standard 3) towards the learning of mathematics

Outcome: Student Attitude	
Treatment Group	-0.01 (0.02)
N	2,050
Control group mean	0.79

This table reports treatment effects on the student's attitude in standard 3 towards the learning of mathematics. The student attitudes are the mean of the following indicator variables: 1) "Enjoyed it a lot" in response to, "How much did you enjoy your mathematics class this school year?"; 2) "No" in response to, "Is Mathematics hard?"; 3) "Yes" in response to, "Do you like explaining to your friends and family how you solved a math problem?" Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1.

Table 29: Treatment effects on student attitudes (standard 4) towards the learning of mathematics

Outcome: Student Attitude	
Treatment Group	0.02 (0.02)
N	2,058
Control group mean	0.78

This table reports treatment effects on the student's attitude in standard 4 towards the learning of mathematics. The student attitudes are the mean of the following indicator variables: 1) "Enjoyed it a lot" in response to, "How much did you enjoy your mathematics class this school year?"; 2) "No" in response to, "Is Mathematics hard?"; 3) "Yes" in response to, "Do you like explaining to your friends and family how you solved a math problem?" Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1.

The head teacher attitudes are measured as the mean of dummy variables for responses to the following questions:

1. "In my school this academic year, mathematics education placed core emphasis on problem-solving and sense-making rather than procedures" (1 = "Strongly agree"; 0 otherwise)
2. "In my school this academic year, teachers incorporated real-world problem-solving and practical applications into their mathematics lessons in standards 1 to 4" (1 = "Strongly agree"; 0 otherwise)
3. "Learners made good progress in their understanding of mathematics" (1 = "Strongly agree"; 0 otherwise)
4. "Learners enjoyed mathematics more than they did at the beginning of the academic year" (1 = "Strongly agree"; 0 otherwise)

The treatment significantly improved head teachers' attitudes toward mathematics teaching and learning (Table 30). The treatment effect on head teacher attitudes is 0.30, significant at 1%. This represents a substantial positive shift relative to the control group mean of 0.29. Summed together, head teachers in the treatment group had an average of 0.59 (on a continuous 0-1 scale) of this index. The results suggest that the intervention meaningfully influenced head teachers to prioritize more engaging and problem-oriented approaches to mathematics education.

Table 30: Treatment Effects on Head Teacher Attitudes towards Mathematics

Outcome: head teacher attitudes towards mathematics	
Treatment Group	0.30*** (0.07)
N	150
Control group mean	0.29

This table reports treatment effects on the head teacher's attitude towards mathematics teaching and learning. Results from Head-teacher surveys. head teacher attitudes are measured as the mean of dummy variables for responses to the following questions:

1. "In my school this academic year, mathematics education placed core emphasis on problem-solving and sense-making rather than procedures" (1 = "Strongly agree"; 0 otherwise)
2. "In my school this academic year, teachers incorporated real-world problem-solving and practical applications into their mathematics lessons in standards 1 to 4" (1 = "Strongly agree"; 0 otherwise)
3. "Learners made good progress in their understanding of mathematics" (1 = "Strongly agree"; 0 otherwise)
4. "Learners enjoyed mathematics more than they did at the beginning of the academic year" (1 = "Strongly agree"; 0 otherwise)

Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1.

Table 31: Treatment Effects on teacher attitudes towards the learning of mathematics

Outcome: teacher attitudes towards mathematics learning	
Treatment Group	0.02 (0.04)
N	542
Control group mean	0.48

This table reports treatment effects on the teacher's attitude towards the learning of mathematics. Results from teacher surveys. Teacher attitudes are measured as the mean of dummy variables for responses to the following questions: 1) "Should mathematics education place core emphasis on problem-solving and sense-making rather than procedures?" (1 = "Strongly agree"; 0 otherwise); 2) "Can all students learn and excel in mathematics regardless of their background?" (1 = "Strongly agree"; 0 otherwise). Standard errors (clustered by strata) in parenthesis. Significance: ***=.01, **=.05, *=.1.

Teacher attitudes are measured as the mean of dummy variables for responses to the following questions:

1. "Should mathematics education place core emphasis on problem-solving and sense-making rather than procedures?" (1 = "Strongly agree"; 0 otherwise)
2. "Can all students learn and excel in mathematics regardless of their background?" (1 = "Strongly agree"; 0 otherwise)

The treatment did not significantly affect teacher attitudes toward mathematics learning, as evidenced by a treatment effect of 0.02, which is small relative to the control mean of 0.48 and not statistically significant. This indicates that while the intervention may have influenced other areas, it did not substantially alter teachers' beliefs about problem-solving, sense-making, or the potential for all students to excel in mathematics.

5. Discussion

This study provides robust evidence of the significant impact of the NNP on numeracy. The estimated treatment effect on overall EGMA scores, at 0.19 standard deviations, is noteworthy, particularly when contextualized within global benchmarks. For example, Lortie-Forgues and Inglis (2019) report an average effect size of only 0.06 SD in a meta-analysis of 141 education RCTs. Similarly, Evans and Yuan (2022) provide a more contextually relevant comparison, analyzing RCTs conducted in 51 low- and middle-income countries between 1981 and late 2017. Across 199 numeracy effect sizes from 70 studies, they found the 80th percentile effect size to be 0.17

SD. Furthermore, they note that effect sizes have remained stable since 2006, suggesting that more recent interventions are unlikely to diverge significantly from this trend.

The moderate numeracy gains observed in standard 2 compared to other grades were anticipated. The standard 2 curriculum implemented during the extended pilot phase remained unrevised from the initial pilot, where challenges had been identified, potentially limiting the programme's immediate impact for this grade.

The overall positive effects of the NNP on numeracy mask important dimensions of heterogeneity. Not all students benefited from the intervention. We find no evidence that the students who scored in the bottom third of the baseline numeracy assessment gained from the programme. Similarly, we find no evidence that the poorest third of students gained from the programme. One potential explanation is low school attendance among these groups, giving them less exposure to the NNP than other students. Another potential explanation is that these students were disproportionately concentrated in schools with large class sizes, making it difficult for the teacher to focus on struggling students. We fail to find evidence for either of these explanations using available data. Schools with feeding programs had larger NNP impacts, consistent with an enabling role for school feeding. However, we cannot interpret this relationship causally, as school feeding programs are not randomly assigned. Unfortunately, we may simply lack good ways to measure whether low attendance or limited school resources explain why some students benefitted less from the programme.

The programme demonstrated notable success in enhancing teacher engagement, curriculum implementation, and teaching quality. Teachers in treatment schools actively participated in professional development activities, including in-person training sessions and mathematics teacher learning circles, indicating that the intervention effectively motivated and supported teachers in their professional growth. Furthermore, the implementation of the curriculum in treatment schools was marked by strong adherence to prescribed practices, such as integrating learner workbooks into lesson plans and encouraging independent learning. Observations revealed that teachers used the provided materials and tools to align classroom activities with the curriculum's goals. The intervention also fostered significant improvements in teaching quality, as evidenced by enhanced instructional practices. These outcomes suggest the program's positive impact on instructional practices and the delivery of mathematics. Nonetheless, there is room to improve implementation fidelity, such as making full use of learner workbooks to complete practice exercises.

The treatment had varied effects on student and educator attitudes towards mathematics. While the treatment did not impact student or teacher attitudes towards learning mathematics, it significantly changed head teachers' attitudes. These head teachers

were more likely to emphasize problem-solving and real-world applications, and they observed positive progress and enjoyment in their students' understanding of mathematics. Additionally, the treatment led to a substantial increase in teachers' participation in professional development activities, such as in-person training and teacher learning circles. However, it had a minimal impact on teachers' attitudes towards mathematics learning. This suggests that while professional engagement increased, beliefs about mathematics education remained relatively unchanged.

The RCT sample was limited to three of Malawi's 34 districts. To what extent are our findings generalizable? This question is of great importance, given the scope of the NNP extended pilot and plans for further scale-up. We offer two directions to address this question going forward.

First, we can re-estimate the programme's impact with statistical adjustments for external validity. Specifically, we can use observed characteristics of schools from the Malawi Education Management Information System (EMIS) to reweight the RCT sample for observational similarity to a broader population of schools (such as schools in the extended pilot, or all primary schools). We plan to do this in the academic journal version of this evaluation.

Second, we can aim for a better understanding of how programme implementation in the RCT may have differed from the extended pilot. In principle, the extended pilot and RCT versions of the NNP were identical. Nonetheless, differences in geographic locations or other factors may have generated different implementation approaches or quality in practice. The implementing partners at Cambridge Education, FCDO, and the Ministry of Education can help shed light on these differences.

6. Recommendations and Conclusion

The National Numeracy Programme was designed to address the critical issue of low numeracy levels among Malawian children, who consistently lagged behind their regional peers in international assessments such as SACMEQ. This evaluation sought to assess the causal impact of the NNP on both Mathematics learning outcomes and teaching quality in Malawi. The findings demonstrate the programme's success, with a 0.19 standard deviation treatment effect on Mathematics learning, equivalent to 0.56 Malawi Equivalent Years of Schooling (EYOS) and 0.25 Learning Adjusted Years of Schooling (LAYS). These results highlight the potential of the NNP to improve foundational learning outcomes, which are essential for bridging the existing learning gap. On teaching quality, the

evaluation found a significant impact of 0.12 on a composite teaching quality index with a 0-1 scale, or an 18 percent increase over the control mean, further affirming the programme's comprehensive impact.

The NNP has already begun a phased national scale-up, starting with Standards 1 and 2, with plans to expand to Standards 3 and 4. While cost effectiveness analysis is outside the scope of this study, the programme's design aligns with evidence from Angrist et al. (2025), which identifies structured pedagogy programs—combining lesson plans, student materials, teacher training, and monitoring—as among the most cost-effective approaches in education. However, scaling up programs successfully requires careful attention to maintaining implementation fidelity. As Kraft (2020) notes, interventions that depend on coordinated efforts among school personnel, such as those involving professional learning communities and new curricula, often encounter significant challenges during scale-up. The NNP shares these characteristics, and it is essential to ensure that the programme's benefits are not diluted but rather enhanced as it expands to reach more students. Importantly, the NNP has already progressed from a pilot to an extended pilot phase, operating in 1,200 schools across the country alongside the treatment schools during the evaluation period. This experience of scaling from a pilot to an extended pilot provides the Ministry of Education and other stakeholders with valuable insights and institutional knowledge to draw upon as they now scale the programme to the national level. We are optimistic that this experience will help mitigate some of the challenges typically associated with scaling interventions, ensuring smoother implementation and maintaining programme fidelity. We propose several actionable steps to support the NNP's continued success.

Despite high levels of implementation fidelity observed in this evaluation, opportunities remain to improve compliance in certain areas. For example, the appropriate use of manipulatives was recorded at 69.9%, while reflection practices were observed at 72%. Learner workbook coverage also showed significant gaps, which must be addressed to sustain and even enhance programme impacts as it scales. Strengthening pre-service and in-service training, coupled with continuous monitoring using a refined version of the mid-term monitoring tool adapted for national scale, will be essential to ensure consistent and effective implementation across all schools.

Another key finding is that the programme's benefits have not reached all groups equally. The poorest students and those in the bottom third of the performance distribution showed limited gains, underscoring the need for targeted support to bring these students along. There is a negative association between poverty levels and baseline numeracy performance (results in the baseline report of this evaluation), suggesting that addressing poverty-related barriers could enhance learning outcomes for the most disadvantaged children. The MOE can make efforts to build this into the government's overall poverty alleviation efforts, while

identifying other support for these disadvantaged students. For instance, our finding of a significant treatment effect among schools with school feeding programs suggest that these programs could be a catalyst for the NNP. While our evaluation has uncovered these patterns, our data to uncover the mechanisms underlying them is limited and inconclusive. Qualitative work may therefore complement our findings to design more effective strategies to reach all students. Future quantitative work could also seek to improve measurement of student attendance to better capture treatment effect heterogeneity in this dimension.

In conclusion, the NNP has made significant strides in improving mathematics learning outcomes and teaching quality in treatment schools. These achievements provide a strong foundation for its nationwide scale-up and demonstrate the programme's potential to close the numeracy gap with regional peers. However, sustaining and enhancing these gains will require careful planning, political commitment, and targeted efforts to address implementation challenges and equity gaps. The findings of this evaluation underscore the importance of continuing to invest in and adapt the NNP, ensuring that all Malawian children have the opportunity to build strong foundational numeracy skills. Stakeholders at all levels must work collaboratively to preserve and amplify the programme's impact as it expands across the nation.

7. Conflict of Interest Statement

Esme Kadzamira, Lizzie Chiwaula, and Elizabeth Meke, three of the authors of this study, are employed by the Malawi Ministry of Education in the Higher Education department. While the current research focuses on basic education and is not directly funded or implemented by the government, the Ministry of Education supervises this project and the findings will inform future national policy decisions regarding the intervention. All other authors declare that they have no financial or non-financial competing interests related to the research presented in this report.

8. References

- Angrist, N., Evans, D. K., Filmer, D., Glennerster, R., Rogers, H., & Sabarwal, S. (2025). How to improve education outcomes most efficiently? A review of the evidence using a unified metric. *Journal of Development Economics*, 172, 103382. <https://doi.org/10.1016/j.jdeveco.2024.103382>
- Athey, Susan, and Guido W. Imbens. "The econometrics of randomized experiments." In *Handbook of economic field experiments*, vol. 1, pp. 73-140. North-Holland, 2017.
- Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli. "Adaptive linear step-up procedures that control the false discovery rate." *Biometrika* 93, no. 3 (2006): 491-507.
- Brombacher, A. (2011). Malawi Early Grade Mathematics Assessment (EGMA): national baseline report 2010. USAID/Malawi and Ministry of Education, Science and Technology. https://pdf.usaid.gov/pdf_docs/pnaec139.pdf
- Brombacher, Aarnout. 2019. "RESEARCH TO INVESTIGATE LOW LEARNING ACHIEVEMENT IN EARLY GRADE NUMERACY (STANDARDS 1–4) IN MALAWI: The Victory of Form over Substance." Oxford, UK: HEART.
- Chaisemartin, Clément de, and Jaime Ramirez-Cuellar. 2024. "At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments?" *American Economic Journal: Applied Economics* 16 (1): 193–212. <https://doi.org/10.1257/app.20210252>.
- Evans, D. K., & Yuan, F. (2022). How Big Are Effect Sizes in International Education Studies? *Educational Evaluation and Policy Analysis*, 44(3), 532–540. <https://doi.org/10.3102/01623737221079646>

Geller, Leanne R. Ketterlin, Lindsey Perry, Linda M. Platas, and Yasmin Sitabkhan. 2018. "Aligning Test Scoring Procedures with Test Uses: A Balancing Act." *Global Education Review* 5 (3): 143–64.

Kazima, M., Longwe, J. and Gobede, F. (2022) Improving teaching and learning mathematics in Malawi Primary schools: A Review of reforms, Interventions, successes and challenges. *Journal of Applied Learning 7 Teaching*, Vol. 5 Special Issue no.2, pp81-89

Kline, R. B. (2009). *Becoming a behavioral science researcher : a guide to producing research that matters*. Guilford Press.

Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale RCTs are often uninformative: Should we be concerned? *Educational Researcher*, 48(3), 158–166.

Malawi Institute of Education (2023) National Numeracy Programme Facilitator Manual, 2023/24 Expanded Pilot Teacher Training Activity.

Piper, B., Ralaingita, W., Akach, L., & King, S. (2016). Improving procedural and conceptual mathematics outcomes: evidence from a randomised controlled trial in Kenya. *Journal of Development Effectiveness*, 8(3), 404–422.
<https://doi.org/10.1080/19439342.2016.1149502>

School to School International (STS) (2023) Pilot Impact Evaluation Endline Report, National Numeracy programme in Malawi. Ministry of Education/Cambridge education/UKAID

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, 47, 150–174.

UNESCO. (2017). *Reading for All: A Guide to Developing Early Grade Reading Programmes in Low-Income Countries*. UNESCO.

UNICEF. (2021). *The State of the World's Children 2021: On my mind: Promoting, protecting, and caring for children's mental health*. UNICEF.

UNICEF. (2022, March). Are children really learning? Exploring foundational skills in the midst of a learning crisis. UNICEF DATA.
<https://data.unicef.org/resources/are-children-really-learning-foundational-skills-report/>

Wedman, J., & Per-Erik Lyrén. (2015). Methods for examining the psychometric quality of subscores: A review and application. *Practical Assessment, Research & Evaluation*, 20(21), 21-.

9. Appendix

9.1. Examples of mathematics curriculum materials

9.1.1. Control group mathematics curriculum

Figure 7: Cover page of Mathematics Learner textbook for standard 2 (Chichewa language)

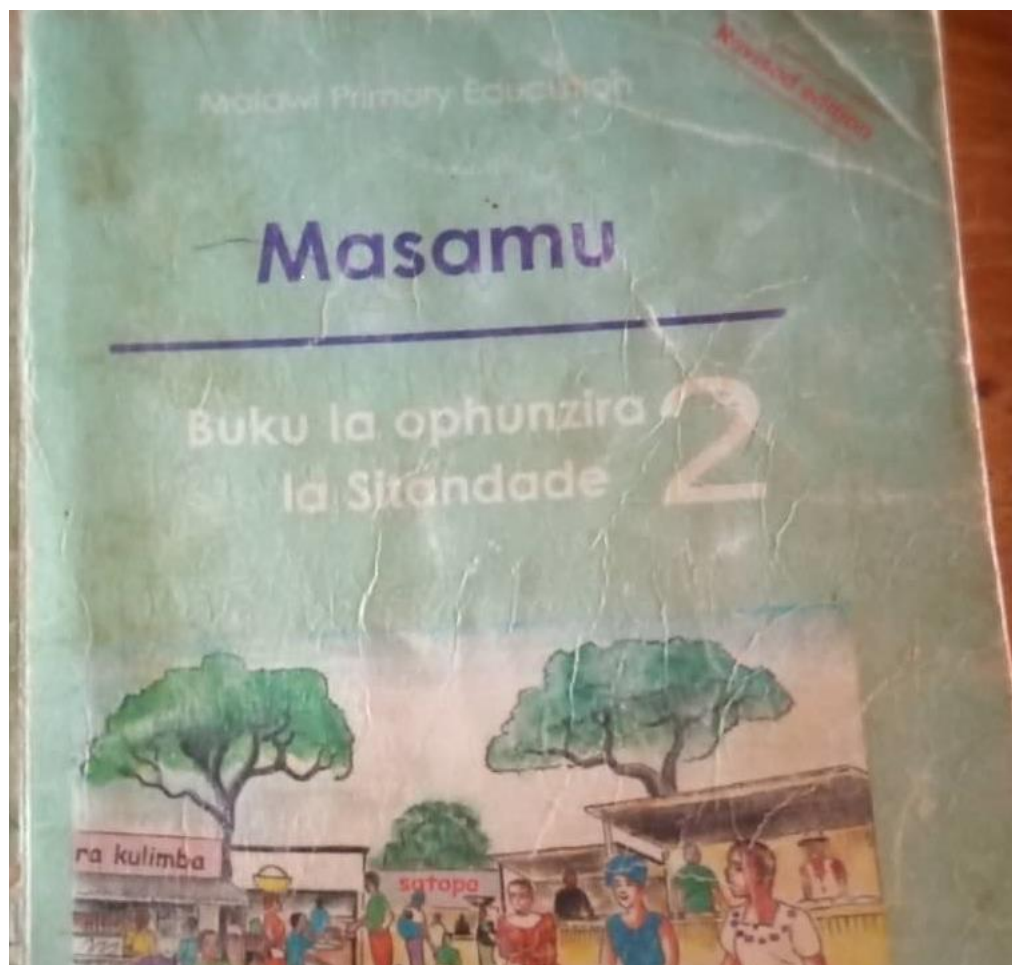


Figure 8: Standard 2, Page 24: Mathematics Exercise on Addition of figures up to 20

Ntchito 2 Kuwonkheisa nambala mpaka 20
Wonkhetsani nambala zotsatirazi.

Chitsanzo	Funso	Yankho
	$\begin{array}{r} 10 \\ + 8 \\ \hline \end{array}$	$\begin{array}{r} 10 \\ + 8 \\ \hline 18 \end{array}$
1	$\begin{array}{r} 13 \\ + 5 \\ \hline \end{array}$	$\begin{array}{r} 6 \\ + 2 \\ \hline \end{array}$
2	$\begin{array}{r} 15 \\ + 1 \\ \hline \end{array}$	$\begin{array}{r} 7 \\ + 7 \\ \hline \end{array}$
3	$\begin{array}{r} 19 \\ + 0 \\ \hline \end{array}$	$\begin{array}{r} 8 \\ + 5 \\ \hline \end{array}$
4	$\begin{array}{r} 17 \\ + 1 \\ \hline \end{array}$	$\begin{array}{r} 9 \\ + 4 \\ \hline \end{array}$
5	$\begin{array}{r} 18 \\ + 1 \\ \hline \end{array}$	$\begin{array}{r} 10 \\ + 7 \\ \hline \end{array}$

Note: The colored part shows an example.

Figure 9: Standard 2 Mathematics exercise on subtraction for figures up to 20

Note: p. 27 of source textbook. The colored part shows an example.

Nichito 2 Kuchotsera nambala mpaka 20
Chotserani nambala zotsatirazi.

Chitsanzo	Funso	Yankho
	$\begin{array}{r} 14 \\ - 3 \\ \hline \end{array}$	$\begin{array}{r} 14 \\ - 3 \\ \hline 11 \end{array}$
1	$\begin{array}{r} 15 \\ - 5 \\ \hline \end{array}$	$\begin{array}{r} 19 \\ - 16 \\ \hline \end{array}$
2	$\begin{array}{r} 17 \\ - 6 \\ \hline \end{array}$	$\begin{array}{r} 18 \\ - 12 \\ \hline \end{array}$
3	$\begin{array}{r} 16 \\ - 4 \\ \hline \end{array}$	$\begin{array}{r} 10 \\ - 10 \\ \hline \end{array}$
4	$\begin{array}{r} 12 \\ - 1 \\ \hline \end{array}$	$\begin{array}{r} 11 \\ - 10 \\ \hline \end{array}$

Figure 10: Standard 2 Mathematics Exercise on Addition of numbers up to 20

MUTU 2 Kuwonkhetsa nambala mpaka 20

Nichito 1 Wonkhetsani nambala zotsatirazi.

Chitsanzo

Funso $9 + 4 =$

Yankho $9 + 4 = 13$

1 $8 + 5 =$	10 $11 + 8 =$
2 $8 + 9 =$	11 $17 + 1 =$
3 $7 + 4 =$	12 $10 + 9 =$
4 $6 + 5 =$	13 $12 + 2 =$
5 $14 + 2 =$	14 $2 + 16 =$
6 $13 + 6 =$	15 $5 + 14 =$
7 $15 + 3 =$	16 $7 + 13 =$
8 $20 + 0 =$	17 $1 + 19 =$
9 $12 + 6 =$	18 $10 + 10 =$

23

Observation: Learners are asked to solve the mathematics on addition on page 23 above.

9.1.2. National Numeracy Programme curriculum

Figure 11: Standard 2 Mathematics Learner workbook for term 2: cover page (English language)

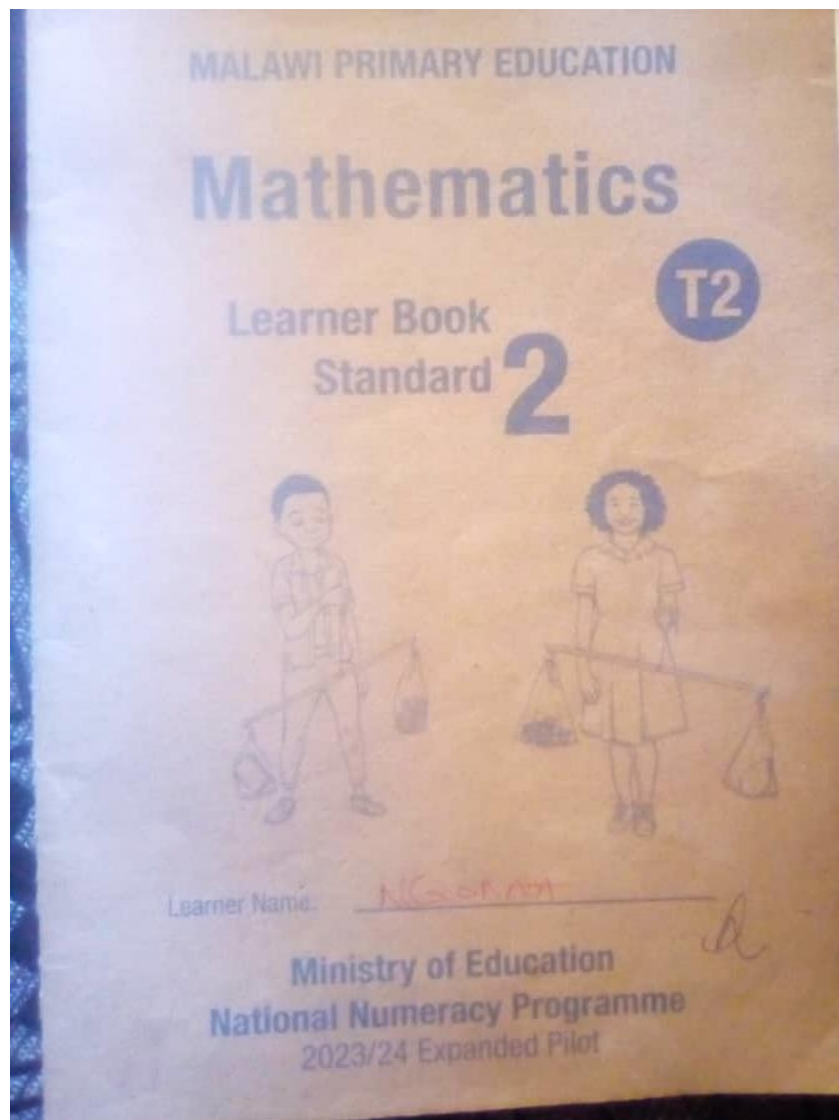
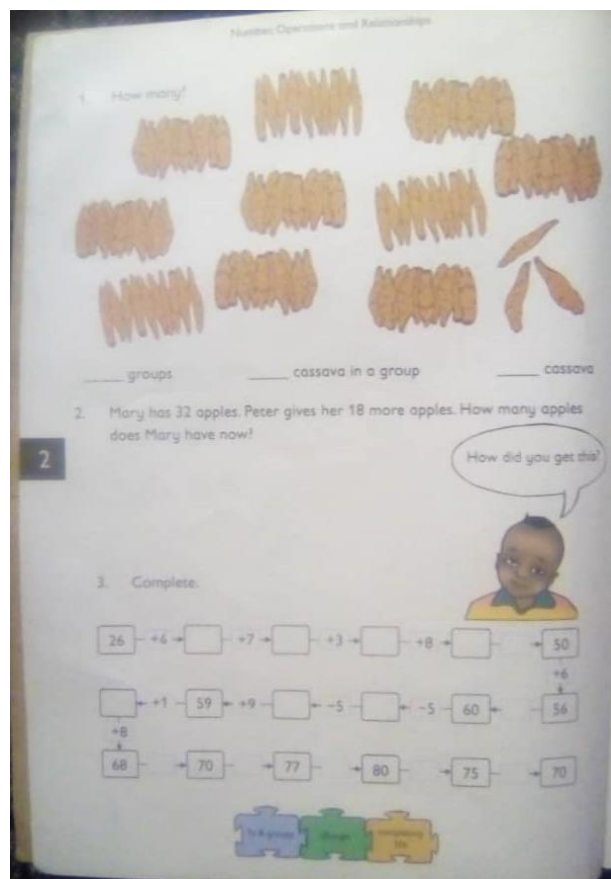


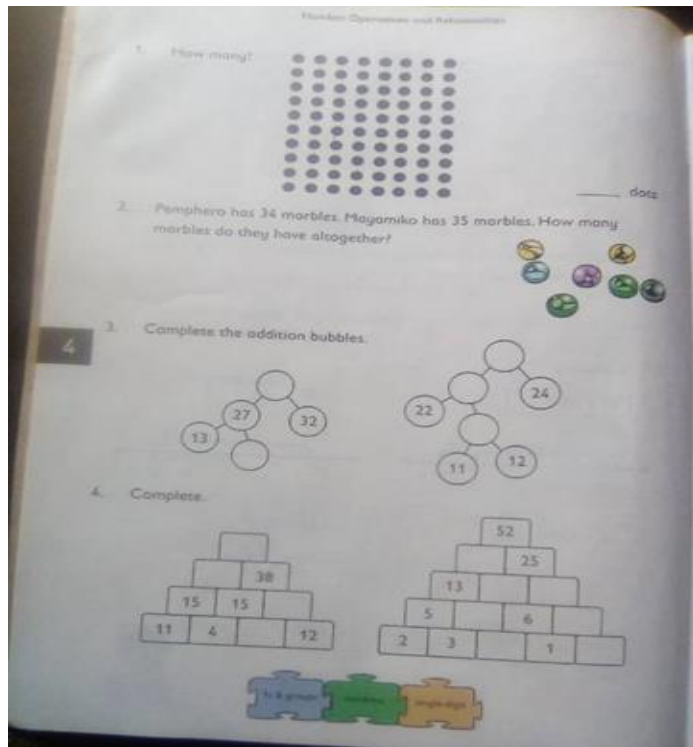
Figure 12: Standard 2 Mathematics Learner workbook for term 2, Page 2: Number Operations and Relationships



Observations:

1. Learners are asked to count a) the groups of cassava, Number of cassava in a group and the total number of cassava.
2. Learners are asked to solve a word problem, which allows them to apply mathematical concepts learnt and encourages reasoning skills.
3. Learners are asked to complete a number chain doing both additions and subtraction.
4. There is a reflection question for learners to explain how they got the answer i.e. "how did you get this?"

Figure 13: Standard 2 Mathematics Learner workbook for term 2, Page 4: Number Operations and Relationships



Observations:

1. Learners are asked to do some counting
2. Learners are asked to solve a word problem
3. Learners are asked to complete the addition bubbles (manipulating numbers)
4. Learners are asked to complete the pyramids (Manipulating numbers).
5. There are puzzle pieces at the bottom of the page (in blue colour, green colour and yellowish colour)

9.2. EGMA Proficiency Benchmarks

In this section, we define EGMA proficiency benchmarks.

No universal set of EGMA proficiency benchmarks exists. The United Nations has defined detailed numeracy benchmarks to measure progress towards Sustainable Development Goal 4 (quality education for all; UNESCO Institute of Statistics, 2020). However, these benchmarks do not align well with the format of EGMA (UNESCO Institute of Statistics, 2023). Our own attempts to map SDG numeracy benchmarks to EGMA reached a similar conclusion. For example, one SDG benchmark for Grade 3 is counting and comparing whole numbers up to 1,000, a value never reached in EGMA.

However, several Sub-Saharan African countries including Ghana, Zambia, and Tanzania, have conducted their own EGMA benchmarking exercises (RTI International 2014a, 2014b, 2015). Building on these standards, we define the following EGMA benchmarks for this study:

Table 32: EGMA Benchmarks for this study

EGMA subtask	Proficiency benchmark
Addition and subtraction Level 1	80%
Missing numbers and Word problems	70%

Source: Author definitions, guided by USAID Benchmarking exercises convened in Ghana, Tanzania, and Zambia. Level 1 consists of addition and subtraction with numbers less than 20.

The missing number subtest consists of simple number sequences from which students must identify the missing number in the sequence. Word problems are addition-subtraction level 1-type problems presented in an applied context.

Our benchmarks do not distinguish by grade level. Although this limits their application to specific standards in our study, it has the advantage of facilitating comparisons of absolute skills across standards and over time.

9.3. A Note on EGMA Score Construction

As described in section 3.5, the Early Grade Mathematics Assessment (EGMA) was used in this study to measure student numeracy outcomes. EGMA consists of eight distinct subtests designed to assess various foundational numeracy skills, each comprising between five and twenty individual items. Among these subtests, three—namely Number Identification, Addition Level 1, and Subtraction Level 1—are administered under timed conditions, generating two types of scores: a speed score, reflecting the number of correct responses per minute, and an accuracy score, indicating the total number of correct items. The remaining five subtests are untimed and exclusively generate accuracy scores.

While there is no universally prescribed methodology for aggregating EGMA subtests, we constructed composite scores to align with established early numeracy conceptual frameworks and guided by methodological conventions detailed in Piper et al. (2016). Two composite scores were developed. The first, Procedural Fluency, aggregates scores from all timed assessments and is reported as the number of correct responses per minute. This construction deliberately preserves the timed dimension of the assessments, emphasizing numeracy fluency, which we posit is crucial both academically and in real-world contexts. The second composite, termed Conceptual Understanding, aggregates the five untimed subtests, expressed as the percentage of correct responses. This score captures higher-order numerical abilities, including tasks such as addition and subtraction (Level 2), number discrimination, missing numbers, and word problems, thus reflecting the focus of the NNP on problem-solving and numerical sense-making.

To verify the robustness and appropriateness of these composite scores, we conducted reliability and distinctiveness analyses using baseline data, collected through identical instruments employed at endline. Reliability was assessed using Cronbach's alpha, a statistical measure of internal consistency which evaluates the extent to which individual items within a composite score measure the same underlying construct. Following Kline (2009), Cronbach's alpha values above 0.7 indicate acceptable reliability, with values

above 0.9 signifying strong reliability. The reliability coefficients for the fluency and conceptual understanding composite scores were 0.83 and 0.84, respectively, indicating moderately strong reliability. This suggests that the items within each composite score consistently measure the intended construct.

Additionally, we evaluated whether each composite score provided distinct information from the other using the approach of disattenuated correlation described by Wedman and Lyren (2015). This technique adjusts observed correlations for measurement error, thereby providing a clearer view of the genuine interrelation between composite scores. According to Sinharay (2010), distinctiveness is sufficient if the disattenuated correlation between composite scores does not exceed 0.8. The disattenuated correlation between our Procedural Fluency and Conceptual Understanding scores was 0.60, indicating substantial distinctiveness and confirming that these two composites measure distinctly separate aspects of numeracy skills.

For the primary analyses, each composite score was standardized relative to the control group's mean and standard deviation within each grade. Subsequently, an overall EGMA score was computed as the mean of these standardized composite scores. This scoring approach implicitly assigns equal weight to Procedural Fluency and Conceptual Understanding composites but inadvertently places disproportionate emphasis on individual subtests within the Procedural Fluency composite, which consists of three subtests, compared to five subtests in Conceptual Understanding. To ensure robustness of our findings, an alternative scoring method was constructed, explicitly assigning equal weights to each individual EGMA subtest, with results presented as part of our sensitivity analyses.

The chosen standardization process inevitably restricts intuitive interpretability since the composite and overall EGMA scores are expressed in standard deviation (SD) units, rather than intuitive metrics such as percentage correct or raw item scores. Nevertheless, expressing outcomes in SD units is standard practice in economics literature due to its facilitation of comparisons across diverse contexts and measurement scales. To enhance interpretability, we further present learning outcomes using two intuitive metrics: Equivalent Years of Schooling (EYOS) and Learning-Adjusted Years of Schooling (LAYS). EYOS translates program impacts into the approximate additional years of schooling required to achieve similar learning gains, providing an intuitive metric for policymakers to appreciate the practical significance of measured outcomes (Evans & Yuan, 2020). LAYS, as proposed by Angrist et al. (2025), adjusts formal schooling years by actual learning outcomes, thus providing an even more realistic measure of educational productivity. Both metrics serve to clarify the practical implications of numeracy improvements, bridging the gap between statistical significance and educational relevance.