Malawi National Numeracy Programme Evaluation

Lontum Nchadze*
Evaewero French*
Esme Kadzamira
Todd Pugatch*
Alperen Acikol*

* Oregon State University
 Center for Education, Training and Research (CERT), University of Malawi

May 3, 2024

Abstract: Despite an increasing focus on foundational literacy and numeracy (FLN) by the international community, numeracy has received much less attention than literacy. Numeracy programs which are successful at scale are particularly hard to find. We evaluate an at-scale foundational numeracy program in Malawi using a randomized controlled trial (RCT). The program shifts the focus of mathematics learning from rote memorization to meaningful problem-solving and application, with teachers receiving training and materials to support this change. Using a matched school pairs design, we randomly assigned half of a sample of 150 government primary schools to the program (treatment group). The control group continued to receive the traditional curriculum. Treatment schools receive the same program concurrently with 1,100 schools in an expanded pilot outside the RCT sample. This evaluation therefore represents a scaled curriculum reform, and informs plans for further scaling the program nationally. Our baseline results confirm balance on numeracy and observable characteristics between treatment and control groups. We also find low baseline levels of numeracy, with 37% of third graders unable to perform two-digit addition or subtraction. Endline data collection will take place in June 2024, with results ready to report at the What Works Hub for Global Education Annual Conference.

Keywords: foundational skills; numeracy; curriculum reform; randomized controlled trials; Malawi

JEL codes: I21; I25; O22

Corresponding author: Lontum Nchadze, Oregon State University, nchadzel@oregonstate.edu.
This study is registered in the American Economic Association RCT Registry (AEARCTR-0011964; Pugatch et al., 2023). We thank Lizzie Chiwaula, Elizabeth Meke, and Chikondi Sepula for their many contributions to this research. 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, Joshua Wakabi, and Arianna Zannolini.

1 Introduction

Foundational literacy and numeracy (FLN) levels among students in low and middle income countries (LMICs) are alarmingly low. In Malawi, the setting of our study, more than 80% of students at the end of grade 2 were unable to read a single familiar word such as *the* or *cat*, or perform two-digit subtraction (World Bank 2018, p. 5). This FLN deficit has drawn increasing attention from the international community (Evans and Hares 2021). The World Bank refers to the LMIC "learning crisis" and "learning poverty" (World Bank 2018); the United Nations adopted a Sustainable Development Goal 4 (education) target to "ensure that all youth…achieve literacy and numeracy" (United Nations 2016); and a Gates Foundation official wrote an influential essay titled, "The pathway to progress on SDG 4 requires the global education architecture to focus on foundational learning and to hold ourselves accountable for achieving it" (Beeharry 2021). Yet despite this increasing focus on FLN, numeracy has received much less attention than literacy.[1]

The challenge of scale compounds the difficulties of addressing FLN deficits. Programs effective at small scale or implemented by capable NGOs often fail when scaled or implemented by governments (Muralidharan and Niehaus 2017; Bold et al. 2018; Vivalt 2020; List 2022). Foundational numeracy programs are no exception. An exhaustive search convened by RTI International among 60 organizations (including NGOs, foundations, bilateral agencies, ministries of education, universities, and think tanks) yielded 28 candidate numeracy programs, of which only six met the criteria for causal impact at scale (RTI International 2023b).

We evaluate an at-scale foundational numeracy program in Malawi using a randomized controlled trial (RCT). The program, known as the National Numeracy Programme (NNP), focuses on problem solving and applications of mathematics, with accompanying teacher training and support. Using a matched school pairs design, we randomly assigned half of a sample of 150 government primary schools to the program (treatment group). The remaining schools (control group) continued to receive the traditional curriculum, which relied on recall of numerical operations. **Treatment schools receive the same program concurrently with 1,100 schools in an expanded pilot outside the RCT sample. This evaluation therefore represents a scaled curriculum reform, and informs plans for further scaling the program nationally.**

In this extended abstract, we describe the program, methodology, and findings of the baseline survey conducted in October 2023. We confirm balance on baseline characteristics between treatment and control groups. Endline data collection will take place in June 2024, with results ready to report at the What Works Hub for Global Education Annual Conference.

We make three main contributions to the literature. First, we contribute to the scarce evidence on how to improve foundational numeracy in LMICs. Research on foundational literacy is relatively abundant; a 2020 meta-analysis of LMIC literacy interventions included 67 randomized controlled trials or quasi-experimental studies. By contrast, foundational numeracy interventions with credible estimates of causal impact are rare [provide evidence]. Some studies have found positive associations between numeracy and labor market outcomes,

[1] A Google Scholar search for "foundational literacy" returns 2,720 results since 2020, compared to 249 results for "foundational numeracy" (search conducted by authors, April 9, 2024).

but not literacy (D□az, Arias, and Tudela 2012 for Peru; Nikoloski and Ajwad 2014 for Tajikistan).

The few RCTs on foundational numeracy suggest a pedagogical intervention like the program we study could have large effects. In particular, the program's focus on problem solving and understanding the meaning of mathematical concepts could be well suited to the context. In India, children working in informal markets could not solve basic arithmetic problems, but performed well when the problems were reframed as market transactions (A. V. Banerjee et al. 2017). Another study in India found numeracy gains from a game-based preschool mathematics curriculum (Dillon et al. 2017). Notably, the gains were limited to symbolic math and did not persist to the mathematics children later encountered in school. The study most similar to ours evaluated an RCT of a primary school numeracy intervention in El Salvador (Maruyama and Kurosaki 2024). Like the program we study, the intervention focused on problem solving skills, distributed textbooks with the new curriculum, and included teacher training. After one year, the progam increased math scores by 0.49 sd. Collectively, the evidence suggests that programs targeted at helping students derive meaning from mathematics can improve foundational numeracy. We add a new data point to this small sample.

Second, we contribute to the evidence on pedagogical and curricular reforms in LMICs. Pedagogical interventions are among the most successful and cost effective intervention category across several systematic reviews (Evans and Popova 2016; Angrist et al. 2020). Classroom practices including promoting curiosity (Alan and Mumcu 2024), "learning to learn" (Nourani, Ashraf, and Banerjee 2023), and teaching at the right level (Muralidharan, Singh, and Ganimian 2019) can increase learning outcomes. Reorienting curricula towards foundational skills can also increase learning (Rodriguez-Segura and Mbiti 2022). The program we study attempts to reorient mathematics pedagogy along similar dimensions. Nonetheless, efforts to increase student engagement with curricular concepts sometimes fail to increase test scores due to low implementation fidelity or misalignment between pedagogy and test content (Berlinski and Busso 2017; M. P. Blimpo and Pugatch 2021; M. Blimpo and Pugatch 2023; de Barros et al. 2023).

Third, we contribute evidence on LMIC education interventions implemented at scale. The challenges of converting small-scale programs into scaled-up policies – in program design, logistics, and politics, among others – are by now well known (Muralidharan and Niehaus 2017; List 2022). Promoting foundational learning has been noted as a particular example of the scale-up challenge (Glewwe and Muralidharan 2016; Beeharry 2021; Evans and Hares 2021). An early grade remedial program in India required numerous iterations before achieving success at scale (A. Banerjee et al. 2017). A recent effort to identify successfully scaled foundational literacy and numeracy programs developed criteria for scale and effectiveness with input from 60 organizations (RTI International 2023b; 2023a). They found just 14 programs (eight for literacy, six for numeracy) worldwide meeting the criteria. The program we study meets the scale criterion (at least 500 schools, spread across at least two subdivisions). This evaluation will determine whether the program also meets the effectiveness criterion.

2 Intervention and research design

2.1 Context

Malawi is a small landlocked country in Southern Africa bordering Tanzania, Zambia, and Mozambique. Malawi remains one of the poorest countries in the World with a GDP (in current price) of US$645.2 well below the Sub-Saharan Africa average of US$1701.2 in 2022 (https://data.worldbank.org/indicator/NY.GDP.PCAP.CD). A large proportion (72%) of the population lives below the poverty line of $2.15 a day (data from 2024; https://www.worldbank.org/en/country/malawi/overview). Malawi's economy is heavily dependent on rainfed agriculture which employs 80% of the population but is extremely vulnerable to climatic shocks.

Malawi's education follows an 8-4-4 system comprising 8 years of primary, 4 years of secondary, and up to 4 years of tertiary. Primary education is compulsory and free. Access to primary education is universal with gross enrolment rates exceeding 100% since 1994, when free primary education was introduced. Nevertheless, recent estimates suggest that nearly 10% of the eligible school-age population (6-13 years old) is not enrolled (Ministry of Education, 2023)[1]. Primary education faces some serious challenges that have undermined the quantitative gains that have been achieved. These include low internal efficiency manifested by high repetition and dropout rates, low completion rates, a lack of progression of students through the system, poor quality of schooling, and low learning outcomes. These problems are more pronounced in lower primary. Repletion rates for example, have remained highest in the first grade, 36% in standard 1 compared to 18% in standard 8 in the 2022/23 school year (Ministry of Education 2023). As a result, primary enrolments are concentrated in the first four grades (about two-thirds), a situation that has hardly changed over the past four decades. There is a persistent learning crisis with evidence that children are progressing through the primary without mastering foundational skills. A recent longitudinal study revealed the extent of learning poverty with 78% of grade 4 students not able to read a simple text with any understanding- an indication of the learning poverty (Asim & Ravender, 2021[2]).

A 2019 scoping study of primary school mathematics instruction in Malawi concluded, "the current dominant enacted pedagogy and learning environment does not support effective learning. The focus on routines and procedures without any attention to understanding, application, and reasoning is of concern… students are not able to apply their mathematical knowledge in a meaningful way" (Brombacher 2019, p. 4). In response, the Malawi Ministry of Education, with technical support from Cambridge Education (a UK-based private firm) and funding from the UK Foreign, Commonwealth and Development Office (FCDO), developed the National Numeracy Programme (NNP).

2.2 Intervention

The National Numeracy Programme seeks to shift mathematics instruction from the traditional paradigm of rote learning to help children understand the mathematics they learn, develop problem-solving and reasoning skills, and apply the mathematics they learn in real life. The programmme consists of a revised mathematics curriculum for standards 1 to 4,

accompanying learning materials, and teacher training. The teacher training includes both initial training before the start of the academic year, and ongoing support through periodic coaching and teacher learning circles (TLCs). The NNP focuses on developing student strategies to solve mathematics problems and understanding the reasoning underlying these strategies, with the eventual goal of solving unfamiliar problems.

The curriculum includes specific child-led activities and teacher-led activities for each topic covered, with the aim of making the teaching and learning of mathematics engaging. Students learn new strategies, including the use of number chains, pyramids, flow diagrams, number lines, and tables, as well as new topics such as data handling, among others (NNP facilitator Manual, 2023/24). Furthermore, each lesson includes a reflection session, where learners are asked to clarify how they came up with solutions to the mathematics problems or are asked to respond to reflection questions 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 the NNP. Prior to the NNP, most schools lacked sufficient materials for each student. The NNP aims to provide each learner with their own mathematics workbook, refreshed each academic 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.

Appendix 7.1 demonstrates differences between the traditional mathematics curriculum and the National Numeracy Programme curriculum. Figure 3-Figure 5 show pages from the textbook of the traditional curriculum. Each page lists a series of arithmetic operations to solve, without context. Figure 7-Figure 8 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?").

The theory of change underlying the intervention begins with NNP inputs, including teacher training and learning materials; to intermediate outcomes (changed pedagogy and better teaching quality); to learning outcomes. Figure 1 shows the theory of change.

The NNP launched in academic year 2021-2022 as a pilot project in 200 schools. A non-experimental evaluation found learning gains of 0.3 sd in Standards 3-4, with no impacts in other standards (School-to-School International 2023). An "expanded pilot" launched in 2022-2023 scaled the program to an additional 1,100 schools. The expanded pilot included revised learning materials and added teacher learning circles to the training. The expanded pilot continued in 2023-2024, in anticipation of national scale-up of the NNP in 2024-2025.

The pilot schools were chosen by the Ministry of Education. We evaluate an RCT of the same version of NNP as the expanded pilot, but in a different set schools. Our evaluation therefore represents an intervention which has already scaled to more than 1,000 schools, and informs plans for further scaling the program nationally.

2.3 Research design

We evaluate the Malawi National Numeracy Programme using a cluster randomized controlled trial (RCT), with schools as the unit of treatment in a matched pair design. The evaluation includes three districts not exposed to the NNP expanded pilot. Within these

districts, we defined eligible schools as schools outside zones (the administrative unit immediately below districts in Malawi) in the initial NNP pilot and outside the sampling frame of other potential confounding activities. We then drew a random sample of 150 eligible schools, split evenly across districts.

From this sample of 150 schools, we formed pairs, matched by district and baseline school characteristics. We randomly assigned schools in each pair to treatment or control. Treatment schools received the NNP curriculum, while control schools received the status quo mathematics curriculum in place before the NNP. We described both curricula in the previous subsection.

Appendix 8.1 presents details of the sampling and randomization. Table 2 shows the geographic distribution of schools in the evaluation, by district and zone. Each of the three districts has 50 schools, by design. Figure 2 shows a map of the schools in the evaluation.

3 Methodology

3.1 Data

Our primary outcome 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. We administered the test in English. 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. Table 2 reports the components of the EGMA. 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). We normalize by the control group mean and standard deviation within each standard (grade). We calculate the overall EGMA score as the mean of the two normalized composite scores.

We collected baseline data in October 2023. Within each school, we administered the EGMA and a short survey to a random sample of 16 students in each of standards 1-4, or 64 students per school. We stratified the student sample by gender. We also surveyed the parent or guardian of each student in the sample, and the head teacher of each school.

We will revisit each school in June 2024 to administer the EGMA to the same students in the baseline survey. We will also conduct follow-up surveys with head teachers and teachers, and observe classrooms.

Although the sample size was determined primarily by budget constraints, it was also informed by power calculations. We have updated our *ex ante* power calculations based on the observed intra-cluster correlation of EGMA scores within control schools. For the full sample, the minimum detectable effect (MDE) is 0.15 standard deviations (sd). For gender-disaggregated analysis, the MDE is 0.16 sd. These MDEs are on par for successful education interventions in LMICs, for which the median effect size is 0.10 sd (Evans and Yuan 2022). See Appendix 7.2 for details of our power calculations.

3.2 Methods

To analyse endline results, we will compare outcomes between treatment and control schools using the following equation, estimated by ordinary least squares (OLS):

$$y_{isg} = \alpha + \beta NNP_{sg} + \theta y_{0isg} + \gamma_g + \varepsilon_{isg} \quad (1)$$

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; y_0 is the baseline outcome (where available); γ is a stratum (i.e., matched pair) fixed effect; and ε is the error term. Our coefficient of interest is β, which measures the effect (intent to treat, or ITT) of assignment to NNP on the outcome. We will cluster standard errors by strata, following recommendations for inference with small strata (de Chaisemartin and Ramirez-Cuellar 2024).

We will estimate equation (1) for all students in the school and separately by grade. In addition to results aggregating all genders, we will also disaggregate results by student gender. We will follow an analysis plan, which we will share on the RCT registry entry for this project.

4 Results

4.1 Baseline balance

Table 1 reports the number of students in the sample, by standard (grade level), sex, and treatment assignment. There are 9,400 students in the sample. The sample is roughly evenly divided by standard, sex, and treatment assignment, as intended.[2]

Table 3 reports balance on school characteristics, using data from the 2021 Education Management Information System (EMIS), an annual school census. Characteristics are similar across treatment and control schools, with one exception (the proportion of students with early childhood development [ECD] center exposure) significant at the 10% level, about what we would expect by chance. These are the same characteristics used to form matched pairs of schools. Balance is therefore not surprising, but nonetheless confirms the matched pair design succeeded at the school level.

Table 4 reports balance on EGMA scores. We normalize scores to the control group mean and standard deviation within each standard. EGMA scores are not statistically distinguishable between treatment and control students, in the full sample or disaggregated by sex. Figure 2 shows the full distributions of EGMA scores by treatment status. The treatment distribution is more dispersed than the control distribution, but otherwise their shapes are similar.

Comparing EGMA results between treatment and control students informs the evaluation design, but offers little insight into numeracy levels within the target population. We address this limitation by benchmarking EGMA results to proficiency levels. No universal set of EGMA proficiency benchmarks exists, nor are we aware of a benchmarking exercise for Malawi. We therefore define proficiency levels ourselves, building on EGMA benchmarks established for

[2] The sample size is close to the target number of 9,600 students (16 students/standard * 4 standards * 150 schools = 9,600). Within each standard, we aimed to sample an equal number of boys and girls.

other Sub-Saharan African countries (RTI International 2014a, 2014b, 2015). See Appendix 7.4 for details.

Table 5 reports the proportion of our sample meeting each benchmark, overall and separately by standard. The proportion of students meeting the proficiency standard is low, even in grades S3 and S4. For instance, only 1% of S1 students are proficient in addition and subtraction Level 2. This proportion rises to 6% for S3 and 15% for S4. By contrast, 37% of S3 students and 15% of S4 students score zero on addition and subtraction Level 2, meaning they cannot solve a single two-digit addition or subtraction problem.

4.2 Main results

[TBC]

4.3 Mechanisms

[TBC]

5 Conclusion

[TBC]

6 References

Alan, Sule, and Ipek Mumcu. 2024. "Nurturing Childhood Curiosity to Enhance Learning: Evidence from a Randomized Pedagogical Intervention." *American Economic Review* 114 (4): 1173–1210. https://doi.org/10.1257/aer.20230084.

Angrist, Noam, David K. Evans, Deon Filmer, Rachel Glennerster, F. Halsey Rogers, and Shwetlena Sabarwal. 2020. "How to Improve Education Outcomes Most Efficiently? A Comparison of 150 Interventions Using the New Learning-Adjusted Years of Schooling Metric."

Asim, Salman and Ravinder Casley Gera, 2021, What matters for learning in education in Malawi: Evidence from the Malawi Longitudinal school survey, Mimeo.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. 2017. "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application." *Journal of Economic Perspectives* 31 (4): 73–102. https://doi.org/10.1257/jep.31.4.73.

Banerjee, Abhijit V., Swati Bhattacharjee, Raghabendra Chattopadhyay, and Alejandro J. Ganimian. 2017. "The Untapped Math Skills of Working Children in India: Evidence, Possible Explanations, and Implications." *Unpublished Manuscript*. https://www.academia.edu/download/79182667/Banerjee˙et˙al.˙2017˙-˙2017-08-17.pdf.

Barros, Andreas de, Johanna Fajardo-Gonzalez, Paul Glewwe, and Ashwini Sankar. 2023. "The Limitations of Activity-Based Instruction to Improve the Productivity of Schooling*." *The Economic Journal*, November, uead099. https://doi.org/10.1093/ej/uead099.

Beeharry, Girindre. 2021. "The Pathway to Progress on SDG 4 Requires the Global Education Architecture to Focus on Foundational Learning and to Hold Ourselves Accountable for Achieving It." *International Journal of Educational Development* 82 (April): 102375. https://doi.org/10.1016/j.ijedudev.2021.102375.

Berlinski, Samuel, and Matias Busso. 2017. "Challenges in Educational Reform: An Experiment on Active Learning in Mathematics." *Economics Letters* 156 (July): 172–75. https://doi.org/10.1016/j.econlet.2017.05.007.

Blimpo, Moussa P., and Todd Pugatch. 2021. "Entrepreneurship Education and Teacher Training in Rwanda." *Journal of Development Economics* 149 (March): 102583. https://doi.org/10.1016/j.jdeveco.2020.102583.

Blimpo, Moussa, and Todd Pugatch. 2023. "Unintended Consequences of Youth Entrepreneurship Programs: Experimental Evidence from Rwanda." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4592983.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur. 2018. "Experimental Evidence on Scaling up Education Reforms in Kenya." *Journal of Public Economics* 168 (December): 1–20. https://doi.org/10.1016/j.jpubeco.2018.08.007.

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.

Díaz, Juan José, Omar Arias, and David Vera Tudela. 2012. "Does Perseverance Pay as Much as Being Smart? The Returns to Cognitive and Non-Cognitive Skills in Urban Peru." *Unpublished Paper, World Bank, Washington, DC*. http://www.iza.org/conference˙files/worldb2014/arias˙o4854.pdf.

Dillon, Moira R., Harini Kannan, Joshua T. Dean, Elizabeth S. Spelke, and Esther Duflo. 2017. "Cognitive Science in the Field: A Preschool Intervention Durably Enhances Intuitive but Not Formal Mathematics." *Science* 357 (6346): 47–55.

Evans, David K., and Susannah Hares. 2021. "Should Governments and Donors Prioritize Investments in Foundational Literacy and Numeracy?" 579. Working Paper. Center for Global Development. https://www.cgdev.org/publication/should-governments-and-donors-prioritize-investments-foundational-literacy-and-numeracy.

Evans, David K., and Anna Popova. 2016. "What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews." *World Bank Research Observer* 31 (2): 242–70.

Evans, David K., and Fei Yuan. 2022. "How Big Are Effect Sizes in International Education Studies?" *Educational Evaluation and Policy Analysis* 44 (3): 532–40.

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.

Glewwe, P., and K. Muralidharan. 2016. "Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." *Handbook of the Economics of Education* 5.

List, John A. 2022. *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*. New York: Currency.

Maruyama, Takao, and Takashi Kurosaki. 2024. "Developing Textbooks to Improve Math Learning in Primary Education: Empirical Evidence from El Salvador." *Economic Development and Cultural Change* 72 (2): 833–66. https://doi.org/10.1086/721768.

Ministry of Education, 2023, Malawi Education Statistics Report.

Muralidharan, Karthik, and Paul Niehaus. 2017. "Experimentation at Scale." *Journal of Economic Perspectives* 31 (4): 103–24. https://doi.org/10.1257/jep.31.4.103.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2019. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." *American Economic Review* 109 (4): 1426–60.

Nikoloski, Zlatko, and Mohamed Ihsan Ajwad. 2014. "Cognitive and Non-Cognitive Skills Affect Employment Outcomes: Evidence from Central Asia." *Unpublished Manuscript, Retrieved from Http://Eprints. Lse. Ac. Uk/60025*. http://www.iza.org/conference˙files/CognitiveSkills˙2014/nikoloski˙z6495.pdf.

Nourani, Vesall, Nava Ashraf, and Abhijit Banerjee. 2023. "Learning to Teach by Learning to Learn." https://www.povertyactionlab.org/sites/default/files/research-paper/WP3647˙Learning-to-teach˙Banerjee-et-al˙Nov2023.pdf.

Pugatch, Todd, Lontum Nchadze, Evaewero French, Alperen Acikol, and Esme Kadzamira. 2023. "Malawi National Numeracy Programme Impact Evaluation." https://doi.org/10.1257/rct.11964-1.0.

Rodriguez-Segura, Daniel, and Isaac Mbiti. 2022. "Back to the Basics: Curriculum Reform and Student Learning in Tanzania," RISE Working Paper Series, , no. 22/099.

RTI International. 2014a. "Proposing Benchmarks for Early Grade Reading and Mathematics in Ghana." RTI International, February 13, 2014 & May 5, 2014.

RTI International. 2014b. "Proposing Benchmarks for Early Grade Reading and Mathematics in Tanzania." RTI International, February 27-28, 2014.

RTI International. 2015. "Proposing Benchmarks and Targets for Early Grade Reading and Mathematics in Zambia." RTI International, July 2-3, 2015.

RTI International. 2023a. "Learning at Scale: Final Report." https://learningatscale.net/wp-content/uploads/2023/10/Learning-at-Scale-Final-Report-.pdf.

———. 2023b. "Numeracy at Scale: Final Report." RTI International. https://learningatscale.net/wp-content/uploads/2023/10/Numeracy-at-Scale˙Final-Report.pdf.

School-to-School International. 2023. "Endline Report: National Numeracy Programme in Malawi." STS International.

UNESCO Institute for Statistics (2020). "Global Proficiency Framework for Mathematics: Grades 1 to 9." UNESCO Institute for Statistics, December 2020.

UNESCO Institute for Statistics (2023). "Feasibility of using the data produced by the Early Grade Reading (EGRA) and Early Grade Mathematics (EGMA) to measure and monitor SDG 4.1.1 by complementing it with other banks of items." UNESCO Institute for Statistics, February 2023.

United Nations. 2016. "Report of the Inter-Agency and Expert Group on Sustainable Development Goal Indicators." United Nations.

Vivalt, Eva. 2020. "How Much Can We Generalize From Impact Evaluations?" *Journal of the European Economic Association* 18 (6): 3045–89. https://doi.org/10.1093/jeea/jvaa019.

World Bank. 2018. *World Development Report 2018: Learning to Realize Education's Promise*. World Bank Washington, DC.

7 Appendix

7.1 Examples of mathematics curriculum materials

7.1.1 Control group mathematics curriculum

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

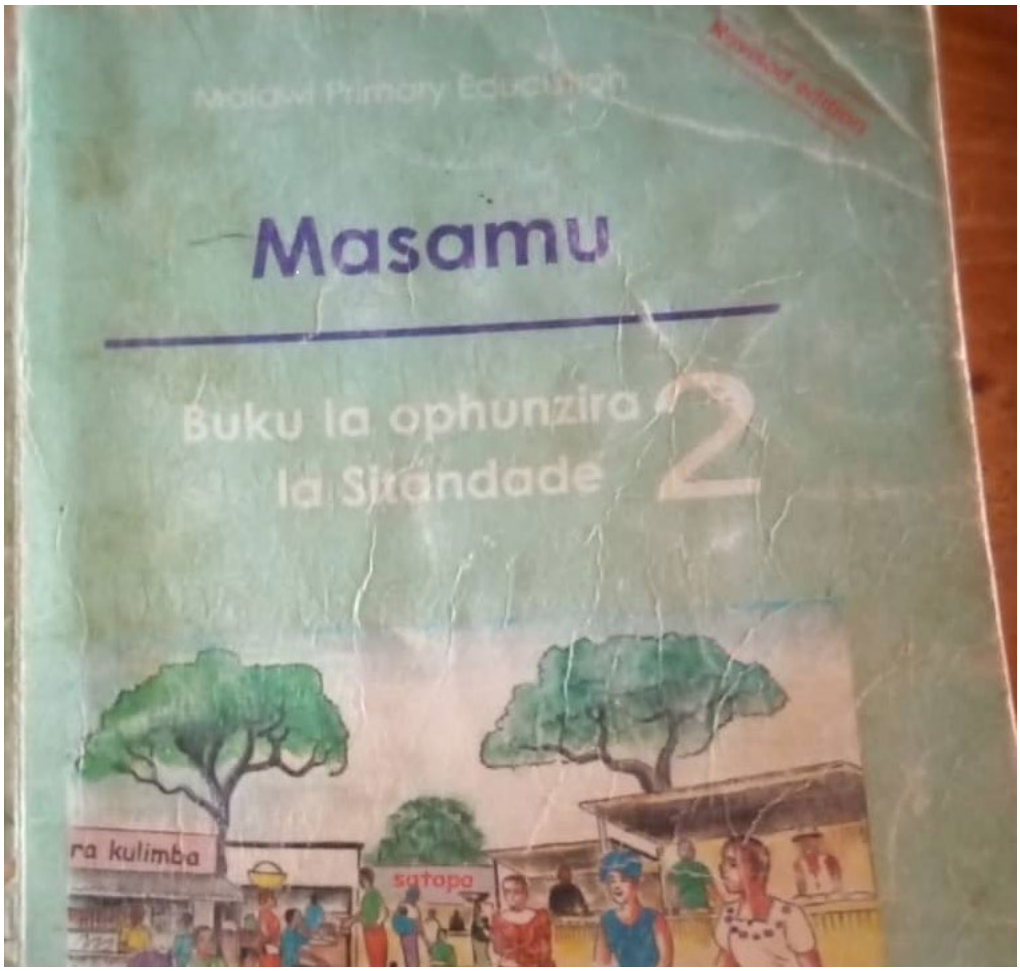
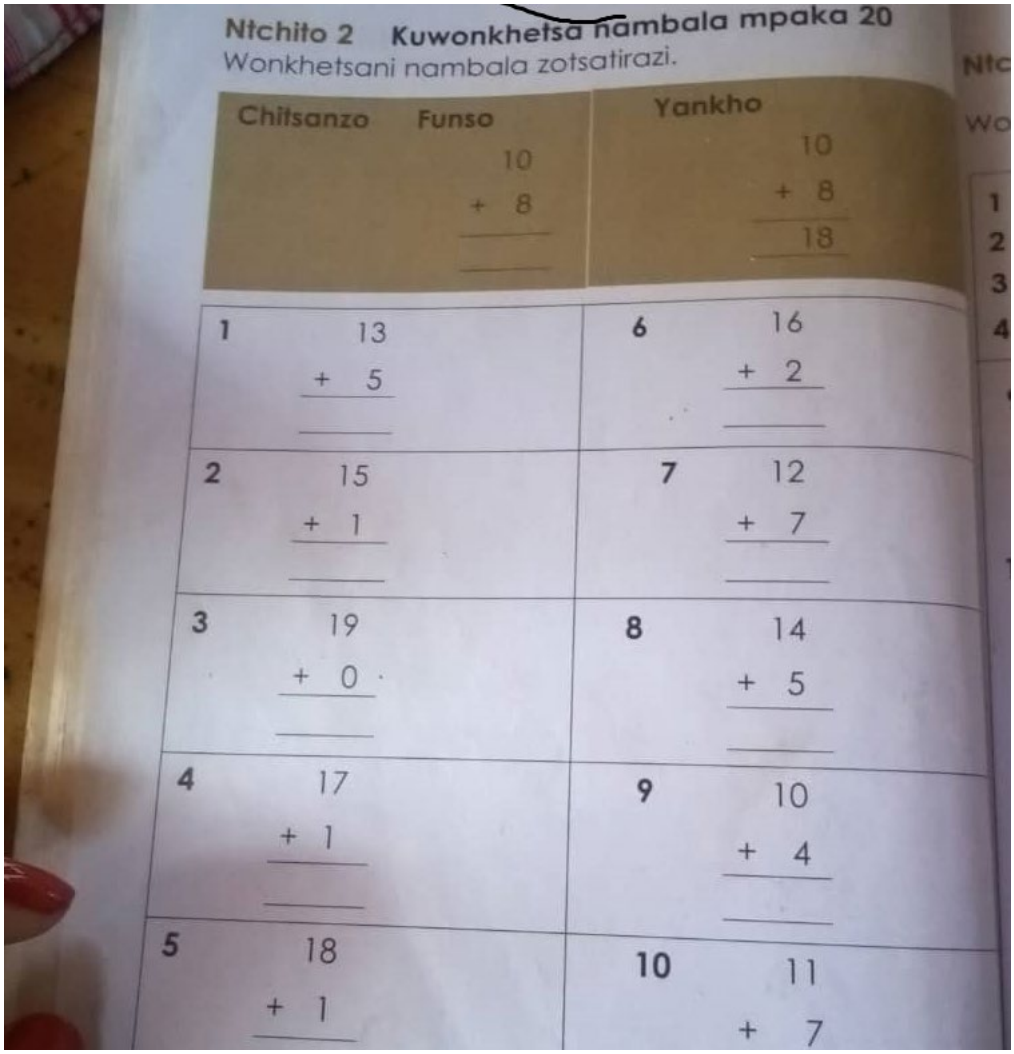


Figure 3: Standard 2, Page 24: Mathematics Exercise on Addition of figures up to 20



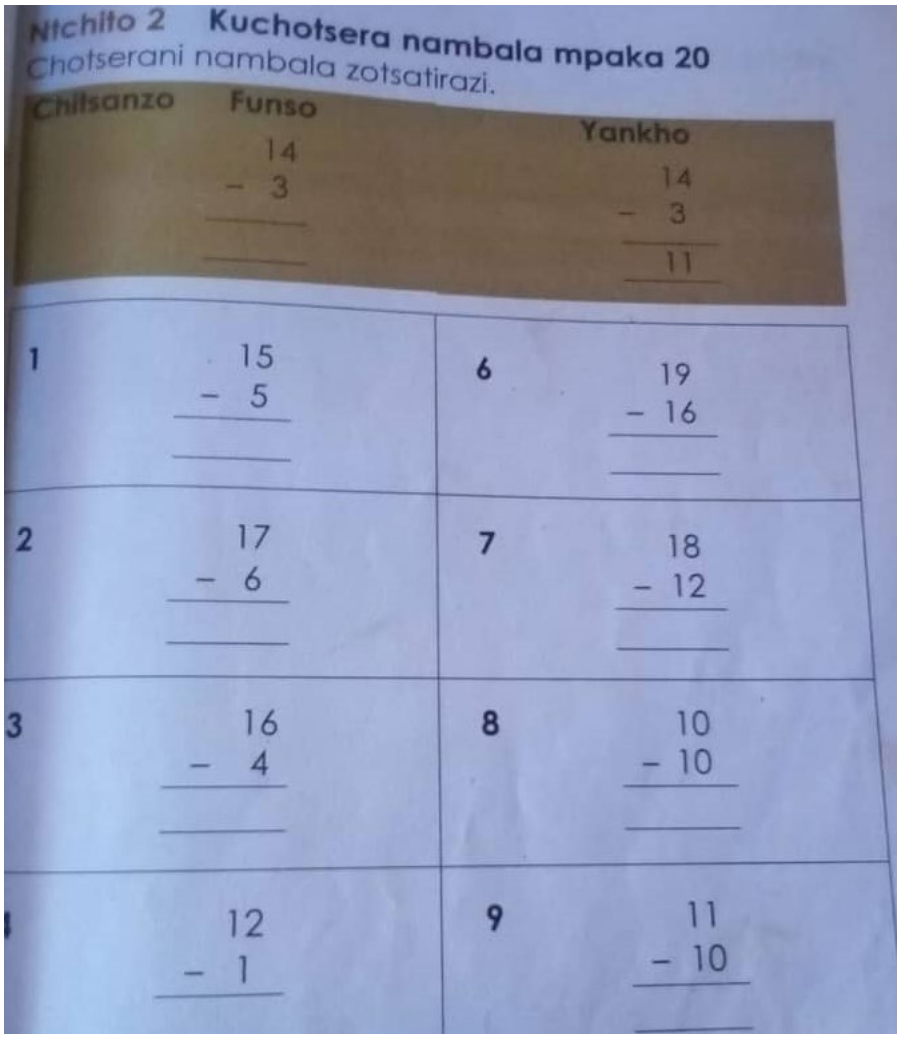
Ntchito 2 Kuwonkhetsa nambala mpaka 20

Wonkhetsani nambala zotsatirazi.

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

Note: The colored part shows an example.

Figure 4: Standard 2 Mathematics exercise on subtraction for figures up to 20



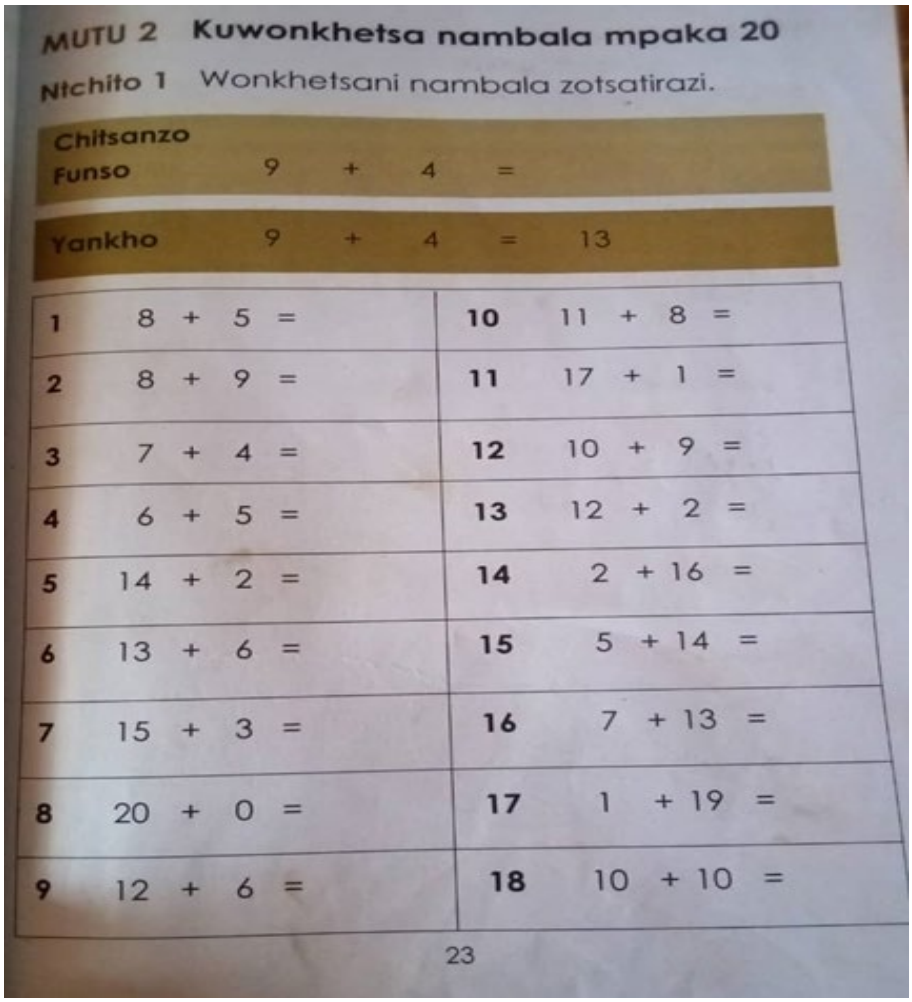
Nchito 2 **Kuchotsera nambala mpaka 20**

Chotserani nambala zotsatirazi.

Chitsanzo	Funso	Yankho
	14 − 3 =	14 − 3 = 11
1	15 − 5 =	
2	17 − 6 =	
3	16 − 4 =	
	12 − 1 =	
6	19 − 16 =	
7	18 − 12 =	
8	10 − 10 =	
9	11 − 10 =	

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

Figure 5: Standard 2 Mathematics Exercise on Addition of numbers up to 20



MUTU 2 **Kuwonkhetsa nambala mpaka 20**

Ntchito 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.

7.1.2 National numeracy programme curriculum

Figure 6: Standard 2 Mathematics Learner workbook for term 2: cover page (English language)

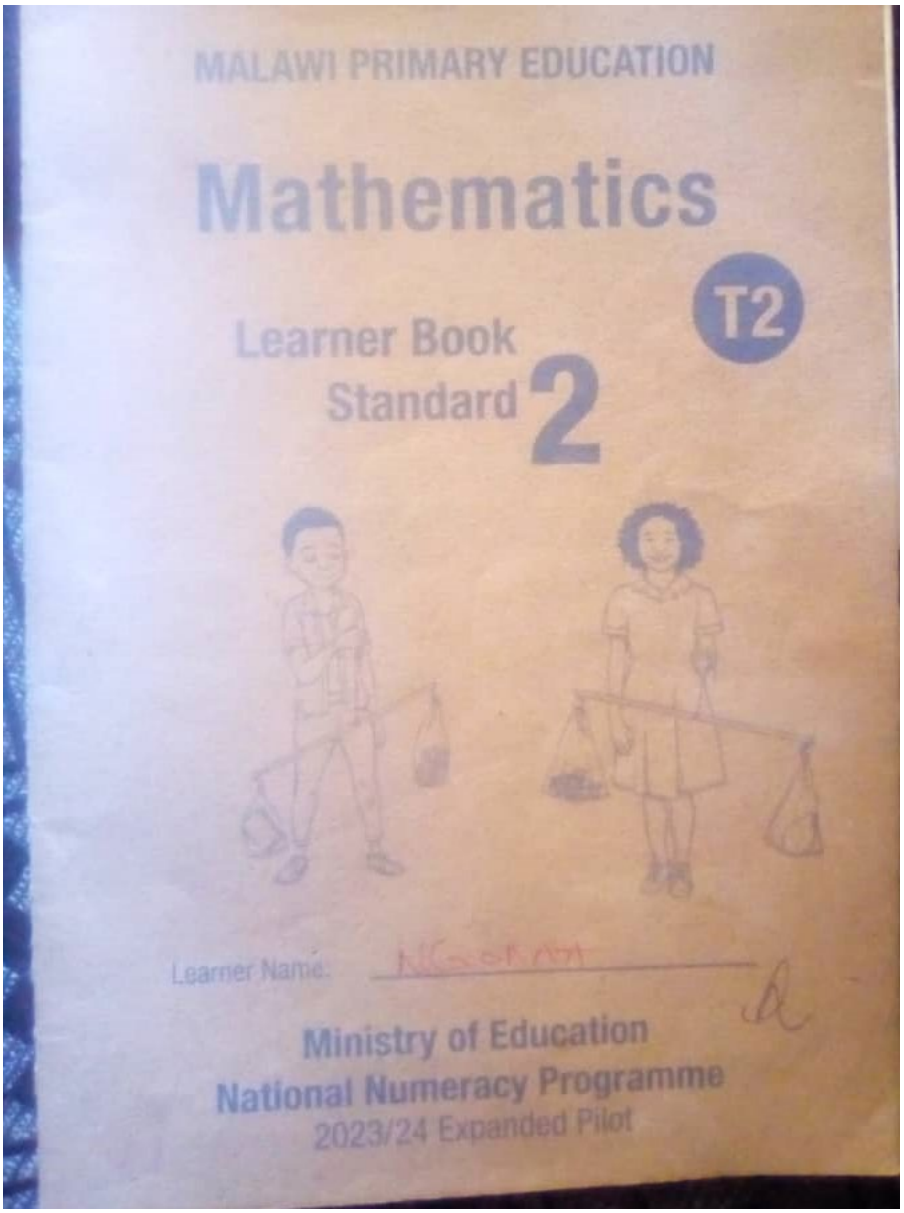
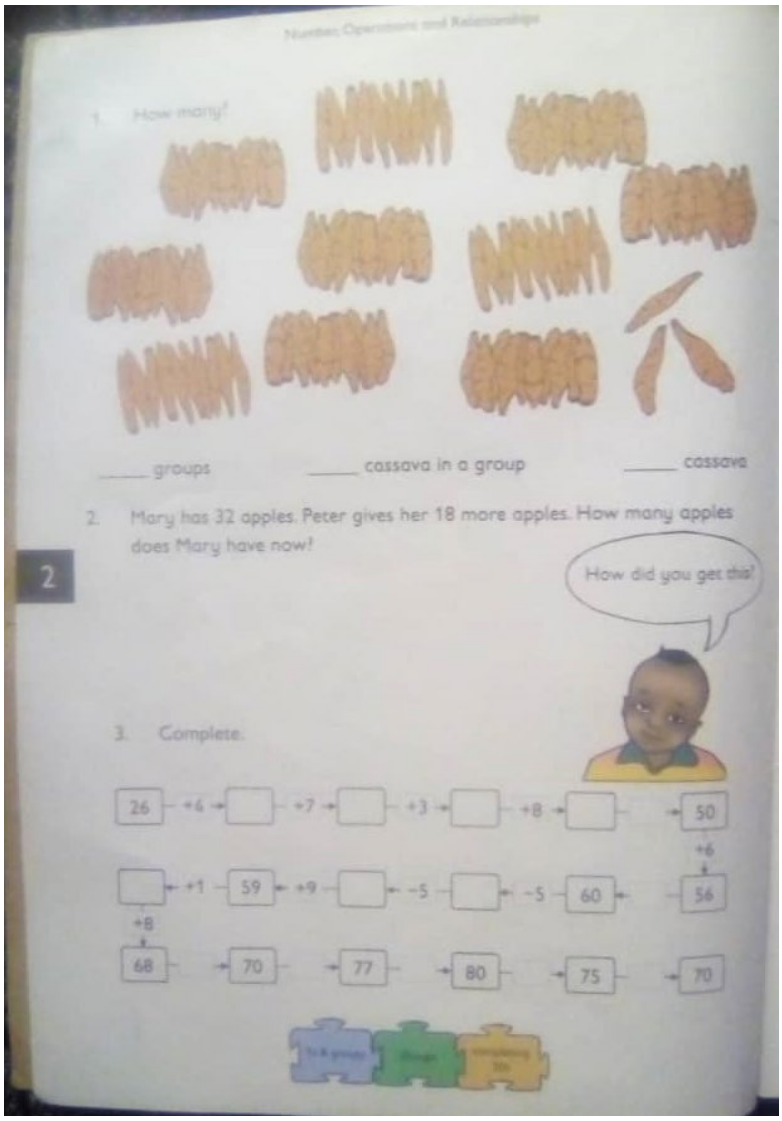


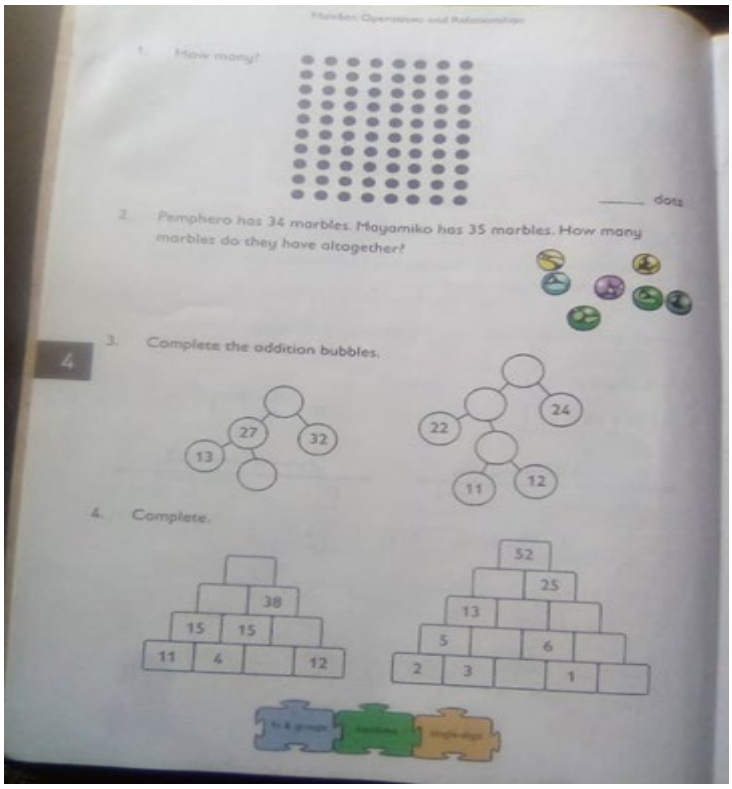
Figure 7: Standard 2 Mathematics Learner workbook for term 2, Page 2: Number Operations and Relationships



Observations:

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

Figure 8: Appendix 3: 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)

7.2 Sampling and randomization

In consultation with the implementing partners, 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. 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)
- Female 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.

Within each group of four, we randomly assigned two schools to the treatment group and two to the control group, following the recommendation by Athey and Imbens (2017). The researchers 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. Two treatment schools were replaced when implementers discovered they were in zones with schools in the NNP pilot.

The RCT therefore follows a matched pair design, with pairs formed based on district location and similarity of baseline school characteristics. The sample includes 150 schools, randomly split between treatment and control.

Table 3 reports sample means and balance tests for the school characteristics used to form matched pairs. There is one statistically significant difference between treatment and control schools (prior ECD exposure) at the 10% level, about what we would expect by chance. The F-test for joint significance of all school characteristics fails to reject the null of no differences between treatment and control.

7.3 Power calculations

Prior to baseline data collection, we calculated the minimum detectable effect (MDE) for the evaluation. We focused on grade-specific student learning outcomes, measured in standard deviation units. We have updated our power calculations, using the intra-cluster correlations (ICCs) observed in the baseline data for overall EGMA scores.

We made the following assumptions:

- Sample size:
  - Cluster RCT with 150 schools, split evenly into treatment and control
  - Within each school: 16 students per grade, evenly split by gender, with 30% attrition
- Power = 80%
- Test size = 5% (two-sided)
- Residual standard deviation after controlling for baseline outcome = 0.8

Table 1 below presents MDEs under these assumptions:

Table 1: Minimum detectable effects

		minimum detectable effect	
Sample	Baseline ICC	all	gender disaggregated
full	0.15	0.15	0.16
S1	0.37	0.24	0.25
S2	0.21	0.20	0.21
S3	0.30	0.22	0.24
S4	0.19	0.19	0.21

Assumptions: outcome = overall EGMA score, calculated as average z-score of timed and untimed sections; 80% power; 5% test size (two-tailed) residual standard deviation = 0.8; 75 schools per treatment arm; 16 students per standard, split by gender; 30% attrition.

MDEs range from 0.15-0.25 standard deviations, on par for successful education interventions in LMICs (Evans and Yuan 2022).

7.4 EGMA Benchmarks

In this section, we define and report 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 2: EGMA Benchmarks for this study

	Proficiency category			
EGMA subtask	**Novice**	**Beginner**	**emergent**	**proficient**
Addition and subtraction L1	0%	Greater than 0% but less than 40%	40%	80%
Addition and subtraction L2	0%	Greater than 0% but less than 40%	40%	80%
Missing numbers	0%	Greater than 0% but less than 30%	30%	70%

Source: Author definitions, guided by USAID Benchmarking exercises convened in Ghana, Tanzania, and Zambia.

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.

Table 31 reports the proportion of our sample meeting each benchmark, overall and separately by standard. The proportion of students meeting the proficiency standard is low, even in S3-S4. For instance, only 1% of S1 students are proficient in addition and subtraction Level 2. This proportion rises to 6% for S3 and 15% for S4.

8 Figures

Figure 1: Theory of change

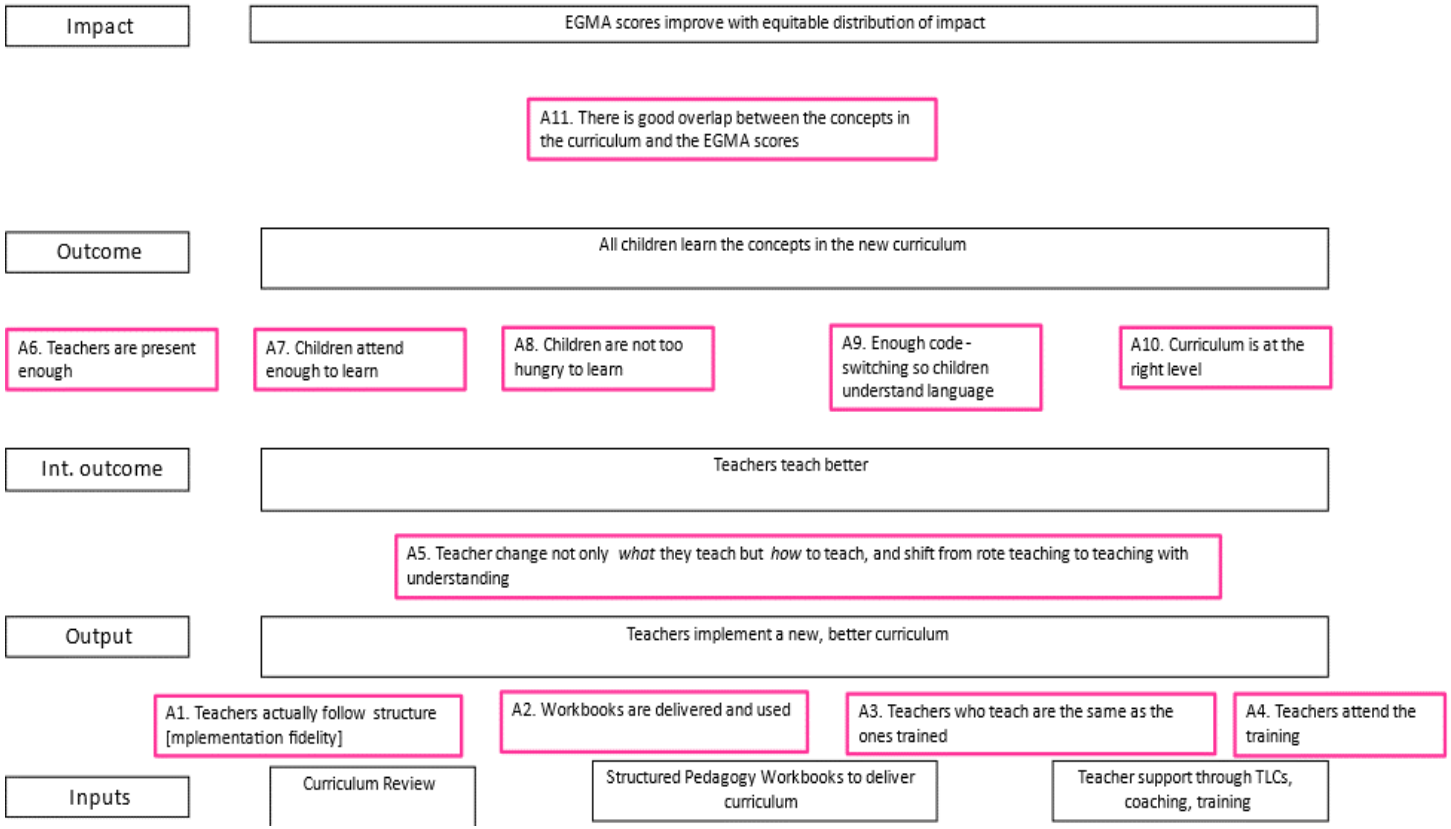


Figure 2: Map of schools in evaluation

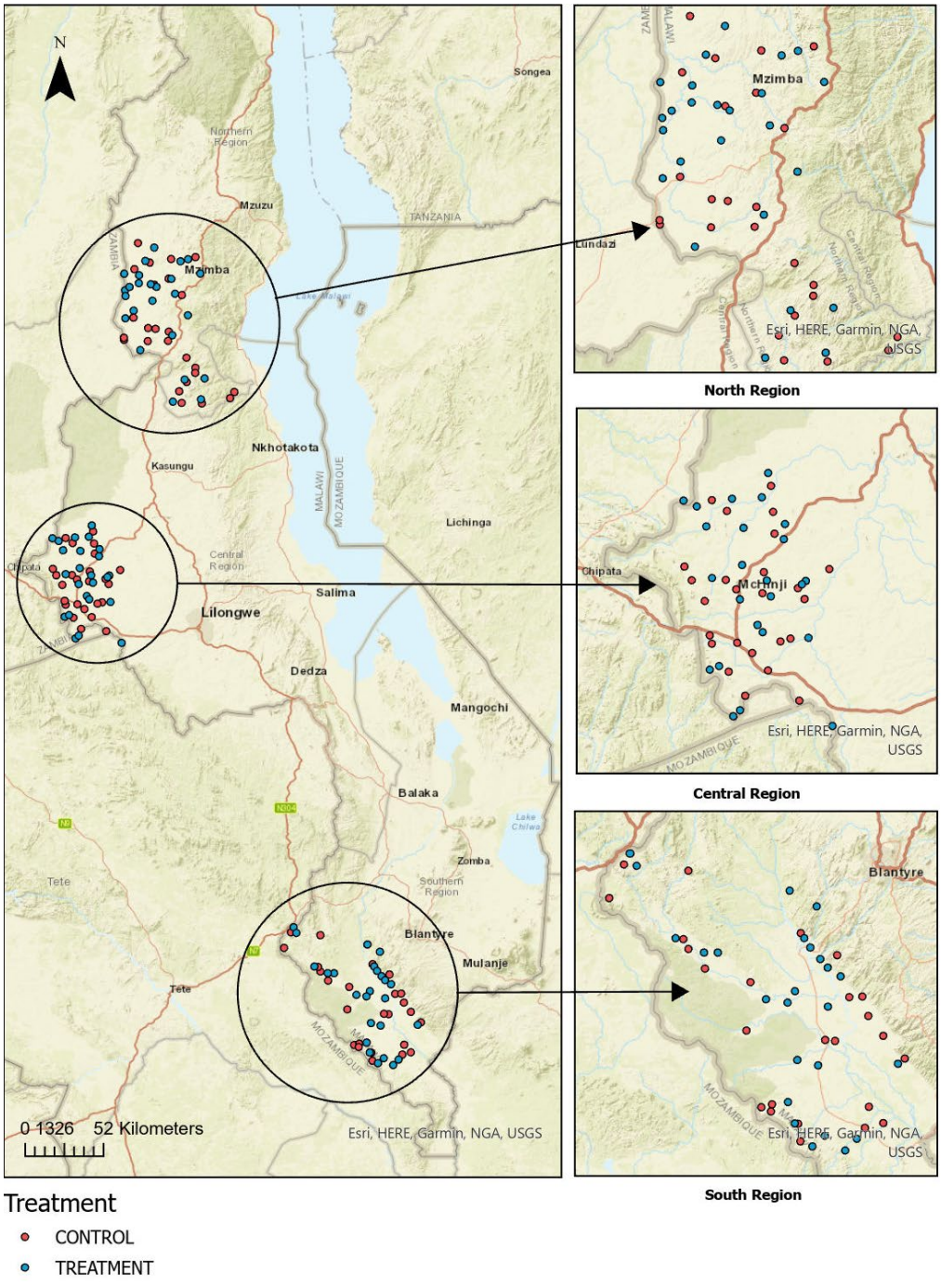


Figure 2: Baseline EGMA scores, full sample

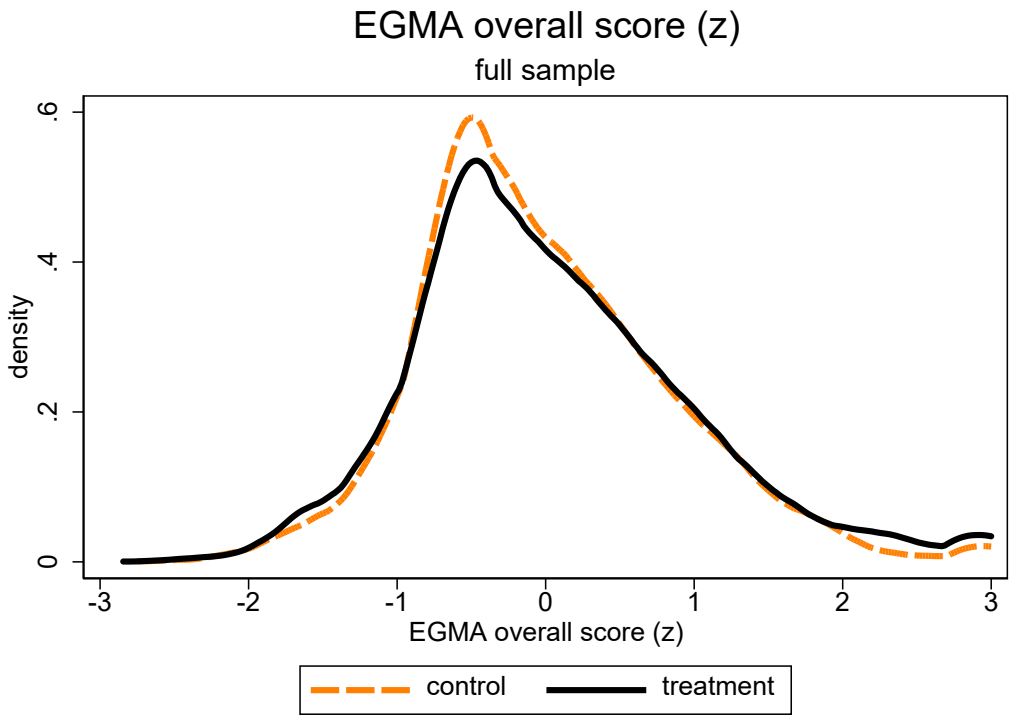


Figure shows kernel density estimates of EGMA overall score (z), calculated as average z-score of timed and untimed EGMA composite scores. Composite scores normalized to mean and standard deviation of control group, separately by standard. Figure top-codes scores at 3 for 83 of 9,400 students (0.9%) for visual purposes.

9 Tables

Table 1: Student sample size, EGMA data

	Control				Treatment				
	Male	Female	Missing	Total	Male	Female	Missing	Total	TOTAL
Standard 1	577	577	37	1,191	587	597	6	1,190	2,381
Standard 2	583	574	42	1,199	585	584	16	1,185	2,384
Standard 3	554	564	46	1,164	572	583	15	1,170	2,334
Standard 4	549	568	35	1,152	574	567	8	1,149	2,301
TOTAL	2,263	2,283	160	4,706	2,318	2,331	45	4,694	9,400

Table reports number of students in baseline EGMA data, by standard, sex, and treatment assignment.

Table 2: Geographic distribution of schools

	control	treatment	total
	(1)	(2)	(3)
Panel A: Mzimba South			
CHASATO	1	0	1
CHIKANGAWA	1	1	2
CHIZUNGU	2	1	3
EDINGENI	0	4	4
EMFENI	1	1	2
ENDINDENI	0	2	2
KABENA	1	1	2
KABUWA	2	0	2
KANJUCHI	1	1	2
KAPHUTA	1	2	3
KAPOLI	0	2	2
KATETE	1	0	1
KAVUULA	1	1	2
KAZINGILIRA	2	1	3
LUVIRI	0	1	1
LUWEREZI	0	1	1
MABIRI	1	1	2
MACHELECHETE	1	0	1
MANYAMULA	1	3	4
MHARAUNDA	0	1	1
MZOMA	3	1	4
UNYOLO	2	0	2
VAZALA	1	0	1
VIBANGALALA	2	0	2
Total	*25*	*25*	*50*
Panel B: Mchinji			
CHIMTEKA	1	2	3
CHIOKO	1	2	3
GUMBA	2	3	5
KALULU	2	4	6
KAMWENDO	3	2	5
KAPIRI	2	2	4
KAVUNGUTI	2	1	3
LUDZI	4	0	4
MIKUNDI	2	3	5
MKANDA	1	3	4
PINDA	2	2	4
TASEKERA	2	0	2
WALIRANJI	1	1	2
Total	*25*	*25*	*50*
Panel C: Chikwawa			
CHANGOIMA	2	2	4
CHIKONDE	1	5	6

DOLO	1	2	3
KAKOMA	3	1	4
KALAMBO	2	2	4
KONZERE	3	2	5
LIVUNZU	3	1	4
MAPELERA	3	2	5
MBEWE	1	3	4
MKUMANIZA	4	2	6
NCHALO	2	3	5
Total	*25*	*25*	*50*
Grand total	**75**	**75**	**150**

Table shows schools in sample, by district and zone.

Table 3: Balance on baseline characteristics, schools (2021 EMIS)

	control	treatment	difference
	(1)	(2)	(1)-(2)
age of school (years)	43.1	39.7	3.4
	[30.4]	[26.0]	
enrolment	705.9	681.8	24.2
	[644.6]	[548.6]	
pupil-teacher ratio	63.5	62.8	0.7
	[24.2]	[20.9]	
female enrolment proportion	0.51	0.51	0.00
	[0.03]	[0.04]	
female teacher proportion	0.24	0.27	-0.03
	[0.21]	[0.20]	
Remoteness category A	0.37	0.47	-0.09
	[0.49]	[0.50]	
Remoteness category B	0.44	0.36	0.08
	[0.50]	[0.48]	
share with ECD background	0.19	0.33	-0.15*
	[0.63]	[0.73]	
repeater share of enrollment, S1-S4	0.29	0.27	0.02
	[0.10]	[0.10]	
share passed PSLCE exam	0.81	0.79	0.02
	[0.23]	[0.19]	
management index	0.84	0.86	-0.02
	[0.17]	[0.16]	
N	75	75	
F-test of joint significance (F-stat)			1.30

Table shows means by treatment group. Remoteness categories from Asim et al, 2019. Category A is considered "most remote," Category B is "moderately remote," and Category C ("not remote") is the omitted category. 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. Standard deviations in brackets. Column (3) shows difference. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level. Final row shows F statistic on test of joint significance.

Table 3: Baseline EGMA scores, full sample

	treatment (1)	control (2)	difference (1)-(2)
Panel A: all students			
timed	0.06	0.00	0.06
	(0.06)	(0.04)	
untimed	0.07	0.00	0.07
	(0.05)	(0.05)	
overall	0.06	0.00	0.06
	(0.05)	(0.04)	
N	4,694	4,706	
Panel B: boys			
timed	0.06	0.00	0.07
	(0.06)	(0.04)	
untimed	0.13	0.04	0.09
	(0.05)	(0.05)	
overall	0.09	0.02	0.08
	(0.05)	(0.04)	
N	2,342	2,364	
Panel C: girls			
timed	0.07	0.00	0.06
	(0.07)	(0.06)	
untimed	0.01	-0.04	0.05
	(0.05)	(0.05)	
overall	0.04	-0.02	0.05
	(0.05)	(0.05)	
N	2,352	2,342	

Table shows baseline EGMA results. Unit of observation is the student. Data collected from 150 schools (75 treatment, 75 control, grouped into strata consisting of matched pairs). Outcomes reported as z-scores, normalized to mean and standard deviation of control group. Overall score is average z-score of timed and untimed composite scores. Standard error in parenthesis. Final column shows difference, controlling for strata fixed effects and clustering standard error by strata. Significance: ***=.01, **=.05, *=.1.

10 Appendix

10.1 Appendix Tables

Table 4: EGMA components

Score	Items	Timed?
Number identification	20	timed
Addition Level 1	20	timed
Subtraction Level 1	20	timed
Number discrimination	10	untimed
Missing numbers	10	untimed
Addition Level 2	5	untimed
Subtraction Level 2	5	untimed
Word problems	6	untimed

Table 5: EGMA results by performance benchmarks

Score		Skills Benchmarks (in percentages)		
Group	Subtest	Zero scores	Emerging	Proficient
Overall	Addition/Subtraction Level 1	19.5	25.3	2.2
	Addition/Subtraction Level 2	56.1	17.8	5.3
	Missing Numbers	41.7	23.0	3.6
Standard 1	Addition/Subtraction Level 1	51.6	2.2	0.00
	Addition/Subtraction Level 2	95.5	0.6	0.00
	Missing Numbers	81.5	4.2	0.1
Standard 2	Addition/Subtraction Level 1	19.3	9.0	1.2
	Addition/Subtraction Level 2	74.9	7.1	0.6
	Missing Numbers	50.3	12.3	0.3
Standard 3	Addition/Subtraction Level 1	5.4	30.5	3.6
	Addition/Subtraction Level 2	37.4	25.5	5.8
	Missing Numbers	24.1	28.1	3.8
Standard 4	Addition/Subtraction Level 1	0.7	61.0	4.3
	Addition/Subtraction Level 2	14.9	38.9	15.0
	Missing Numbers	9.6	48.3	10.5

Table reports baseline EGMA results by performance benchmarks for the entire sample and by grade. Unit of observation is the student. Data collected from 9,400 students across 150 schools (75 treatment, 75 control, grouped into strata consisting of matched pairs).