CES ifo
Working Papers

12542
2026

March 2026

Targeting Foundational Skills at Scale: Skill Specificity and Transfer

Andreas de Barros, Theresa Lubozha

CES ifo

Imprint:

CESifo Working Papers

ISSN 2364-1428 (digital)

Publisher and distributor: Munich Society for the Promotion
of Economic Research - CESifo GmbH

Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740

Email office@cesifo.de
https://www.cesifo.org

Editor: Clemens Fuest

An electronic version of the paper may be downloaded free of charge
· from the CESifo website: www.ifo.de/en/cesifo/publications/cesifo-working-papers
· from the SSRN website: www.ssrn.com/index.cfm/en/cesifo/
· from the RePEc website: https://ideas.repec.org/s/ces/ceswps.html

Targeting Foundational Skills at Scale: Skill Specificity and Transfer

Andreas de Barros
Theresa Lubozha

March 18, 2026

Targeting Foundational Skills at Scale: Skill Specificity and Transfer*

Andreas de Barros† Theresa Lubozha‡

March 18, 2026

Whether targeted foundational instruction yields broad, long-term human capital gains is central to education policy but largely untested. We provide causal evidence from Zambia's government-run foundational skills program in public primary schools. After two years, a randomized trial shows the program increases literacy by 0.10 and numeracy by 0.15 standard deviations. In mathematics, effects on targeted skills are 2.6 times larger than on comprehensive assessments, without detectable transfer to adjacent domains. Adding professional development doubles per-pupil costs without additional learning gains. Despite limited short-run transfer, event-study estimates show positive effects on grade-7 language and mathematics exam scores in early adolescence.

Keywords: field experiment; foundational skills; human capital; skill formation; skill transfer; long-term effects.

JEL codes: C93; H52; I21; I28; J24.

*This article follows a prospective, registered analysis plan available on the American Economic Association's registry for randomized controlled trials (study ID: AEARCTR-0014922). It was approved by the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology, the University of Zambia Biomedical Research Ethics Committee, Zambia's National Health Research Authority, and the Ministry of Education. This research was supported by the Global Partnership for Education Knowledge and Innovation Exchange (KIX), a joint endeavor with the International Development Research Centre, Canada, through Grant No. 109295-001 to the Massachusetts Institute of Technology. The Belgian Agency for International Cooperation provided additional funding. de Barros acknowledges support from the National Academy of Education (NAEd) and the NAEd/Spencer Postdoctoral Fellowship Program. The authors thank J-PAL, Pratham, Teaching at the Right Level Africa, VVOB – education for development, and Zambia's Ministry of Education for making this study possible. The authors thank the Examinations Council of Zambia for making data available, especially the Department of Research, Planning, and Information. The Centre for Promotion of Literacy in Sub-Saharan Africa, Innovations for Poverty Action, and Palm Associates assisted with data collection. The authors thank Prince Muraguri and Victor Olajide for their excellent research assistance. Dennis Kyalo and Jacqueline Mathenge provided research management. Jacobus Cilliers, Adrienne Lucas, and Isaac Mbiti volunteered to serve on the study's research advisory committee. The IDELA team at Save the Children, the TEACH team at the World Bank, and Ben Weidmann and Yixian Xu at the Harvard Kennedy School Skills Lab generously shared data collection instruments. Jenny Beth Aloys provided expert training on classroom observations. We also thank Anjali Adukia, Kodjo Aflagah, Drew Bailey, Abhijit Banerjee, Rukmini Banerji, Arja Dayal, Greg Duncan, Caroline Elliot, Alejandro Ganimian, Joshua Gilbert, Maimuna Ginwalla, Rachel Glennerster, Paola Guerrero-Rosada, Junita Henry, Chavi Jain, Mridul Joshi, Miranda Moolenaar, Shadrick Mwaba, Daniele Ressler, Katherine Rhodes, Sharon Schroen, Lena Shi, Tavneet Suri, Vikas Varma, and Nico Vromant. Lubozha is employed by VVOB – education for development, and has no other conflicting interests to declare. de Barros has no conflicts of interest to declare and retains the final editing rights among the two authors. In compliance with common journal policies, the authors acknowledge the use of Artificial Intelligence (AI) for AI-assisted copy editing, code development, and image generation for one survey instrument—they did not use AI for generative editorial work or autonomous content creation.

†Assistant Professor, University of California, Irvine. E-mail: `adb@uci.edu`.

‡Strategic Education Advisor, VVOB – education for development.

1 Introduction

Does a targeted focus on foundational skills shape broader learning? Standard frameworks in the economics of education treat early literacy and numeracy as building blocks for later learning, where gains in foundational competencies strengthen the base on which subsequent, more advanced skills are built (Cunha and Heckman, 2007). This view implies that improving a child's foundational skills should generate returns that extend beyond the specific skills targeted. Yet, there is limited direct evidence on two questions central to this view: whether targeted gains in specific foundational skills remain skill-specific or transfer to adjacent domains, and whether a focus on foundational skills translates into broader academic performance in later grades.

Research on learning transfer in cognitive science suggests that the first question has a predictable answer: gains in narrowly targeted skills rarely generalize to broader competencies, even within the same content domain, without explicit bridging instruction (Barnett and Ceci, 2002). The second question—whether a narrow focus on foundational skills can nonetheless generate broader academic benefits over longer horizons—has received less direct empirical attention. Both questions matter for models of human capital production and for how researchers and policymakers evaluate educational investments.

This paper provides evidence on both questions by studying Zambia's *Catch Up* program, a national adaptation of the Teaching at the Right Level (TaRL) methodology. The program operates across approximately 5,000 primary schools, reaches more than one million students annually, and is implemented entirely by regular government teachers without additional staff. In mathematics, the program targets number recognition and procedural arithmetic—a subset of the content domains that comprise internationally recognized definitions of foundational numeracy.[1] In literacy, the program's content aligns with standard definitions of foundational literacy, with no notable narrowing of focus. This asymmetry in instructional targeting across subjects provides a useful contrast for interpreting differences in short-run skill gains and longer-run academic outcomes.

We employ two complementary research designs. The first is a preregistered cluster-randomized trial in which we randomly assigned 182 administrative zones, and the 1,115 government-run primary schools within them, to treatment or control. Within treatment zones, we further randomized one school per zone to receive an enhanced version of the program that doubled per-student costs (from \$9.63 to \$19.97) by adding continuous professional development (CPD) for teachers. For primary data collection, we

[1] Related descriptive evidence from India suggests that mastery of number recognition and arithmetic can substantially exceed mastery of other foundational mathematics domains, such as geometry, fractions, and measurement, with these gaps widening across grades (de Barros and Ganimian, 2023).

subsampled 273 schools from these zones and followed 8,025 grade-3 students for two years. The trial measures both the specific skills the program targets and a comprehensive set of foundational skills, aligned with internationally recognized definitions and Zambia's national curricular framework. The second design is an event study exploiting the program's staggered rollout across 116 districts. Using restricted data covering the universe of public grade-7 primary school leaving examinations from 2014 to 2025—approximately 4.4 million test scores per subject—we examine whether exposure to remedial instruction in grades 3 through 5 increases performance on the wider set of language and mathematics skills tested on the grade-7 leaving examinations.

On the first question, our experimental results confirm pronounced skill-specificity in mathematics, with narrowly targeted procedural gains accompanied by only limited transfer to adjacent foundational skills. After two years of intervention, the program improved foundational literacy by 0.10 SD and foundational mathematics by 0.15 SD. In mathematics, effects on the specific skills the program targets, number recognition and procedural arithmetic, were substantially larger, at 0.40 SD. Effects on the remaining foundational math skills were small (0.03 SD) and not statistically distinguishable from zero. Even within the same content domains, effects on number sense and applied arithmetic, which require reasoning beyond procedural execution, were only 0.05 SD, compared with 0.40 SD on targeted skills. In literacy, where the program covers the full range of foundational skills, no such divergence arises.

On the second question, however, our event-study results point to a different pattern. Despite a lack of short-run transfer, the program produces positive effects on grade-7 primary school-leaving exams in early adolescence—comprehensive exams that test skills well beyond those the program targets.[2] Among cohorts expected to have been fully exposed to the program, it improved performance by 0.14 SD in language ($p < 0.001$) and 0.11 SD in mathematics ($p = 0.023$). The contrast between a lack of short-run transfer and positive long-run effects on distal assessments is difficult to reconcile with standard accounts of learning transfer, which hold that broad generalization is typically *less* likely than gains on proximal skills (Barnett and Ceci, 2002; Detterman, 1993).

The divergence in effects across narrowly targeted and comprehensive foundational mathematics skills also has direct implications for how researchers assess program effectiveness. Using only the subset of targeted skills, the program's cost-effectiveness in mathematics is \$2.43 per 0.1 SD in the short run; using the comprehensive assessment, it is \$6.38 per 0.1 SD—a factor of 2.6. Because many evaluations of remedial programs use proximal assessments aligned with program content (Banerjee et al., 2017; Ardington et al.,

[2]Only 3.6 percent of grade-7 mathematics test items cover the skills the program targets.

2026), reported effect sizes may overstate gains in the broader foundational competencies these programs aim to develop.

These short- and long-run effects emerged under the constraints typical of a national program operating at scale. Unlike many evaluations of remedial instruction, which study NGO-implemented interventions, summer camps, or programs with dedicated facilitators, this paper investigates a government-run program delivered by regular civil-service teachers. On any given school day, only one-third of the students in our experimental sample attended a remedial class. Doubling per-student costs to provide teachers with continuous professional development yielded no additional learning gains, despite strong implementation: 95.6 percent of eligible schools participated in the CPD component, and 86.8 percent of participating teachers engaged with it more than five times. The program thus operates in a context in which student exposure is limited, additional investments in teacher capacity do not translate into additional learning, and in-school delivery depends entirely on existing government systems.

These findings contribute to research on human capital formation in three ways.

First, we document an empirical pattern that informs models of early skill formation: large contemporaneous gains in targeted procedural skills coexist with limited transfer to adjacent foundational skills in the short run, yet cohorts exposed to the program show broad achievement gains on comprehensive exams in early adolescence. Standard frameworks model foundational competencies as aggregate inputs whose returns are realized through dynamic complementarity (Cunha and Heckman, 2007; Cunha et al., 2010). Our experimental results show no detectable contemporaneous transfer to non-targeted mathematical competencies—consistent with evidence from cognitive science that procedural gains rarely generalize without explicit bridging instruction (Barnett and Ceci, 2002; Perkins and Salomon, 1992)—while event-study estimates indicate positive effects on grade-7 examinations covering competencies well beyond the program's targeted content. Taken together, these results suggest that the contribution of early procedural instruction to broader academic performance operates over longer horizons than a two-year trial window and through mechanisms that may depend on subsequent learning opportunities.

Second, we show that short-run and long-run evaluations of the same program can identify fundamentally different estimands, with first-order consequences for inference about the returns to foundational skill investments. Bailey et al. (2020) document that "overalignment" between intervention content and outcome measures is a pervasive concern in education evaluations. Our evidence provides a direct test within the same program, using two complementary designs. The 2.6-fold difference in implied cost-effectiveness across outcome measures is large enough to affect cross-study

comparisons that inform global policy recommendations (GEEAP, 2023). Yet the event-study estimates suggest that the narrow short-run pattern may not reflect a ceiling on what the program achieves. These results imply that overalignment in short-run evaluations can coexist with positive effects on broader, distal assessments that may materialize only over longer horizons—a possibility that has received limited attention in discussions of treatment-outcome alignment (Bailey et al., 2017, 2020).

Third, we contribute evidence on the education production function at scale. A large experimental literature documents positive effects of targeted instruction in lower-income settings, but predominantly in NGO-implemented interventions, summer camps, or programs with dedicated facilitators (Angrist and Meager, 2023). Evidence from teacher-led, in-school programs is more limited, and existing findings are mixed (Banerjee et al., 2017; Duflo et al., 2024; Ardington et al., 2026). In our setting, the program generates positive effects despite limited per-student exposure and delivery through existing government systems. The null effect of the added CPD component aligns with previous literature: collaborative professional development has shown limited returns in lower-income countries (Popova et al., 2022), while level-based instruction and explicit instructional routines have produced positive effects in similar contexts (Angrist et al., 2024). Together, these results expand the evidence base on what government-run programs can achieve, showing that positive effects are attainable even with limited per-student exposure, without parallel delivery structures, and without added peer-based professional support.

2 Context and interventions

2.1 Teaching at the Right Level in Zambia

Public education, governed by the Ministry of Education, is the most common form of primary school education in Zambia. Publicly financed schools currently serve more than 95 percent of primary school students in the country. Enrollment is free, and primary schools operate in a region's local language.[3] Primary school enrollment numbers are high, at 85.6 percent net enrollment. This positive development contrasts with high levels of learning poverty among Zambia's children. In the 2023 UNESCO AMPL assessments of foundational numeracy and literacy skills, only 12.7 percent of grade-4 students reached

[3]There are seven official languages of instruction in Zambia. As there are more than 70 local languages spoken in the country, a school's language of instruction may differ from a child's home language. After our study, in 2025, a curricular reform introduced English as the language of instruction.

the global minimum proficiency levels articulated by the Sustainable Development Goals (SDG 4.1.1).

To address these low learning levels, Zambia's Ministry of Education has embraced the Teaching at the Right Level (TaRL) methodology since the end of the 2016 academic year (Lipovsek et al., 2023).[4] The Zambian TaRL program focuses on grades three to five, and it is locally known as *Catch Up*. As of 2025, it covered nine of Zambia's ten provinces, serving approximately 5,000 primary schools and just over 1,000,000 learners annually.[5] The program divides children into groups based on their learning needs and pace and adds extra remedial lessons during which teachers provide differentiated mathematics and literacy instruction to each group. For mathematics, compared to the content domains covered by the Global Proficiency Framework and commonly accepted definitions of foundational learning, the program focuses more explicitly on domains related to number recognition and procedural arithmetic (de-emphasizing geometric shapes, measurement, and data display). For literacy, there are no notable differences. All program implementers are regular teachers and inspectors commonly assigned to the public primary schools, with no additional staff assigned to schools; NGO involvement is limited to technical assistance.[6]

2.2 Additional continuous professional development for TaRL teachers

The continuous professional development (CPD) program investigated in this research rests on a prior mixed-methods study that set out to identify what drives Zambian public school teachers to change their instruction (de Barros et al., 2024). Using primary qualitative data from 78 Zambian education personnel from the school to the provincial level, this study combined qualitative thematic analysis with an unsupervised machine-learning technique (Natural Language Processing; topic modeling) to identify drivers of pedagogical shifts. It then combined qualitative analyses with linear probability models to uncover their associations with teacher professional development. Its findings suggest that teaching practices are malleable, with change being predominantly driven by on-site continuous professional development opportunities relating to team-based problem-solving, verbal

[4]The Ministry launched a "pre-pilot" activity at the end of 2016, piloted alternative implementation modalities in 2017, and launched the intervention described in this article in 2018. For a more detailed description of the interventions' Theory of Change, see Appendix B.

[5]The tenth province started phasing in the program in 2026.

[6]To describe the intervention in greater detail, in Appendix C, we draw on classroom observation data and child surveys from our trial's endline. Remedial classes in program schools feature more group work, use more teaching and learning materials, and exhibit higher scores on structured classroom observation measures of teaching quality. At the same time, class sizes and teacher effort in remedial classes are comparable to those in regular classes. Program schools also exhibit lower levels of corporal punishment compared to non-program schools. Within program schools, remedial classes exhibit lower levels of corporal punishment than regular literacy and math classes.

discussions, and skills acquisition. The study highlighted the potential of school-based CPD opportunities as a means to alter teaching practices, and it motivated the development of the CPD program.

Based on the findings from the mixed-methods study and a subsequent year of iterative piloting on a small scale, a research-practice partnership with the Ministry of Education co-developed a continuous professional development program to better support *Catch Up* teachers. This program complements the standard *Catch Up* program by establishing and supporting communities of practice among teachers (both within schools and between schools via WhatsApp). During regular professional development meetings, teachers engage in discussions about the *Catch Up* program, receive additional guidance documents and videos that align with their responsibilities throughout the academic year, and are invited to collaborate with their colleagues to participate in "mastery challenges." In addition, the Ministry recognizes teachers' successful participation through non-monetary incentives and issues formal letters of commendation.

3 Research design

This article employs two complementary research designs using distinct data sources: a large, preregistered cluster-randomized trial with a waitlist design, and an event study. In the randomized trial, we randomly assigned 91 of 182 zones and all public primary schools in those zones to receive the *Catch Up* program.[7] The remaining zones and their public primary schools were assigned to continue with business as usual until 2025 (or later). In addition, in each of the 91 zones assigned to receive the program, we randomly assigned one public primary school to receive the continuous professional development (CPD) program for teachers. Random assignment to the interventions allows us to study the causal effect of being assigned to the intervention after two years (the intent-to-treat, or "ITT" effect) and to compare this effect across the two program variants (with vs. without the CPD program component). In turn, the event study leverages the program's staggered roll-out across the country and reports on the long-run effects of a district's exposure to the program on students' primary school leaving exam scores (at the end of grade 7). Figure 1 depicts the program's phased rollout and provides a map of the subsample of schools in the randomized trial, along with their assignment to the three experimental groups.

[7] Above the school level, "zones" reflect the smallest administrative subdivision of Zambia's public education system. Zones are nested within districts, and districts are nested within provinces.

3.1 Sample and representativeness

For the randomized trial, our sampling strategy followed a three-step process. First, we identified a sample of 182 zones. These zones were slated for a potential program roll-out but had yet to receive the *Catch Up* program. They are located in eleven districts in Central province, one district in Southern province, and six districts in Western province.

Second, we determined the sample of 273 schools. If a zone was assigned to receive *Catch Up*, all publicly-funded schools in that zone were targeted by the program (government-run schools and government-supported "community" schools). Yet, for the study's data collection activities, we drew a random subsample of government schools (excluding community schools).[8] With access to microdata on all schools in the country, we first constructed a list of government schools across the study zones (1,115 overall). In a random half of the zones, we then randomly sampled one government school per zone; in the remaining half of the zones, we randomly sampled two of these schools.

In the third and final step, during school visits at baseline, we randomly subsampled third-graders from among those who were present on the day of the baseline visit. We stratified our sampling by gender and selected up to 40 students per school (not all schools had 40 students present). At baseline, we successfully surveyed 8,025 students (4,091 girls and 3,934 boys), or about 29 students per school.[9] These students represent the study's sample we tracked over two years; at endline, we successfully re-surveyed 86.8 percent of these students.

Appendix Table A1 assesses the representativeness of our sample of schools along key observable dimensions, including enrollment numbers, gender ratio, rurality, and accreditation status. Panel A compares schools in our sample of 182 zones to all other public primary schools nationwide. Panel B confirms the representativeness of our random subsample of 273 government schools relative to all 1,115 government schools within the study zones. We document no meaningful differences in observable characteristics and, in both panels, joint F-test p-values exceed 0.1. These results indicate that the study zones and sampled schools are broadly similar to other public primary schools nationwide on key observed characteristics and support the relevance of our experimental findings for government-run public primary schools in the country.

[8] While recognized community schools in Zambia may receive government support, they are typically community- or NGO-managed rather than government-run.

[9] Of the formally enrolled students, 11.8 percent had not been to school in the past four weeks. Of those who had come to school in the past four weeks, 23.6 percent were absent on the day of the visit. Among those present and sampled at random, 2.9 percent could not be surveyed (e.g., they left the school before the survey team completed their school visit).

For the event study, we use restricted data from all public grade-7 national assessment examination centers across Zambia from 2014 to 2025 (12 years). This sample includes the universe of 6,618 public examination centers with all exams administered at these sites (4,439,847 language test scores and 4,461,130 mathematics scores). The program rolled out at the district level across 116 districts at different times during this period, with control zones from the randomized trial remaining untreated throughout. Each observation in our event-study sample represents a center-by-sex-by-year mean, weighted by the number of students tested in that cell.

3.2 Randomization

We created three experimental groups of schools for the randomized controlled trial. We began by randomly assigning half of the 182 zones to either receive the *Catch Up* program or not receive the program (and continue with business as usual).[10] We randomized zones within strata of four zones each. We generated these strata by grouping zones that (a) shared the same district and (b) had similar levels of average academic performance.[11] After that, in the zones assigned to the program, we randomly assigned one school to receive the program with the additional continuous professional development (CPD) program. We refer to these three sets of schools as the "Control", "regular *Catch Up*", and "*Catch Up* with CPD" groups, respectively.

Appendix Figure A1 summarizes the study's sampling and randomization procedures. Table 1 and Appendix Figure A2 confirm that randomization successfully led to balanced groups of schools and students whose differences in observable characteristics do not exceed what can be expected by chance. Attrition rates are balanced as well, and so are the observable characteristics of the non-attriting students.

3.3 Data

This article uses primary data collected across three data collection rounds, additional administrative data, data on program costs, and restricted data on Zambia's primary school leaving exams. All primary data was collected with independent research teams, not by the Ministry or NGOs. Data collection included a "baseline" and "endline" in all 273 schools

[10] Those zones assigned to the *Catch Up* program are the same in which we sampled two government schools for data collection; in the remaining zones not assigned to the program, we sampled one school (see above).

[11] To establish which zones shared similar performance levels, we used test scores from Zambia's official grade-7 exams and ranked zones by their average performance in math and language. If a district's number of zones was not divisible by four, we grouped the remainder of schools across districts; also, as 182 is not divisible by four, one stratum has only two zones.

sampled for the trial (July to November 2022; September to November 2024), and one round of unannounced visits to all 273 schools (July 2024). We administered all assessments and interviews with students and teachers in the official, local language of a given school (Bemba, Lozi, Nyanja, and Tonga).[12]

To capture the main outcomes of the randomized trial, we measured students' foundational skills in literacy and mathematics with one-on-one assessments. The instruments consisted of two components: (1) A standard "ASER" test that covers select math domains (number recognition and procedural arithmetic) and select literacy domains (letter recognition and reading), and (2) additional test questions that focus on the remaining domains of foundational mathematics and literacy not measured by the ASER test. Both test components were adaptive; they only tested more advanced skills if students had the respective prerequisites (e.g., students who could not read letters were not asked to attempt a reading comprehension task).

To construct the assessments, we used a blueprint with a clear mapping of test questions to content and cognitive domains. They follow common definitions of "foundational skills," which are recognized internationally and in Zambia.[13] The blueprint also maps the test questions to grade-level expectations, following Zambia's official curriculum framework.[14] The assessments recorded students' responses to all test questions (or test "items") for both test components. We use a hybrid item response theory (IRT) model to aggregate these responses and generate continuous estimates of student ability (a two-parameter logistic model for binary items and a graded response model for ordinal items). We generate one overall score per subject and ensure comparability across waves using anchor items with fixed parameters (see Appendix D for technical details). We standardize all scores at the student level using the control group distribution at endline. At endline, the average conditional reliability of the literacy and mathematics measures is 0.93 and 0.90, respectively.

As a prespecified, secondary outcome for the trial, we also created a measure of mathematics skills explicitly targeted by the program. In mathematics, we generated one continuous, standardized score for these skills (i.e., number recognition and procedural

[12]Translations and local adaptations included multiple field pilots, discussions during enumerator training sessions, and an expert review of assessments with native speakers who had taught in the given language (one expert per language).

[13]These definitions align with Zambia's mathematics syllabus for the early grades and with the national literacy framework. In mathematics, they also align with the UNESCO global proficiency framework. In literacy, the national literacy framework (and our assessments) cover skills that go beyond the UNESCO global proficiency framework for reading (such as writing, for example).

[14]In 2025, after our study, Zambia's curriculum for public primary schools changed; here, we are referring to the curriculum in place during the study period. The authors thank expert reviewers at Zambia's Ministry of Education for their evaluation of the assessments and confirmation that they align with Zambia's "Literacy and Numeracy Education Framework."

arithmetic) as opposed to the remaining content domains captured by the assessments (e.g., geometry). As before, we estimate each student's score using a hybrid item response theory model and standardize scores with respect to the control group at endline.[15] Using endline data, the average conditional reliability of the measure of skills explicitly targeted by the program is 0.80.

As additional secondary outcomes in the randomized trial, we also measured students' creativity, their socio-emotional skills, and working memory. To explore mechanisms, we measured students' attitudes towards school, literacy, and mathematics, whether students studied after school, and teachers' participation in and exposure to continuous professional development activities with their colleagues. In Appendix D, we provide a detailed description of each of these measures and report on their measurement properties.

The administrative data includes backend data on all teacher interactions with the CPD communication system, records on training attendance, teacher payroll data, and school-level data from Zambia's Education Management Information System (EMIS). To capture information on program costs, we compiled detailed records following J-PAL's templates for cost-effectiveness analysis. Finally, our event-study results rely on restricted data from Zambia's Examinations Council covering the universe of grade-7 primary school leaving examinations administered in public schools from 2014 to 2025. This dataset contains center-level statistics (means, standard deviations, and student counts), disaggregated by subject and sex, from all public examination centers nationwide. The data includes location identifiers that allow us to link each center to the timing of program rollout, based on program implementation records. In Appendix D, for mathematics, we provide an item-level mapping of all grade-7 test questions, showing that only 3.6 percent of test questions cover the procedural, foundational number and operations skills specifically targeted by the intervention.

3.4 Compliance and exposure

School-level compliance with the program was high, but actual student exposure was substantially lower. Appendix Table A2 reports implementation and exposure measures for the pooled treatment group and separately by program variant.

Nearly all schools assigned to the program implemented it (Table A2, Panel A). Among schools assigned to treatment, 97.8 percent submitted diagnostic assessment data to the Ministry, 95.6 percent ran the program in the second term of the 2024 school year, and 86.8 percent did so in the third term. Schools scheduled *Catch Up* classes on 4.3 out of

[15] For completeness, we also report on the proportion of students who have mastered discrete levels of ability, as per the ASER tests (we focus on whether students can at least do two-digit subtraction with borrowing).

5 weekdays on average, and 86.8 percent had a teacher serving as a dedicated program mentor. About half of the teachers on a school's payroll had attended a multi-day, off-site training; others may have been trained on-site by their school's mentor.[16]

Despite this high compliance, student exposure was low (Table A2, Panel B). Only about half of students in treatment schools had attended school the previous day (50.4 percent), and only one-third of all students (33.3 percent) had attended any *Catch Up* class. Students are expected to attend one remedial class per day in one of the two subjects, so dosage fell short on two margins: many students were absent from school entirely, and even among those present, only roughly two-thirds attended a remedial class.

Lastly, of the 91 program schools assigned to receive the additional CPD component, virtually all participated (see Appendix Figure A3). Almost all had at least one teacher who had attended an off-site CPD training (93.4 percent), and nearly all participated in at least one of the mastery challenges (95.6 percent). Again, our data on teacher-level take-up rates are limited, as teachers could participate in teams, and we cannot distinguish teachers teaching the targeted grades (3 to 5) from those teaching the remaining four grades. Among *all* the teachers working in the assigned schools, including those not expected to participate, approximately one-third participated in mastery challenges (34.6 percent). Among those participating, many sustained their engagement in mastery challenges, with 86.8 percent of participating teachers completing more than five and 60.9 percent completing more than ten submissions.

4 Experimental results

Following our pre-analysis plan, to estimate the two-year effects of the program, our identification strategy rests on the trial's random assignment of schools to experimental groups. We exploit this random assignment to estimate the causal effects of being assigned to the interventions through linear regressions, with the following specification:

$$Y_{isr} = \alpha_r + \beta_1 T_{sr} + \beta_2 D_{sr} + \beta_3 L_{isr,0} + \beta_4 M_{isr,0} + \boldsymbol{\delta}' \boldsymbol{X}_{isr,0} + \epsilon_{isr} \tag{1}$$

Here, Y_{isr} is the outcome of interest for student i in school s and randomization stratum r, measured at endline. In our primary analyses, Y_{isr} represents test scores. The α_r terms are randomization stratum fixed effects, T_{sr} is the treatment assignment dummy for the regular *Catch Up* program, D_{sr} is a dummy indicating a school's random assignment to the program

[16]This percentage includes all teachers employed by the school, including those assigned to teach grades not targeted by the program.

with continuous professional development, and ϵ_{isr} is the residual. To increase precision, following Cilliers et al. (2024), all specifications include students' baseline performance in literacy and mathematics ($L_{isr,0}$, $M_{isr,0}$) and $\boldsymbol{X}_{isr,0}$ as covariates. Measured at baseline, $\boldsymbol{X}_{isr,0}$ is a vector of controls selected by a post-double selection (PDS) Lasso procedure on student and school characteristics, partialing out the strata fixed effects and baseline measures of literacy and mathematics skills (Belloni et al., 2014).[17] Following Abadie et al. (2022), we cluster standard errors at the zone level (182 clusters).

The coefficients β_1 and β_2 identify intent-to-treat effects. As discussed below (in section 4.2), we find that effects for the two program variants (with or without the additional CPD component) do not differ meaningfully, both in terms of their substantive magnitude and statistical significance. As preregistered, we therefore pool treatment arms in our analyses of overall program effects, estimating a version of Equation 1 with a single treatment indicator.

Accounting for our preregistered hierarchy of primary, secondary, and exploratory research hypotheses, we adjust for multiple hypothesis testing by computing the sharpened false discovery rate (FDR)-adjusted q-values. We prioritize the study's statistical tests in the registered priority order and do not adjust p-values for our two main outcomes of interest (i.e., foundational learning in literacy and mathematics). In turn, we calculate q-values within two prespecified families of tests: tests related to secondary research questions (e.g., concerning the subset of foundational math skills targeted by the program), and tests related to potential mediators and mechanisms.

4.1 Program effects

We present the overall program effects on students' foundational skills in Column (1) of Table 2. At endline, assignment to the program improved students' foundational literacy skills by 0.10 standard deviations (SD; $p < 0.001$) and foundational mathematics skills by 0.15 SD ($p < 0.001$). Effects on the subdomain of math skills targeted by the program (number recognition and basic arithmetic) were substantially larger, at 0.40 SD ($q = 0.001$). For reference, over the two-year period between baseline and endline, control-group students who participated in both assessment rounds gained 1.10 SD in literacy, 1.31 SD in mathematics, and 1.56 SD in the subdomain of targeted math skills, respectively (see Column (1) of Table 1).

[17] We prespecified the control variable input set for the PDS algorithm in our pre-analysis plan (Table 1). Following Cilliers et al. (2024), we use the default, "plug-in" penalty parameter of Stata's *pdslasso* command (not cross-validation).

We document the effects on secondary outcomes and potential mechanisms in Column (1) of Table 3. The program improved students' working memory by 0.07 SD (q = 0.038). In contrast, the point estimates for effects on creativity and socio-emotional skills are close to zero and are not statistically significant at conventional levels. We also find a small improvement in student attitudes towards school, literacy, and mathematics, of 0.04 SD, but we cannot reject the null hypothesis of no impact (q = 0.156). As a result of the program, students in the treatment schools were 3.9 percentage points (p.p.) more likely to have studied at home (q = 0.005). In program schools, teachers were also more likely to collaborate with their colleagues and receive feedback (an increase of 0.13 SD), though this outcome is measured at the teacher level, estimates are noisier, and we cannot reject the null (q = 0.156).

4.2 Effects of additional continuous professional development

Despite teachers' exposure to and take-up of the additional continuous professional development component, we do not find evidence that it improved student learning relative to the common program without it. Columns (2) and (3) of Table 2 show that the estimated effects for the two experimental groups are nearly identical, and Column (4) indicates that their difference is close to zero. If anything, Table 3 suggests that the effect on teacher collaboration may have been *lower* in the CPD group than in the group without this added component (0.07 SD vs. 0.19 SD; a -0.12 SD difference). Although this difference is imprecisely estimated, we can rule out improvements larger than 0.09 SD based on the upper bound of the 95 percent confidence interval (or improvements exceeding 0.05 SD using a one-sided 95 percent upper confidence bound).

4.3 Heterogeneity

Following our pre-analysis plan, we examine heterogeneous treatment effects using two complementary approaches: confirmatory subgroup analysis and exploratory machine learning methods. Our subgroup analysis focuses on two dimensions of heterogeneity. First, we examine effects by gender, motivated by Duflo et al. (2024), who find larger effects of TaRL for girls than boys in Ghana (despite the intervention not being designed to favor either gender). Second, we examine effects by baseline performance, as instruction targeted to learning levels should particularly benefit those students who lag furthest behind.

Table 4 presents treatment effects for two prespecified subgroups. For girls, program effects of 0.12 SD in literacy (q = 0.004) and 0.16 SD in mathematics (q = 0.002) are similar to the overall effects reported in Table 2. For students with low baseline performance, we find

differential patterns across subjects: in mathematics, effects for the bottom quartile (0.14 SD, q = 0.018) are similar to those for the full sample, but in literacy, effects are concentrated among low performers (0.15 SD, q = 0.008).[18]

To explore potential heterogeneity further, we employ causal forests (Athey and Imbens, 2016). For this additional exploration, we prespecified two hypotheses: that effects would be larger in schools with greater within-school dispersion of learning levels (measured by each school's Gini coefficient of baseline scores), and that within schools, students who lag behind their peers would particularly benefit. [19] Appendix Table A3 reports the results. In both subjects, within-school inequality is the variable most frequently selected for splitting, and estimated conditional average treatment effects are larger in high-inequality schools than in low-inequality schools—consistent with our first hypothesis, though imprecisely estimated. By contrast, within-school performance ranks low in variable importance and shows no consistent gradient in effects, providing little support for our second hypothesis. More broadly, the forest's overall ranking does not capture meaningful heterogeneity: the AIPW-estimated group-average effects for the above- and below-median predicted-CATE halves are inverted for literacy and only modestly ordered for mathematics, and a cross-validated RATE test (Chernozhukov et al., 2025) does not reject the null of no heterogeneity for either subject (literacy $p = 0.364$; mathematics $p = 0.612$).

Taken together, program effects in literacy are largest among low-performing students, and the causal forests suggest that school-level dispersion in learning levels may moderate program effects—while a student's position within their school's achievement distribution does not appear to systematically moderate effects. Yet the additional, exploratory findings remain suggestive because they are imprecisely estimated, and the causal forests do not detect significant overall heterogeneity in either subject.

4.4 Extensions

We complement the above findings with two sets of additional analyses related to skill specificity and teacher beliefs (not among the trial's prioritized hypotheses). First, in light of the domain-specific heterogeneity in program effects, Panel A of Table 5 documents program effects for those domains of foundational mathematics not specifically targeted by the program. These analyses confirm that the program's effects on foundational mathematics are almost exclusively driven by improvements in students' ability to answer

[18] Following our pre-analysis plan, we report effects *within* subgroups rather than testing for differential effects *across* subgroups.

[19] The latter within-school measure—a student's rank relative to schoolmates—differs from the overall baseline performance quartile examined above: a student in the bottom quartile of the full sample may nonetheless rank above peers in a low-performing school, and vice versa.

questions related to number recognition and procedural arithmetic. For the remaining math skills, the program effects are close to zero (0.03 SD) and do not reach conventional levels of statistical significance (q = 0.490). Even for items belonging to the same content domains as the targeted skills yet capturing applied arithmetic (vs. procedural arithmetic) and number sense (vs. number recognition), the program produced more muted effects (of 0.05 SD; q = 0.271) compared to the large program effects of 0.40 SD for the prespecified, targeted domain. Appendix Figure A4 presents item-level program effects, confirming that all but two of the math questions with a positive effect reflect number recognition and procedural arithmetic skills.

Second, we explore additional mechanisms at the teacher level using interview data collected during school visits in July 2024 (shortly before the launch of the study's endline assessments in September 2024). Panel B of Table 5 documents no notable program impacts on whether teachers attribute their students' learning to factors within their control vs. other outside factors (their perceived "locus of control").[20] Moreover, Panel B shows impacts on the extent to which teachers overestimate students' skills. Specifically, we document the gap between their estimates for grade-5 students attending their school and the observed student performance in that school at endline. In the control group, teachers overestimated the share of students who can read a short paragraph (by 24.4 percentage points) and the share of students who can solve the subtraction question (by 32.0 percentage points). In literacy, where the program did not substantially increase the percentage of students who could read a paragraph, teachers did not update their beliefs. In math, the program's impact on procedural arithmetic skills coincided with a sharp decrease in teachers' overestimates (closing teachers' overestimates by 14.1 percentage points; q = 0.001). Thus, teachers updated their beliefs only for the specific procedural math skills that improved as a result of the program, while maintaining their prior beliefs about students' literacy skills and their own ability to affect student learning more broadly.

5 Event-study results

To investigate the long-run effects of the program, we estimate dynamic treatment effects using Gardner's (2022) two-stage difference-in-differences estimator, which avoids the negative-weighting bias that can afflict conventional two-way fixed effects estimation under staggered adoption. We stack male and female observations as separate rows within each center-year, yielding a center-by-sex-by-year panel. This allows us to absorb gender-specific

[20] We caution that our locus of control measure exhibits substantial noise, which, after investigating data for the control group only, led us to de-prioritize this outcome in our pre-analysis plan before accessing the full endline dataset.

time shocks and gender-specific district intercepts while pooling both sexes for estimation of the treatment effect. In the first stage, we estimate district-by-sex and year-by-sex fixed effects using only untreated (not-yet-treated and never-treated) observations:

$$Y_{cst} = \alpha_{d(c),s} + \lambda_{ts} + \epsilon_{cst} \quad \text{for all } (c, s, t) \text{ where } D_{u(c)t} = 0 \tag{2}$$

where Y_{cst} represents standardized test scores for center c, sex s, in year t; $d(c)$ denotes the district containing center c; $\alpha_{d(c),s}$ is a district-by-sex fixed effect; λ_{ts} is a year-by-sex fixed effect; and $D_{u(c)t}$ is an indicator for treatment status of unit $u(c)$, the treatment unit containing center c. Treatment is assigned at the district level for most districts, but at the zone level for districts participating in the randomized controlled trial (zones are nested within districts). In the second stage, we regress the residualized outcomes $\tilde{Y}_{cst} = Y_{cst} - \hat{\alpha}_{d(c),s} - \hat{\lambda}_{ts}$ on event-time indicators:

$$\tilde{Y}_{cst} = \sum_k \beta_k \cdot \mathbf{1}(K_{u(c)t} = k) + \nu_{cst} \tag{3}$$

where the sum runs over all event-time periods excluding the omitted reference period $k = -1$, and never-treated units are absorbed by the residualization. $K_{u(c)t}$ denotes event time relative to program rollout to treatment unit $u(c)$, and the coefficients β_k capture treatment effects for each event-time period k. Because treatment is assigned at the district or zone level and does not vary by sex, the treatment-effect coefficients β_k are pooled across genders; the sex-interacted fixed effects serve only to flexibly control for gender-specific level differences and trends. We weight observations by the number of students tested and cluster standard errors at the district level (116 clusters).

Note that, in the first two years of a school's exposure (periods $k \in \{0, 1\}$), students taking the grade-7 exams are not expected to have benefited from the program, which focuses on supporting students as they progress from grades 3 to 5. Unless there are spillover effects to untargeted grade levels, we do not expect program impacts on exam scores during its first two years; in the subsequent two years (periods $k \in \{2, 3\}$), we may observe grade-7 students who were partially exposed to it, yet not for the full, expected three years. To assess the program's impact after three years of exposure in the targeted grade levels, we report the pooled effect for periods $k \in \{4, 5, 6, 7\}$, corresponding to students who, absent grade retention, were in grade 3 or below at rollout and can thus be expected to have been fully exposed to the program as they progressed through its target grade levels.[21]

[21] Following Gardner (2022), the pooled effect is the equal-weighted average of the four event-study coefficients $\hat{\beta}_4$ through $\hat{\beta}_7$, with the standard error obtained from the estimated variance-covariance matrix. We do not argue that all students were fully exposed to the program, nor that there is no grade retention among students. Rather, the reported intent-to-treat effect based on cohort eligibility represents the policy-relevant parameter for a planner who designs and evaluates educational programs based on typical grade progression.

Figure 2 presents the results, with effects expressed in student-level standard deviations. We find no evidence of systematic pre-trends for either subject. For Zambian language scores, the pooled effect for cohorts expected to have been fully exposed to the program is 0.14 standard deviations ($p < 0.001$). For mathematics, the pooled effect is 0.11 standard deviations ($p = 0.023$). Across these fully exposed cohorts, point estimates appear somewhat larger at longer event-time horizons (i.e., $k = 6$ and $k = 7$), suggesting that effects may strengthen with additional years since program rollout. This phenomenon may reflect program maturation (schools becoming more proficient); however, it may also reflect spill-overs from the program to untargeted grades (teaching practices being used in grades 1 and 2). In the long run, the intervention thus improved the comprehensive literacy and mathematics skills captured by the Zambian primary school leaving exams. Appendix Figure A5 documents that, for both subjects, these positive effects are largely uniform across girls and boys.

Finally, Appendix Figure A6 indicates no meaningful change in the number of grade-7 examinees tested in public exam centers, alleviating concern that achievement gains reflect changes in the composition of examinees rather than genuine learning gains.[22] As shown in Appendix Figure A7, the results are also not sensitive to removing from the analytical sample the four districts that received the program in 2018 and had participated in a small-scale implementation pilot the year prior (2017).

6 Magnitudes and cost-effectiveness

6.1 Expressing impacts in equivalent years of learning

Following Evans and Yuan (2019), to facilitate interpretation, we translate the standard deviation impacts observed in the randomized trial into equivalent years of learning. This approach compares treatment effects to learning gains observed in the control group over the two-year study period, leveraging our IRT-based vertical linking of baseline and endline scores.[23]

Table 1 presents control-group learning trajectories for the balanced panel of students observed at baseline and endline. Control students gained 1.10 SD in literacy and 1.31 SD in mathematics over two years. Assuming that standard deviation gains accumulate

[22] Students enrolled in public primary schools cannot take their leaving exam in private test centers.

[23] While recent literature has proposed Learning-Adjusted Years of Schooling (LAYS) as an alternative metric (Angrist et al., 2024), we express effects relative to observed control-group trajectories. This approach avoids the additional assumptions about how test-score differences map into long-run productivity required by LAYS and directly measures learning progression using our panel data structure rather than imputing from cross-sectional patterns.

approximately proportionally over time, our treatment effects of 0.10 SD and 0.15 SD thus represent 2.2 months and 2.8 months of additional learning over the two-year period, respectively (approximately 1.1-1.4 months per year). For targeted mathematics skills, where control students gained 1.56 SD, the treatment effect of 0.40 SD translates to 6.1 months of additional learning (approximately one-quarter of the two-year period).

6.2 Cost-effectiveness

Appendix E summarizes our cost-effectiveness analysis, with Table E1 reporting program costs by variant. The regular *Catch Up* program cost $9.63 per student over the two-year intervention period. Adding the continuous professional development component increased costs by $10.34 per student, for a total of $19.97.

Cost-effectiveness varies with the choice of outcome measure in the short run.[24] For the two-year randomized trial, the regular program costs $6.38 per 0.1 SD in mathematics and $9.53 per 0.1 SD in literacy. For the targeted subset of foundational mathematics skills, the cost falls to $2.43 per 0.1 SD.[25] These findings highlight how outcome measurement shapes conclusions about program effectiveness. In the short run, using proximal outcomes for mathematics yields cost-effectiveness estimates that are 2.6 times more favorable than those from comprehensive assessments of foundational skills.

Finally, the added CPD component doubled per-student costs without improving effectiveness. In the randomized trial, we found no difference in learning gains between the $9.63 basic program and the $19.97 enhanced version, suggesting that additional teacher support did not improve student learning outcomes, despite teachers' engagement with the added program component.

7 Conclusion

This paper asked whether narrowly targeted foundational instruction generates broad, long-term human capital gains. We studied Zambia's national remedial program, delivered by government teachers across public primary schools, using two complementary designs: a cluster-randomized trial and an event study spanning the universe of public grade-7 examinations. After two years of intervention, the experimental estimates show that the program improved foundational literacy by 0.10 standard deviations and numeracy by 0.15

[24] Because our cost data were collected for the locations and period of the randomized trial, we do not extend these calculations to the event-study estimates reported in Section 5.

[25] Cost-effectiveness (cost per 0.1 SD) is calculated as program cost per student divided by effect size, multiplied by 0.1.

standard deviations. In mathematics, short-run gains were concentrated in the procedural skills the program directly targets: effects on these skills were 2.6 times larger than on comprehensive assessments, with near-zero transfer to adjacent domains. Doubling per-pupil costs through continuous professional development for teachers yielded no additional learning gains. Despite this short-run pattern of narrow, domain-specific effects, event-study estimates that exploit the program's staggered district-level rollout show positive effects on primary school-leaving exam scores in early adolescence. Among cohorts expected to have been fully exposed to the program, scores improved by 0.14 standard deviations in literacy and 0.11 standard deviations in mathematics on assessments that cover competencies well beyond what the program targets.

Taken together, these results yield three sets of implications. First, the contrast between short-run specificity and long-run breadth is consistent with a temporal dimension of foundational skill formation that has received limited direct empirical attention in the literature. The short-run pattern is consistent with a production technology in which procedural competencies and broader numerical reasoning are complements realized across grade levels rather than within a single assessment window. Our findings cannot isolate the precise channel; dynamic skill complementarity, changes in classroom instruction, and shifts in student engagement are difficult to distinguish with the available data. But the pattern implies that short-run assessments of non-targeted skills may understate a program's long-run value, just as proximal assessments of targeted skills may overstate its breadth. Second, the divergence between targeted and comprehensive assessments has implications for how researchers evaluate and compare educational interventions. Within the same program setting, the estimated cost-effectiveness differs by a factor of 2.6 depending on the outcome measure, a magnitude large enough to matter for cross-study comparisons that pool estimates from studies using different assessment instruments. Third, the positive effects of a government-run, teacher-led program operating at scale under real-world constraints indicate that the positive results documented in NGO-implemented evaluations are not confined to settings with dedicated facilitators or parallel delivery structures. At the same time, the null effect of the professional development add-on we tested suggests that, in this setting, the base program was sufficient to generate learning gains, whereas additional peer-based support did not improve outcomes further.

Tables and figures

Table 1: *Balancing checks between experimental groups*

	Control mean	Differences		
		T vs C	No CPD vs C	CPD vs C
	(1)	(2)	(3)	(4)
Panel A: Full baseline sample of students				
Literacy	-1.092	-0.005	-0.012	0.002
	[0.910]	(0.044)	(0.048)	(0.048)
Math	-1.303	-0.054	-0.041	-0.067
	[1.016]	(0.051)	(0.056)	(0.057)
Targeted math skills	-1.558	-0.028	-0.020	-0.036
	[0.884]	(0.045)	(0.051)	(0.051)
Student is female	0.511	-0.000	0.009	-0.010
	[0.500]	(0.009)	(0.011)	(0.011)
Asset index	0.000	-0.030	-0.029	-0.031
	[1.000]	(0.053)	(0.056)	(0.059)
Best friend in school, attends the same grade	0.691	-0.007	0.006	-0.021
	[0.462]	(0.017)	(0.019)	(0.021)
Home language different from schools' language	0.322	0.011	0.021	0.001
	[0.467]	(0.030)	(0.032)	(0.034)
Attrition	0.124	0.014	0.016	0.012
	[0.330]	(0.011)	(0.012)	(0.012)
F-statistic		0.746	0.761	1.015
p-value		0.651	0.637	0.426
Panel B: Baseline students observed at endline				
Literacy	-1.095	-0.001	-0.008	0.005
	[0.910]	(0.046)	(0.050)	(0.051)
Math	-1.313	-0.043	-0.028	-0.058
	[1.019]	(0.052)	(0.058)	(0.058)
Targeted math skills	-1.562	-0.020	-0.008	-0.033
	[0.887]	(0.046)	(0.052)	(0.052)
Student is female	0.506	-0.006	0.008	-0.020
	[0.500]	(0.010)	(0.012)	(0.012)
Asset index	0.000	-0.033	-0.032	-0.034
	[1.004]	(0.054)	(0.058)	(0.060)
Best friend in school, attends the same grade	0.694	-0.010	0.002	-0.022
	[0.461]	(0.018)	(0.020)	(0.021)
Home language different from schools' language	0.318	0.017	0.029	0.005
	[0.466]	(0.030)	(0.032)	(0.034)
F-stat		0.593	0.488	1.299
p-value		0.761	0.843	0.253

Notes. This table compares students in the control and treatment groups. Panel A includes all students sampled at baseline; Panel B includes students observed at baseline and endline (i.e., non-attriting students). The table shows the control-group mean and corresponding standard deviations for each variable (in brackets), and it compares both experimental groups, including randomization-strata fixed effects, showing the mean difference and corresponding standard errors (clustered at the zone level, in parentheses). Baseline test scores are expressed on the same scale as the distribution of endline scores; they are standardized such that the control group at endline has a mean of zero and a standard deviation of one.

Table 2: *Intent-to-treat effects on foundational skills*

	Pooled ITT effect	ITT effects by program variant		
		Without CPD	With CPD	Effect of CPD
	(1)	(2)	(3)	(4)
Literacy	0.101	0.109	0.092	-0.017
	(0.033)	(0.037)	(0.039)	(0.040)
Math	0.151	0.152	0.150	-0.003
	(0.042)	(0.045)	(0.049)	(0.045)
Targeted math skills	0.397	0.389	0.405	0.015
	(0.039)	(0.044)	(0.050)	(0.051)
	[0.001]			

Notes. This table shows the impact of the intervention on assessments of literacy, math, and a preregistered subdomain of math that the intervention focuses on. Estimates come from regressions of endline test scores on a treatment indicator with controls for randomization strata and baseline characteristics chosen with post-double Lasso selection. Endline scores are standardized to have a mean of zero and a standard deviation of one in the control group at endline. Standard errors, in parentheses, are clustered at the zone level. Simes FDR q-values, in brackets, follow a prespecified order of tests across the impacts investigated in the study; we do not provide adjustments for the two main prespecified outcomes. "Without CPD" and "With CPD" refer to the two treatment arms without and with the continuous professional development component for teachers; "Effect of CPD" refers to their difference. As pre-registered, since there is no meaningful difference in effects, we focus the paper (and its adjustment of hypothesis tests) on effects that pool the two treatment arms.

Table 3: *Intent-to-treat effects on additional outcomes and mechanisms*

	Pooled ITT effect	ITT effects by program variant		
		Without CPD	With CPD	Effect of CPD
	(1)	(2)	(3)	(4)
Panel A: Additional outcomes				
Working memory	0.065	0.032	0.100	0.068
	(0.030)	(0.033)	(0.034)	(0.029)
	[0.038]			
Creativity	-0.016	-0.029	-0.003	0.026
	(0.034)	(0.037)	(0.041)	(0.040)
	[0.628]			
Socio-emotional skills	-0.026	0.002	-0.054	-0.055
	(0.036)	(0.041)	(0.043)	(0.041)
	[0.562]			
Panel B: Mechanisms				
Attitudes	0.044	0.003	0.087	0.085
	(0.031)	(0.036)	(0.037)	(0.038)
	[0.156]			
Studied at home (yesterday)	0.039	0.032	0.047	0.015
	(0.013)	(0.016)	(0.016)	(0.017)
	[0.005]			
Teacher collaboration and feedback	0.132	0.189	0.070	-0.119
	(0.087)	(0.101)	(0.103)	(0.105)
	[0.156]			

Notes. This table shows the impact of the intervention on secondary outcomes (Panel A) and potential mechanisms (Panel B). Estimates come from regressions of outcomes on a treatment indicator with controls for randomization strata and baseline characteristics chosen with post-double Lasso selection. Outcomes are standardized to have a mean of zero and a standard deviation of one in the control group at endline. Standard errors, in parentheses, are clustered at the zone level. Simes FDR q-values, in brackets, follow a pre-specified order of tests across the impacts investigated in the study; we do not provide adjustments for the two main, pre-specified outcomes. "Without CPD" and "With CPD" refer to the two treatment arms without and with the continuous professional development component for teachers; "Effect of CPD" refers to their difference. As pre-registered, since there is no meaningful difference in effects, we focus the paper (and its adjustment of hypothesis tests) on effects that pool the two treatment arms.

Table 4: *Intent-to-treat effects among pre-specified subgroups of students*

	Pooled ITT Effect (1)
Panel A: Literacy	
Girls	0.124
	(0.040)
	[0.004]
Low-performing quartile	0.151
	(0.056)
	[0.008]
Panel B: Math	
Girls	0.161
	(0.047)
	[0.002]
Low-performing quartile	0.139
	(0.058)
	[0.018]

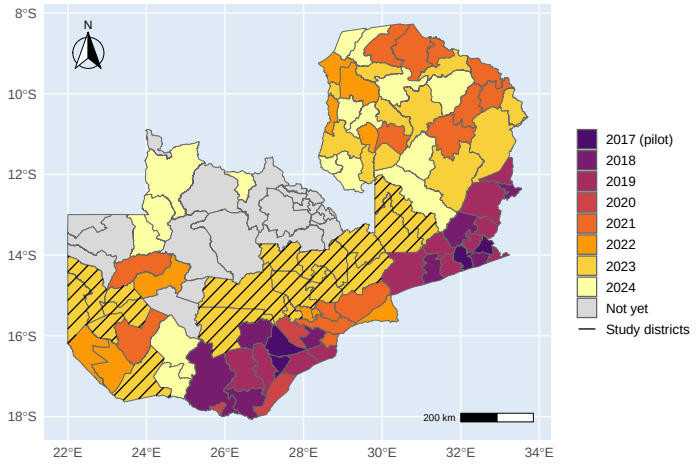
Notes. This table shows the impact of the intervention on assessments of literacy and math among two pre-specified subgroups: girls and students who, at baseline, performed in the bottom quartile of the performance distribution of the respective subject. Estimates come from regressions of endline test scores on a treatment indicator and treatment-subgroup interactions, with controls for randomization strata and baseline characteristics chosen with post-double Lasso selection. Endline scores are standardized to have a mean of zero and a standard deviation of one in the control group at endline. Standard errors, in parentheses, are clustered at the zone level. Simes FDR q-values, in brackets, follow a pre-specified order of tests across impacts investigated in the study. As pre-registered, since there is no meaningful difference in effects, we focus the paper (and its adjustment of hypothesis tests) on effects that pool the two treatment arms.

Table 5: *Treatment effects on non-targeted math skills and teacher beliefs*

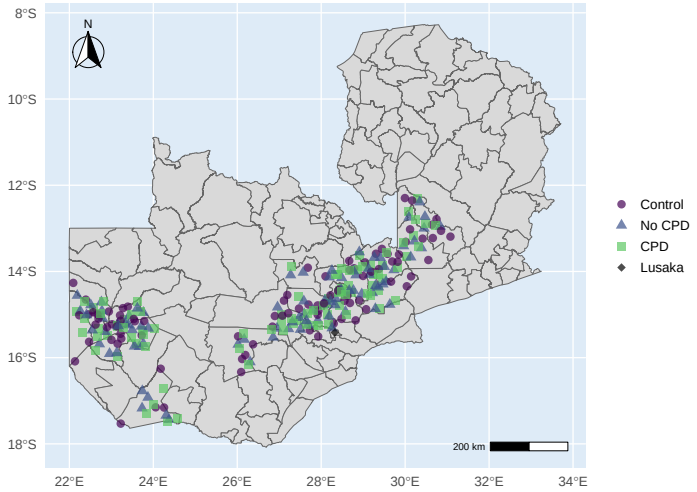
	Pooled ITT Effect (1)
Panel A: Non-targeted math skills	
Non-targeted math skills	0.028
	(0.041)
	[0.490]
Non-targeted number and arithmetic skills	0.047
	(0.038)
	[0.271]
Panel B: Teacher beliefs	
Teacher locus of control	0.059
	(0.070)
	[0.457]
Overestimate (vs. ASER paragraph level)	-0.000
	(0.028)
	[1.000]
Overestimate (vs. subtraction question)	-0.141
	(0.030)
	[0.001]

Notes. This table shows the impact of the intervention on non-targeted math skills and teacher beliefs. Panel A shows intent-to-treat effects on the subdomains of math skills not targeted by the program (i.e., the complement of the prespecified targeted math domain). Panel B shows intent-to-treat effects on additional teacher-level outcomes (which were de-prioritized in the pre-analysis plan). Endline scores and the measure of locus of control are standardized to have a mean of zero and a standard deviation of one in the control group at endline. "Overestimate" measures the gap between a teacher's estimate of the proportion of students who can solve a task (as elicited during process monitoring) and the observed student performance in her school (as measured at endline). For literacy, this comparison focuses on students' ability to read a short paragraph included in the ASER test; for math, it focuses on a subtraction item. For reference, in the control group, teachers overestimate the proportion of students who can read a short paragraph by 0.244 and the proportion of students who can solve the subtraction question by 0.320. Estimates come from regressions of outcomes on the treatment indicator, with controls for randomization strata and baseline characteristics chosen with post-double Lasso selection. Standard errors, in parentheses, are clustered at the zone level. In Panel A, Simes FDR q-values (in brackets) add to the tests adjusted for in Table 2. In Panel B, they add to tests adjusted for in Panel B of Table 3.

Figure 1: *Geographic scope of the study*



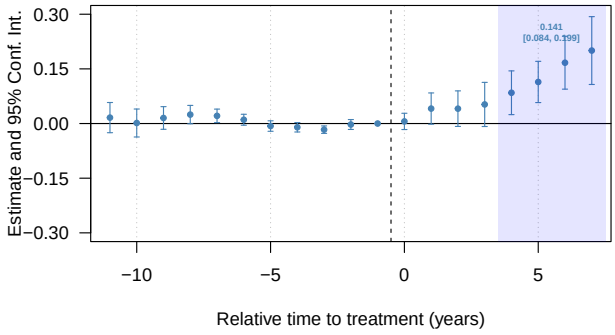
(a) *Staggered program roll-out in Zambia*



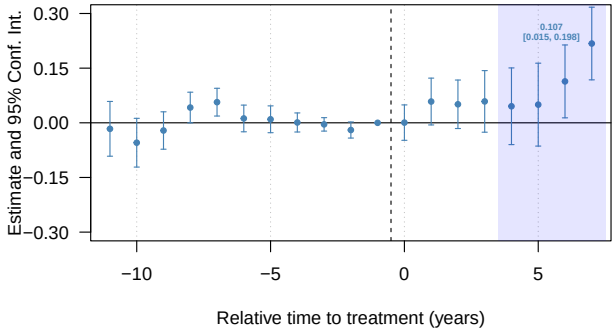
(b) *Subsample of study schools by random assignment status*

Note: This figure shows the geographic scope of the study. The top panel maps the program's phased rollout across Zambia's 116 districts. Colors indicate the year the program was first introduced in each district. In 2017, some schools in four districts participated in a small-scale pilot that tested alternative implementation modalities; these districts first received the program described in this paper in 2018. Cross-hatching indicates the districts of the randomized trial, in which 91 zones were randomly assigned to receive the program in 2023 (treatment) and 91 zones were assigned to receive it in 2025 or later (control). In 2025, control zones outside Western Province received the program; control zones in Western Province have yet to receive it. In 2025, no new districts were phased in. The bottom panel maps the 273 schools in the trial's primary data collection sample. Each dot represents a school, with colors indicating experimental groups. "CPD" and "No CPD" refer to the two treatment arms—with and without the continuous professional development component, respectively.

Figure 2: *Event-study estimates of effects on grade-7 exam scores*



(a) *Language (Zambian)*



(b) *Mathematics*

Note: This figure reports event-study estimates of the program's intent-to-treat effect on grade-7 primary school leaving exam scores using Gardner's (2022) two-stage difference-in-differences estimator and exploiting the program's staggered roll-out (see Figure 1). Panel (a) shows results for scores in the Zambian language exam (languages differ by district); panel (b) shows results for exam scores in mathematics. The horizontal axis represents years relative to program roll-out, where period 0 indicates the first year a district or zone received the program. The vertical axis shows the treatment effect in student-level standard deviations. Blue shading indicates cohorts expected to have been in grades 1 to 3 when the program rolled out; as the program aims to support learners as they move from grade 3 to 5, absent grade retention, these cohorts can be expected to have been fully exposed to the program. Points represent coefficient estimates for each event-time period, with vertical bars indicating 95-percent confidence intervals based on standard errors clustered at the district level (116 clusters). We also report the pooled effects across the four cohorts expected to have been fully exposed to the program (and provide the 95-percent confidence interval in brackets). Test scores are standardized using the pooled student-level standard deviation calculated within each year across all examination centers nationally, accounting for both within- and between-center variation. Each observation represents a center-year-level mean, weighted by the number of students tested. The sample includes all public grade-7 national assessment examination centers from 2014 to 2025 (6,618 unique centers, with 4,439,847 language scores and 4,461,130 mathematics scores). Program rollout occurred at the district level, except for control zones from the randomized controlled trial (some of which received the intervention in 2025). The reference category consists of period -1 (one year before treatment), as well as all never-treated observations.

References

Abadie, A., Athey, S., Imbens, G.W., Wooldridge, J.M., 2022. When Should You Adjust Standard Errors for Clustering? The Quarterly Journal of Economics 138, 1–35. doi:10.1093/qje/qjac038.

Angrist, N., Evans, D.K., Filmer, D., Glennerster, R., Rogers, H., Sabarwal, S., 2024. How to improve education outcomes most efficiently? A review of the evidence using a unified metric. Journal of Development Economics , 103382doi:10.1016/j.jdeveco.2024.103382.

Angrist, N., Meager, R., 2023. Implementation Matters: Generalizing Treatment Effects in Education. Technical Report. Annenberg Institute at Brown University. URL: https://edworkingpapers.com/ai23-802.

Ardington, C., Menendez, A., Thunde, J., 2026. The Limits of a Great Buy: What Rigorous Evidence Reveals about Teaching at the Right Level. Technical Report. University of Cape Town. Cape Town. URL: https://www.datafirst.uct.ac.za/sites/default/files/media/documents/dfru_uct_ac_za/2872/aflearn-report-tarl.pdf.

Athey, S., Imbens, G., 2016. Recursive Partitioning for Heterogeneous Causal Effects. Proceedings of the National Academy of Sciences 113, 7353–7360. doi:10.1073/pnas.1510489113.

Athey, S., Wager, S., 2019. Estimating Treatment Effects with Causal Forests: An Application. Observational Studies 5, 37–51. doi:10.1353/obs.2019.0001.

Bailey, D.H., Duncan, G.J., Cunha, F., Foorman, B.R., Yeager, D.S., 2020. Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. Psychological Science in the Public Interest 21, 55–97. doi:10.1177/1529100620915848.

Bailey, D.H., Duncan, G.J., Odgers, C.L., Yu, W., 2017. Persistence and fadeout in the impacts of child and adolescent interventions. Journal of Research on Educational Effectiveness 10, 7–39. doi:10.1080/19345747.2016.1232459.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., Walton, M., 2017. From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application. Journal of Economic Perspectives 31, 73–102. doi:10.1257/jep.31.4.73.

Barnett, S.M., Ceci, S.J., 2002. When and where do we apply what we learn? A taxonomy for far transfer. Psychological Bulletin 128, 612–637. doi:10.1037/0033-2909.128.4.612.

de Barros, A., Ganimian, A.J., 2023. The Foundational Math Skills of Indian Children. Economics of Education Review 92, 102336. doi:10.1016/j.econedurev.2022.102336.

de Barros, A., Henry, J., Mathenge, J.W., 2024. What Drives Teachers to Change Their Instruction? A Mixed-Methods Study from Zambia. Comparative Education Review 68, 562–585. doi:10.1086/733519.

Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on Treatment Effects after Selection among High-Dimensional Controls. The Review of Economic Studies 81, 608–650. doi:10.1093/restud/rdt044.

Chernozhukov, V., Demirer, M., Duflo, E., Fernández-Val, I., 2025. Fisher–Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India. Econometrica 93, 1121–1164. doi:10.3982/ECTA19303.

Cilliers, J., Elashmawy, N., McKenzie, D., 2024. Using Post-Double Selection Lasso in Field Experiments. Working Paper 10931. The World Bank. Washington, D.C. URL: https://openknowledge.worldbank.org/entities/publication/0cde089d-33ba-4f51-8c03-b25b5114d41a.

Cunha, F., Heckman, J., 2007. The Technology of Skill Formation. American Economic Review 97, 31–47. doi:10.1257/aer.97.2.31.

Cunha, F., Heckman, J.J., Schennach, S.M., 2010. Estimating the technology of cognitive and noncognitive skill formation. Econometrica 78, 883–931. doi:10.3982/ECTA6551.

Detterman, D.K., 1993. The case for the prosecution: Transfer as an epiphenomenon, in: Detterman, D.K., Sternberg, R.J. (Eds.), Transfer on Trial: Intelligence, Cognition, and Instruction. Ablex, Norwood, NJ, pp. 1–24.

Duflo, A., Kiessel, J., Lucas, A.M., 2024. Experimental Evidence on Four Policies to Increase Learning at Scale. The Economic Journal , ueae003doi:10.1093/ej/ueae003.

Evans, D., Yuan, F., 2019. Equivalent Years of Schooling: A Metric to Communicate Learning Gains in Concrete Terms. Working Paper WPS8752. The World Bank. Washington, D.C. URL: http:

//documents.worldbank.org/curated/en/123371550594320297/
Equivalent-Years-of-Schooling-A-Metric-to-Communicate-Learning-Gains-in-Concre

Gardner, J., 2022. Two-Stage Differences in Differences. doi:10.48550/arXiv.2207.05943. arXiv:2207.05943 [econ].

GEEAP, 2023. Cost-Effective Approaches to Improve Global Learning: What Does Recent Evidence Tell Us Are Smart Buys for Improving Learning in Low- and Middle-Income Countries? Technical Report. The World Bank. Washington, D.C. URL: https://thedocs.worldbank.org/en/doc/231d98251cf326922518be0cbe306fdc-0200022023/related/GEEAP-Report-Smart-Buys-2023-final.pdf.

Lipovsek, V., Poswell, L., Morrell, A., Pershad, D., Vromant, N., Grindle, A., 2023. Reflections on Systems Practice: Implementing Teaching at the Right Level in Zambia, in: Faul, M., Savage, L. (Eds.), Systems Thinking in International Education and Development. Edward Elgar Publishing, Cheltenham, UK. Political Science and Public Policy, pp. 27–46. doi:10.4337/9781802205930.00012.

Perkins, D.N., Salomon, G., 1992. Transfer of learning, in: Husén, T., Postlethwaite, T.N. (Eds.), International Encyclopedia of Education. 2nd ed.. Pergamon Press, Oxford.

Popova, A., Evans, D.K., Breeding, M.E., Arancibia, V., 2022. Teacher Professional Development around the World: The Gap between Evidence and Practice. The World Bank Research Observer 37, 107–136. doi:10.1093/wbro/lkab006.

Appendices

Appendix A: Additional tables and figures

Table A1: *Representativeness of schools in the randomized controlled trial*

	Sample		Other		Difference
	n	mean [s.d.]	*n*	mean [s.d.]	
	(1)	(2)	(3)	(4)	(5)
Panel A: Schools in study zones vs. other schools in the country					
Grade-3 enrollment	1,047	63.172	5,114	60.481	2.690
		[51.153]		[52.214]	(5.292)
Grade-4 enrollment	1,047	63.023	5,114	60.043	2.980
		[52.439]		[59.360]	(5.549)
Grade-5 enrollment	1,047	60.703	5,114	57.283	3.420
		[57.716]		[65.933]	(6.148)
Enrollment, total	1,047	471.351	5,114	472.214	-0.862
		[388.682]		[433.932]	(45.175)
Proportion female	1,047	0.504	5,114	0.503	0.000
		[0.048]		[0.050]	(0.002)
Started as non-government school	1,114	0.349	5,641	0.286	0.063
		[0.477]		[0.452]	(0.032)
Rural	1,114	0.912	5,641	0.890	0.022
		[0.283]		[0.313]	(0.045)
F-statistic					1.721
p-value					0.111
Panel B: Subsample of schools vs. other schools in study zones					
Grade-3 enrollment	263	64.605	784	62.691	1.913
		[52.304]		[50.786]	(3.506)
Grade-4 enrollment	263	64.943	784	62.379	2.564
		[55.863]		[51.261]	(3.503)
Grade-5 enrollment	263	61.700	784	60.369	1.331
		[58.629]		[57.441]	(3.644)
Enrollment, total	263	487.981	784	465.773	22.208
		[393.436]		[387.167]	(25.228)
Proportion female	263	0.505	784	0.503	0.002
		[0.057]		[0.044]	(0.004)
Started as non-government school	273	0.344	841	0.351	-0.006
		[0.476]		[0.477]	(0.031)
Rural	273	0.927	841	0.907	0.019
		[0.261]		[0.290]	(0.018)
F-statistic					0.852
p-value					0.546

Notes. This table reports on the representativeness of schools in the randomized controlled trial, using data from Zambia's education management information system (EMIS, as of 2020). Panel A compares all 1,115 randomized schools in the sample of 182 study zones with all other government primary schools in the country. Panel B compares the study's random subsample of 273 schools with the remaining government primary schools in the study's 182 zones. While all study schools can be matched to the EMIS, the EMIS lacks enrollment data for some; effective sample sizes are shown in columns (1) and (3). Columns (2) and (4) show the mean and corresponding standard deviations (in brackets) for each variable. Column (5) shows their difference. The EMIS does not include reliable information on zones; this information is only available for study zones. Therefore, standard errors (shown in parentheses) are clustered at the district level in Panel A and clustered at the zone level in Panel B. *F* tests rely on list-wise deletion of observations with missing information; the *F* test in Panel A includes 6,161 (out of 6,755) schools in the country; the corresponding test in Panel B includes 1,047 (out of 1,115) schools in study zones.

Table A2: *Program implementation and exposure*

	Pooled treatment mean	By program variant		
		Without CPD	With CPD	Effect of CPD
	(1)	(2)	(3)	(4)
Panel A: Implementation of TaRL				
School submitted TaRL assessment data	0.978	0.989	0.967	-0.022
	[0.147]	[0.105]	[0.180]	(0.025)
School implemented program in term 2	0.956	0.978	0.934	-0.044
	[0.206]	[0.147]	[0.250]	(0.036)
School implemented program in term 3	0.868	0.890	0.846	-0.044
	[0.339]	[0.314]	[0.363]	(0.057)
TaRL mentor available in school	0.868	0.846	0.890	0.044
	[0.339]	[0.363]	[0.314]	(0.057)
Teacher trained on TaRL (among any teachers)	0.475	0.473	0.476	0.000
	[0.499]	[0.500]	[0.500]	(0.000)
Number of weekdays with TaRL classes scheduled	4.283	4.124	4.440	0.301
	[1.562]	[1.744]	[1.352]	(0.263)
Panel B: Student attendance (yesterday)				
Attended school	0.504	0.488	0.520	0.030
	[0.500]	[0.500]	[0.500]	(0.026)
Attended any TaRL class	0.333	0.327	0.340	0.013
	[0.471]	[0.469]	[0.474]	(0.037)
Attended TaRL literacy class	0.259	0.258	0.260	-0.004
	[0.438]	[0.438]	[0.439]	(0.039)
Attended TaRL math class	0.215	0.218	0.211	-0.016
	[0.411]	[0.413]	[0.408]	(0.039)

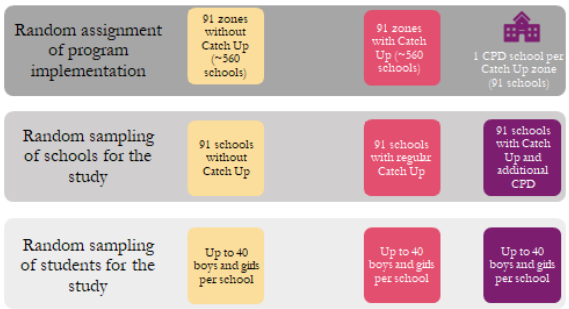
Notes. This table reports program implementation measures (Panel A) and student attendance and exposure to the program on the previous school day (Panel B). Columns (1) to (3) show the treatment-group means and corresponding standard deviations (in brackets), for the pooled sample of treatment schools and by program variant. Column (4) reports the coefficient on an indicator for assignment to the CPD variant from a regression estimated on the treated sample only with randomization-strata fixed effects; standard errors, clustered at the zone level, are reported in parentheses. Implementation measures in Panel A are at the school level, except for the teacher-training measure, which is at the teacher level. The latter includes all teachers on a school's payroll (grades 1 to 7), while the program and its CPD component targeted grades 3 to 5. Attendance and exposure measures in Panel B are at the student level. The sample comprises all students sampled at baseline, with students who have left the school treated as absent. Panel A draws on primary data collected during monitoring, teacher training data, and endline headteacher surveys; Panel B draws on enumerator-collected data gathered during the endline school visits.

Table A3: *Exploration of heterogeneous treatment effects on foundational skills*

	Importance (1)	Subgroup indicator (2)	CATE (3)	Standard error (4)	Weak group (5)	Strong group (6)	Difference (7)
Panel A: Literacy							
Gini inequality	0.366	Low	0.091	0.072	0.291	0.217	-0.074
		High	0.188	0.079	0.124	0.369	0.244
Asset index	0.167	Low	0.145	0.055	0.220	0.322	0.102
		High	0.075	0.055	0.294	0.199	-0.095
Baseline literacy	0.161	Low	0.135	0.062	0.207	0.293	0.086
		High	0.059	0.048	0.321	0.185	-0.136
Rural	0.095	Yes	0.106	0.035	0.905	0.914	0.009
		No	0.134	0.145	0.095	0.086	-0.009
Within-school performance	0.094	Low	0.101	0.052	0.237	0.280	0.043
		High	0.110	0.055	0.281	0.195	-0.086
Different home language	0.048	Yes	0.098	0.053	0.337	0.320	-0.016
		No	0.114	0.037	0.663	0.680	0.016
Female	0.037	Yes	0.128	0.043	0.455	0.549	0.094
		No	0.089	0.038	0.545	0.451	-0.094
Best friend in same school, grade	0.031	Yes	0.106	0.037	0.707	0.675	-0.032
		No	0.116	0.048	0.293	0.325	0.032
Group-average ITT effect					0.164	0.054	-0.110
					(0.041)	(0.043)	(0.059)
RATE p-value							0.364
Panel B: Mathematics							
Gini inequality	0.420	Low	0.177	0.075	0.241	0.261	0.020
		High	0.267	0.062	0.129	0.360	0.231
Asset index	0.121	Low	0.154	0.065	0.288	0.254	-0.034
		High	0.170	0.061	0.209	0.284	0.075
Rural	0.120	Yes	0.153	0.043	0.935	0.884	-0.052
		No	0.215	0.198	0.065	0.116	0.052
Baseline math	0.109	Low	0.140	0.060	0.283	0.216	-0.067
		High	0.126	0.062	0.273	0.225	-0.047
Within-school performance	0.102	Low	0.099	0.060	0.313	0.205	-0.108
		High	0.153	0.059	0.269	0.205	-0.064
Best friend in same school, grade	0.052	Yes	0.140	0.046	0.729	0.653	-0.076
		No	0.201	0.061	0.271	0.347	0.076
Different home language	0.051	Yes	0.153	0.057	0.391	0.266	-0.124
		No	0.161	0.050	0.609	0.734	0.124
Female	0.025	Yes	0.163	0.048	0.496	0.509	0.013
		No	0.154	0.048	0.504	0.491	-0.013
Group-average ITT effect					0.126	0.191	0.066
					(0.054)	(0.044)	(0.069)
RATE p-value							0.612

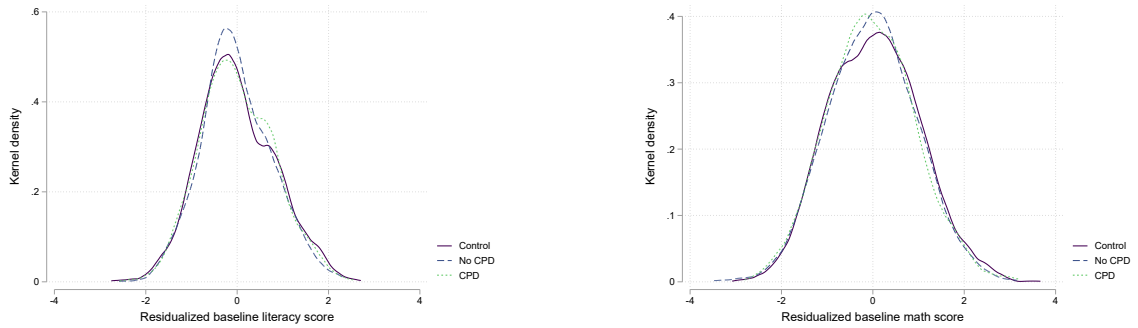
Notes. Using causal forests, this table explores additional heterogeneity in intent-to-treat effects on students' foundational literacy (Panel A) and mathematics skills (Panel B), going beyond the pre-registered subgroup analyses presented in Table 4. Column (1) reports the variable importance from the causal forest, representing the proportion of times each variable is selected for splitting when building trees in the forest. Column (2) indicates the subgroup, where for quartile variables, "Low" refers to students in the bottom quartile (Q1) and "High" refers to students in the top quartile (Q4). The inequality measure captures within-school heterogeneity in baseline achievement, calculated as the Gini coefficient of baseline test scores for each school; students are then categorized into quartiles based on their school's Gini coefficient, with "Low" representing students in schools with the most equal baseline achievement and "High" representing students in schools with the most unequal baseline achievement. Results for the remaining two quartiles are omitted from the table. Column (3) reports the conditional average treatment effect (CATE) for each subgroup, and column (4) reports its standard error (clustered at the zone level). In columns (5) and (6), "weak group" refers to students whose predicted CATE is below the median of the forest's out-of-bag predictions, and "strong group" to students whose predicted CATE is above the median. The column entries report the proportions of students belonging to a row's subgroup; column (7) reports the difference in proportions. "Group-avarage ITT effect" at the bottom of each panel reports the AIPW-estimated intent-to-treat effect within each group (Athey and Wager, 2019), with standard errors in parentheses; the final column reports the difference, with its standard error computed assuming independence across the two non-overlapping groups. "RATE p-value" reports the two-sided p-value from a sequential cross-validation test of the rank-weighted average treatment effect (AUTOC), using cluster-respecting folds ($K = 5$), which provides a formal test of whether the forest's treatment effect ranking captures genuine heterogeneity (Chernozhukov et al., 2025).

Figure A1: *Sampling and randomization procedure*



Note: This figure summarizes the study's sampling and randomization procedure. Within the trial's sample of 182 zones, we randomized half the zones to the Catch Up program and the other half to the control group. Within each control zone, we randomly sampled one school for the study. Within each program zone, we randomly sampled two schools for the study. We randomly assigned one of these two Catch Up schools to receive the program with the additional continuous professional development (CPD) component. Within each sampled school, we randomly sampled up to 40 boys and girls for the study (stratified by gender).

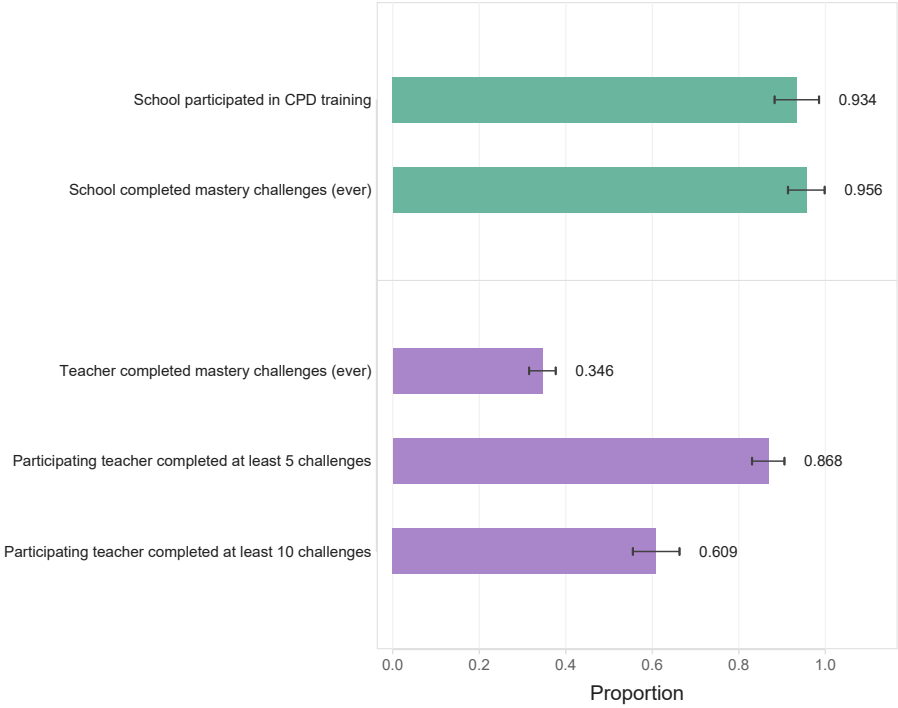
Figure A2: *Balance on baseline test scores*



(a) *Literacy* **(b)** *Mathematics*

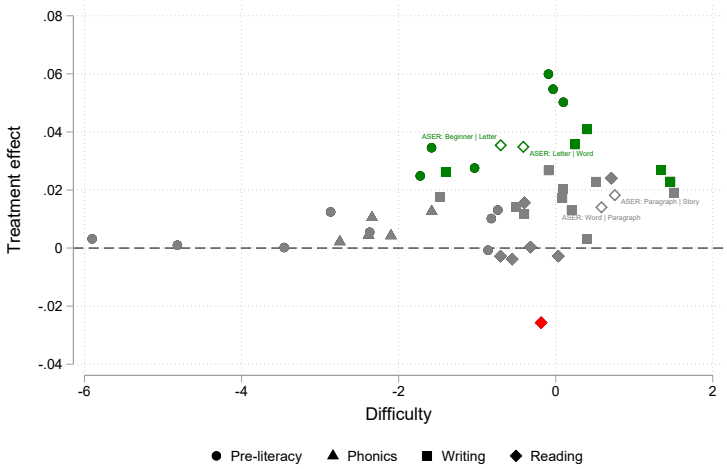
Note: This figure reports on the sample's balance across the three experimental groups, as per the baseline tests in literacy and mathematics. Test scores are aggregated using a hybrid item response theory (IRT) model—two-parameter logistic for dichotomous items and graded response for ordinal items—then standardized and centered with respect to the control group at endline. Each panel shows kernel density plots, by treatment status, of residuals from a regression of baseline test scores on strata fixed effects. "CPD" and "No CPD" refer to the two treatment groups assigned to receive the program—with and without the continuous professional development component, respectively. The left panel reports results for literacy; the right panel reports results for mathematics.

Figure A3: *Take-up of continuous professional development (CPD) activities*

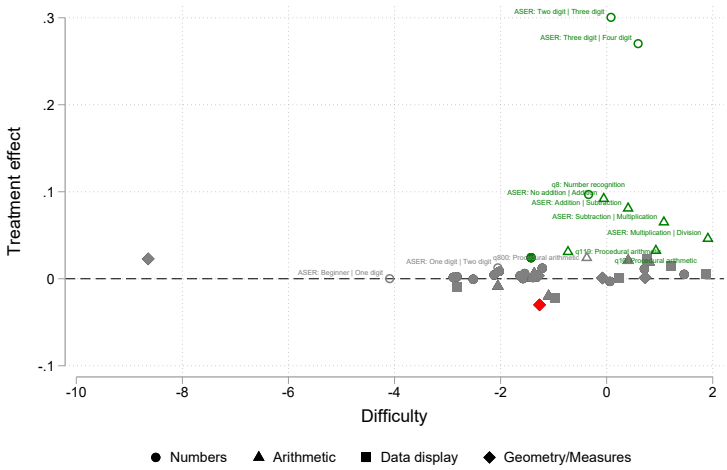


Note: This figure presents participation rates in continuous professional development (CPD) activities. The top panel displays school-level indicators showing the proportion of schools with at least one teacher participating in each CPD activity (per school, one deputy headteacher and one teacher were invited to off-site CPD trainings; they were then asked to onboard the remaining teachers in their school). The bottom panel displays teacher-level indicators showing the proportion of individual teachers participating in each activity. Teachers were encouraged to complete the mastery challenges as a team, which could result in one submission for multiple teachers. Horizontal bars represent means with 95-percent confidence intervals indicated by error bars. The sample includes all 91 schools assigned to receive the program with the additional CPD component, and all teachers employed in these schools (as per official payroll data). The sample includes teachers across all grade levels (1 to 7), while the program and its CPD component targeted specific grades (3 to 5), which attenuates teacher-level participation rates.

Figure A4: *Item-wise program effects*



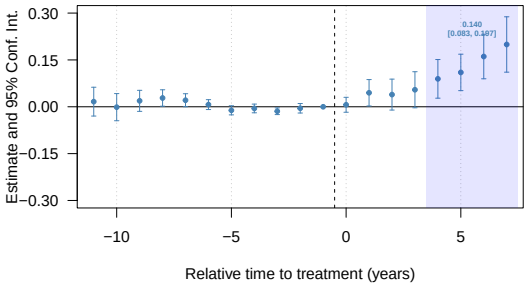
(a) *Literacy*



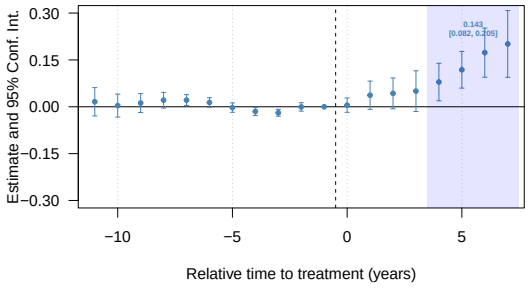
(b) *Mathematics*

Note: This figure reports on the program's pooled intent-to-treat effect on a student's probability of solving each assessment item correctly (for binary items) or passing a given skill threshold (for ordinal items), for literacy (Panel A) and mathematics (Panel B). The y-axis reflects these program effects; the x-axis reflects each item's or threshold's difficulty parameter from the study's hybrid item response model (compare to Appendix Tables D2 and D3). We omit one literacy item with a very low difficulty parameter (-21.564). The shape of markers reflects each item's content domain; we group listening comprehension, phonemic awareness, and vocabulary as pre-literacy skills. For literacy, hollow markers and marker labels denote thresholds related to ASER tests; for mathematics, they capture the preregistered subdomain of number recognition and procedural arithmetic (including the ASER thresholds and four other items). Estimates come from linear probability models of indicators on a treatment indicator with controls for randomization strata and baseline characteristics chosen with post-double Lasso selection. Standard errors are clustered at the zone level. Green color indicates positive and significant effects, gray indicates non-significant effects, and red indicates negative significant effects at conventional levels ($p < 0.05$). Item-wise results and their statistical significance should be considered with caution—they were not prespecified and do not account for multiple hypothesis testing.

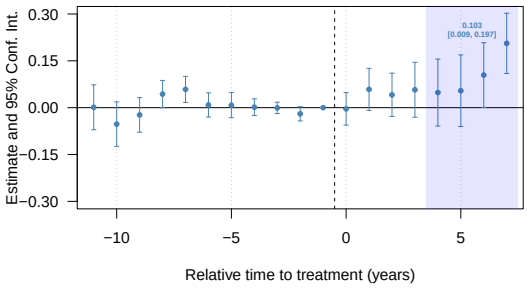
Figure A5: *Event-study estimates of effects on grade-7 exam scores, by gender*



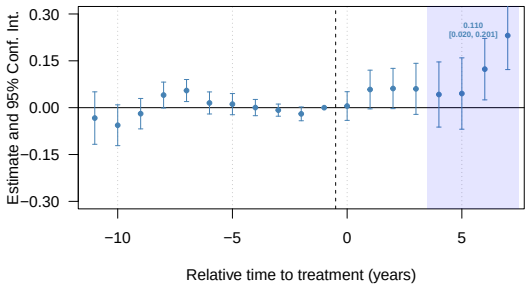
(a) *Language (Zambian) – Girls*



(b) *Language (Zambian) – Boys*



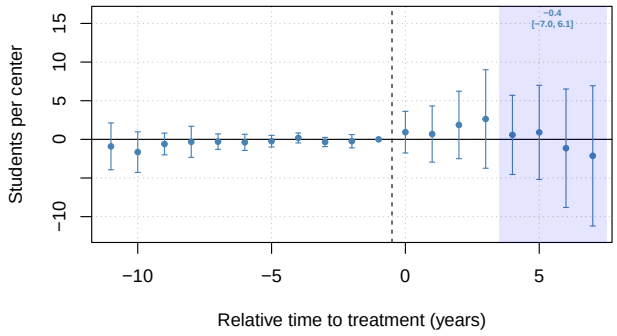
(c) *Mathematics – Girls*



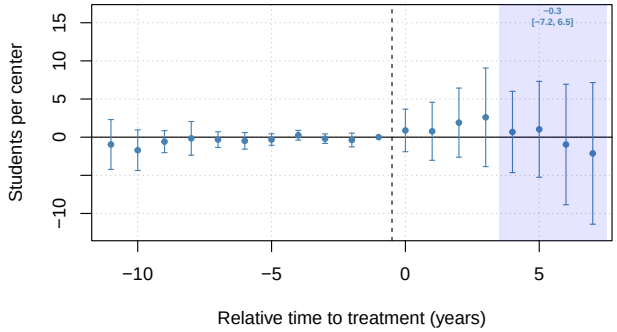
(d) *Mathematics – Boys*

Note: This figure reports event-study estimates of the program's intent-to-treat effect on grade-7 primary school leaving exam scores using Gardner's (2022) two-stage difference-in-differences estimator and exploiting the program's staggered roll-out (see Figure 1). Panels (a) and (b) show results for scores in the Zambian language exam (languages differ by district) for girls and boys, respectively; panels (c) and (d) show results for exam scores in mathematics for girls and boys, respectively. The horizontal axis represents years relative to program roll-out, where period 0 indicates the first year a district or zone received the program. The vertical axis shows the treatment effect in student-level standard deviations. Blue shading indicates cohorts expected to have been in grades 1 to 3 when the program rolled out; as the program aims to support learners as they move from grade 3 to 5, absent grade retention, these cohorts can be expected to have been fully exposed to the program. Points represent coefficient estimates for each event-time period, with vertical bars indicating 95-percent confidence intervals based on standard errors clustered at the district level (116 clusters). We also report the pooled effects across the four cohorts expected to have been fully exposed to the program (and provide the 95-percent confidence interval in brackets). Test scores are standardized using the pooled student-level standard deviation calculated within each year across all examination centers nationally, accounting for both within- and between-center variation. Each observation represents a center-year-level mean, weighted by the number of students tested. The sample includes all public grade-7 national assessment examination centers from 2014 to 2025 (6,618 unique centers, with 4,439,847 language scores and 4,461,130 mathematics scores). Program rollout occurred at the district level, except for control zones from the randomized controlled trial (some of which received the intervention in 2025). The reference category consists of period -1 (one year before treatment), as well as all never-treated observations.

Figure A6: *Event-study estimates of effects on grade-7 test taking*



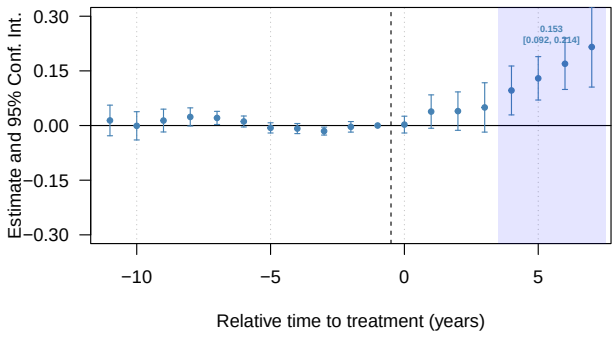
(a) *Language (Zambian)*



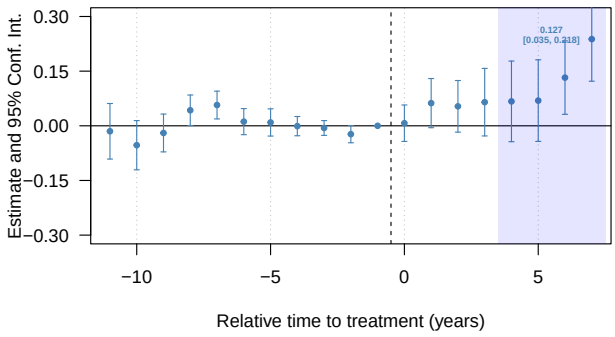
(b) *Mathematics*

Note: This figure reports event-study estimates of the program's intent-to-treat effect on grade-7 primary school leaving exam test taking using Gardner's (2022) two-stage difference-in-differences estimator and exploiting the program's staggered roll-out (see Figure 1). Panel (a) shows results for a test center's number of students sitting the Zambian language exam (languages differ by district); panel (b) shows results for mathematics. The horizontal axis represents years relative to program roll-out, where period 0 indicates the first year a district or zone received the program. The vertical axis shows the treatment effect in the number of students per test center. Blue shading indicates cohorts expected to have been in grades 1 to 3 when the program rolled out; as the program aims to support learners as they move from grade 3 to 5, absent grade retention, these cohorts can be expected to have been fully exposed to the program. Points represent coefficient estimates for each event-time period, with vertical bars indicating 95-percent confidence intervals based on standard errors clustered at the district level (116 clusters). We also report the pooled effects across the four cohorts expected to have been fully exposed to the program (and provide the 95-percent confidence interval in brackets). Each observation in our event-study sample represents a center-by-sex-by-year count (i.e., separate counts for male and female students at each center-year). The sample includes all public grade-7 national assessment examination centers from 2014 to 2025 (6,618 unique centers, with 4,439,847 language scores and 4,461,130 mathematics scores). Program rollout occurred at the district level, except for control zones from the randomized controlled trial (some of which received the intervention in 2025). The reference category consists of period -1 (one year before treatment), as well as all never-treated observations.

Figure A7: *Event-study estimates of effects on grade-7 exam scores (excluding pilot locations)*



(a) *Language (Zambian)*



(b) *Mathematics*

Note: This figure reports event-study estimates of the program's intent-to-treat effect on grade-7 primary school leaving exam scores using Gardner's (2022) two-stage difference-in-differences estimator and exploiting the program's staggered roll-out (see Figure 1), excluding examination centers located in pilot districts and zones. Panel (a) shows results for scores in the Zambian language exam (languages differ by district); panel (b) shows results for exam scores in mathematics. The horizontal axis represents years relative to program roll-out, where period 0 indicates the first year a district or zone received the program. The vertical axis shows the treatment effect in student-level standard deviations. Blue shading indicates cohorts expected to have been in grades 1 to 3 when the program rolled out; as the program aims to support learners as they move from grade 3 to 5, absent grade retention, these cohorts can be expected to have been fully exposed to the program. Points represent coefficient estimates for each event-time period, with vertical bars indicating 95-percent confidence intervals based on standard errors clustered at the district level. We also report the pooled effects across the four cohorts expected to have been fully exposed to the program (and provide the 95-percent confidence interval in brackets). Test scores are standardized using the pooled student-level standard deviation calculated within each year across all examination centers nationally, accounting for both within- and between-center variation. Each observation represents a center-year-level mean, weighted by the number of students tested. The sample includes all public grade-7 national assessment examination centers from 2014 to 2025, excluding those in pilot locations. Program rollout occurred at the district level, except for control zones from the randomized controlled trial (some of which received the intervention in 2025). The reference category consists of period -1 (one year before treatment), as well as all never-treated observations.

Appendix B: Theory of Change

B.1 The Teaching at the Right Level program in Zambia's public schools

The Teaching at the Right Level intervention seeks to address four separate but related challenges common to low and middle-income countries: (1) despite their enrollment in school, many students do not acquire foundational mathematics and literacy skills during their early grades, (2) learning levels within classrooms are very heterogeneous, (3) curricular expectations are very high, and (4) teachers are largely unaware of their students' low skill levels. We verified that these are indeed pressing needs among study participants by measuring related indicators at baseline (for a description of these measures, see section 3.3).

The inputs offered by the intervention to address these challenges are teacher trainings and guidebooks on how to implement the program, following a government order for teachers to do so (for a description of the intervention, see section 2). We check that these inputs are being delivered by accessing records on training attendance. Other inputs into the education system are held constant; the program is implemented with the existing resources, with no additional staff or pay allocated to program schools.

The expected outputs are that teachers diagnose students' learning levels, group them according to their ability, and hold one additional *Catch Up* class per day. We tracked students' school attendance and their exposure to these classes, as also mentioned in section 3.3.

The expected outcomes are that students improve the foundational mathematics and literacy skills targeted by the program. We also expect students to improve their attitudes towards school, literacy, and mathematics, and to improve their after-school study habits. We checked whether this is the case by measuring these outcomes at endline (for details on these measures, see section 3.3).

Lastly, the expected impact is that children improve their foundational mathematics and literacy skills. For literacy, there is no distinction between these skills and those targeted by the program. For mathematics, in turn, the program explicitly targets a subset of foundational skills (e.g., emphasizing basic arithmetic over foundational spatial skills). We also expect secondary impacts on students' creativity, socio-emotional skills, and working memory. We measured these outcomes at endline.

B.2 Supporting TaRL teachers with continuous professional development

This additional program component responds to the challenge of how to change instructional behavior among teachers in public primary schools, and especially how to successfully implement the Teaching at the Right Level intervention at scale. Previous attempts to implement the program with public school teachers during the regular school year either did not lead to significant impacts on student learning or yielded mixed results (for details, see section 1).

The inputs for the additional program component consist of establishing and supporting communities of practice for teachers, additional guidance documents and videos for teachers, and mastery challenges that prompt teachers to collaborate with each other; in addition, the Ministry recognizes participating teachers through formal letters of commendation (for a description of the added program component, see section 2). We have access to a record of all messages sent to teachers and thus track whether the inputs are delivered as planned.

The expected outputs are that teachers participate in the mastery challenges and engage in the communities of practice. We track teachers' enrollment and their participation with complete access to the CPD intervention's backend data. That is, we have data on all submissions of mastery challenges, teachers' enrollment status in WhatsApp groups, and a record of all the messages sent in the communities of practice.

The expected outcomes are that teachers increase their team-based problem-solving, engage in discussions with their colleagues, receive verbal encouragement, and have the opportunity to acquire additional instructional skills by witnessing practical demonstrations. As a result, we expect that teachers improve their implementation of the *Catch Up* program. We checked whether this is the case by measuring related outcomes at endline (for details on these measures, see section 3.3).

Lastly, the expected impact is that children taught by teachers assigned to the added CPD component improve their foundational mathematics and literacy skills more strongly compared to their peers who are taught by teachers in *Catch Up* schools that do not receive the additional program component.

Appendix C: How did classrooms differ?

C.1 Insights from classroom observations

To provide additional insights on how the remedial classes differ from regular instruction, this Appendix reports on classroom observation data collected during the school visits to all 273 schools in July 2024. In control schools, we observed grade-5 classes in regular math and literacy; in the program schools, we observed both regular grade-5 classes and remedial classes.

We relied on the "Teach Primary" classroom observation tool (Molina et al., 2020), a structured instrument that assesses teaching practices across three key areas: classroom culture, instruction, and practices expected to promote the development of socio-emotional skills.[26] Ratings for each subdimension are on a three-point scale; we report on average ratings (by domain and overall); we also use Item Response Theory (IRT) and estimate standardized scores with a graded response model. If raters provided multiple ratings during a class, we aggregate these ratings to the class-level mean.

We provide this information as additional descriptive and comparative background only and advise caution. First, teachers could either refuse to be observed or not hold a class on the day of a given school visit; therefore, we lack classroom observation data from 40 schools. Second, due to concerns of statistical power and multiple hypothesis testing, we did not pre-register the analyses presented here. Third, out of 631 observations conducted, we removed 63 because surveyors indicated that the observed class appeared "staged" (e.g., it did not occur as per the timetable, or teachers asked for extra time to prepare the class).

Table C1 investigates systematic differences between remedial classes in program schools and regular classes in control schools, as well as between regular classes in program schools and control schools. Our analytical approach closely follows the strategy described in the main text (see section 3.4). Columns (1) to (4) report raw means; the remaining columns report IRT scores (which we standardize relative to the distribution of classes in the control group).

Panel A shows that instructional quality is overall higher in remedial classrooms (by 0.20 SD; $p = 0.035$). This systematic difference is of similar size for the subdimension of classroom culture (by 0.19 SD; $p = 0.063$), which captures whether teachers create a welcoming learning environment and set positive behavioral expectations. We do not observe systematic differences for the subdimension of instructional practices, which includes indicators related to lesson facilitation, checks for understanding, feedback provided to students, and teaching that requires critical thinking. The difference between

[26] All raters were trained and certified as per the World Bank's training requirements for the instrument.

remedial classes in program schools and instruction in control schools was particularly pronounced in terms of instructional practices that are expected to increase students' socio-emotional skills (by 0.55 SD; $p < 0.001$), including practices that give greater autonomy to students, promote perseverance, and foster social and collaborative skills.

Panel B shows that remedial classes are characterized by a stark increase in group work (by 37.8 percentage points over a base of 31.7 percent; $p < 0.001$). The program itself does not provide additional materials to teachers yet asks them to create such materials themselves; in line with this expectation, we also observe a strong increase in teachers' use of teaching and learning materials (by 40.4 percentage points over a base of 38.1 percent; $p < 0.001$). Consistent with the program's mandate to teach all students, not just a subset, we do not find a notable difference in class size. We also do not find a meaningful difference in whether teachers were "on task" (i.e., engaged in teaching activities rather than other, non-teaching activities).

Finally, we observe some, yet limited, evidence of spillovers of teaching practices into regular classes in program schools. While noisy, point estimates for the measures of teaching quality are positive; the measure capturing instructional practices expected to promote socio-emotional skills is most pronounced (0.19 SD; $p = 0.105$).

Table C1: *Comparison of remedial and regular classes*

	Control (regular class)	Differences (mean scores)		Differences (IRT scores)	
	Mean	Remedial (T) vs. regular (C)	Regular (T) vs. regular (C)	Remedial (T) vs. regular (C)	Regular (T) vs. regular (C)
	(1)	(2)	(3)	(4)	(5)
Panel A: Teaching quality					
Overall	2.092	0.079	0.033	0.195	0.093
	[0.358]	(0.033)	(0.040)	(0.093)	(0.111)
Classroom culture	2.293	0.070	0.018	0.186	0.039
	[0.385]	(0.038)	(0.045)	(0.100)	(0.115)
Instruction	2.173	-0.012	0.011	-0.011	0.041
	[0.442]	(0.040)	(0.046)	(0.093)	(0.105)
Socioemotional	1.822	0.211	0.075	0.554	0.189
	[0.426]	(0.042)	(0.048)	(0.104)	(0.117)
Panel B: Other observable characteristics					
Group work	0.317	0.378	0.054		
	[0.467]	(0.049)	(0.049)		
Teaching and learning materials used	0.381	0.404	0.058		
	[0.487]	(0.041)	(0.049)		
Class size	30.007	-2.711	2.730		
	[18.323]	(2.114)	(2.115)		
Teacher on task	0.938	-0.002	0.017		
	[0.145]	(0.014)	(0.015)		

Notes. This table investigates systematic differences in classrooms across the program schools and control schools. Column (1) reports means for regular classes in control schools. Columns (2) and (4) compare remedial classes in treatment schools to regular classes in control schools; Columns (3) and (5) compare regular classes in treatment schools to regular classes in control schools. Panel A relies on the "Teach Primary" classroom observation tool (a structured instrument that measures teaching practices across three key areas: classroom culture, instruction, and practices expected to promote socioemotional skills). Panel B provides additional observable characteristics. The sample consists of 568 classroom observations conducted in July 2024. Estimates come from regressions of outcomes on the treatment indicator, with controls for randomization strata and baseline characteristics chosen with post-double Lasso selection. Columns (1) to (4) report raw means; the remaining columns report IRT scores from a graded response model (standardized relative to the distribution of classes in the control group). Standard deviations are shown in square brackets; standard errors, in parentheses, are clustered at the zone level.

C.2 Corporal punishment

While not included in our study's pre-analysis plan, our one-on-one child surveys at endline included questions about the prevalence of corporal punishment in schools. Specifically, surveyors referred to any literacy and math classes from the current term of

the school year, including *Catch Up* classes in the treatment schools, and administered the following question: "[H]ave any of your teachers hit, pinched, or slapped a child? This could have happened to you or to any other child". If the child answered affirmatively, we also asked whether the teacher had punished the interviewed child, another child, or both.

Although Zambian law prohibits corporal punishment of children (as per the Children's Code Act No. 12 of 2022), about one in three children in the control group (31.1 percent) reported that one of their language or math teachers had hit, pinched, or slapped a child in the current term of the school year; about one in four children (23.0 percent) reported that they had experienced such punishment themselves.

Table C2 shows that the program reduced the prevalence of corporal punishment in language and math classes by 3.5 percentage points (q = 0.033). In addition, it reduced the proportion of children who reported being personally hit, pinched, or slapped by 2.7 percentage points (q = 0.066).

Finally, among students in program schools who reported experiences of corporal punishment, we asked whether it occurred in their regular class(es), their *Catch Up* class(es), or both. Corporal punishment was less common in remedial classes: 64.7 percent of children reported it had occurred exclusively in regular classes, 12.4 percent reported it had occurred only in remedial classes, and 22.8 percent reported it had occurred in both types of classes.

References

Molina, E., Fatima, S.F., Ho, A.D., Melo, C., Wilichowski, T.M., Pushparatnam, A., 2020. Measuring the quality of teaching practices in primary schools: Assessing the validity of the Teach observation tool in Punjab, Pakistan. Teaching and Teacher Education 96, 103171. doi:10.1016/j.tate.2020.103171.

Table C2: *Intent-to-treat effects on the prevalence of corporal punishment*

	Pooled ITT effect	ITT effects by program variant		
		Without CPD	With CPD	Effect of CPD
	(1)	(2)	(3)	(4)
Anyone	-0.035	-0.042	-0.029	0.013
	(0.015)	(0.017)	(0.018)	(0.015)
	[0.033]			
Self	-0.027	-0.029	-0.025	0.003
	(0.014)	(0.015)	(0.015)	(0.013)
	[0.066]			

Notes. This table shows the impact of the intervention on the prevalence of corporal punishment—i.e., whether a literacy or math teacher had hit, pinched, or slapped a child during the current term of the school year. Estimates come from regressions of outcomes on a treatment indicator with controls for randomization strata and baseline characteristics chosen with post-double Lasso selection. Outcomes reflect proportions. Standard errors, in parentheses, are clustered at the zone level. Simes FDR q-values are reported in brackets. Because corporal punishment outcomes were not pre-registered, these q-values are computed by appending the two outcomes to the pre-specified order of mechanism tests reported in Table 3. "Anyone" includes any child, including the respondent; "Self" refers to the respondent. "Without CPD" and "With CPD" refer to the two treatment arms without and with the continuous professional development component for teachers; "Effect of CPD" refers to their difference. As pre-registered, since there is no meaningful difference in effects, we focus the paper (and its adjustment of hypothesis tests) on effects that pool the two treatment arms. In the control group, 31.1 percent of children reported that one of their language or math teachers had hit, pinched, or slapped a child in the current term of the school year, whereas 23.0 percent reported that they had experienced this punishment themselves.

Appendix D: Measurement

D.1 Measurement in the randomized controlled trial

D.1.1 Primary outcomes

Foundational skills in literacy and mathematics. We measured students' foundational skills in literacy and mathematics with one-on-one assessments. The instruments consisted of two components: (1) A standard ASER test that covers select math domains (number recognition and procedural arithmetic) and select literacy domains (letter recognition and reading), and (2) additional test questions that focus on the remaining domains of foundational mathematics and literacy not measured by the ASER test. Both test components were adaptive; they only tested more advanced skills if students had the respective prerequisites (e.g., students who could not read letters were not asked to attempt a reading comprehension task).

To construct the assessments, we used a blueprint with a clear mapping of test questions to content and cognitive domains. They follow common definitions of "foundational skills," which are recognized internationally and in Zambia.[27] In mathematics, the assessments capture four content domains: basic arithmetic, data display, geometric shapes and measurement, and numbers (number recognition and number sense). They also capture two cognitive domains that cut across the content domains: applied or higher-order thinking skills, and procedural or lower-order thinking skills. In literacy, the assessments capture six domains: phonemic awareness, vocabulary, listening comprehension (the three of which may be jointly considered as "pre-literacy skills"), phonics, writing, and reading (including basic reading, reading fluency, and reading comprehension).[28]

The assessments recorded students responses to all test questions (or test "items") for both test components. To aggregate these responses, we use a hybrid item response theory (IRT) model, with a two-parameter logistic (2PL) model for binary items and a graded response model (GRM) for ordinal items. More specifically, we estimate a two-group model at endline that constrains the item parameters across experimental groups but allows for

[27] These definitions align with Zambia's mathematics syllabus for the early grades and with the national literacy framework. In mathematics, they also align with the UNESCO global proficiency framework. In literacy, the national literacy framework (and our assessments) cover skills that go beyond the UNESCO global proficiency framework for reading (e.g., writing).

[28] The blueprint also maps the test questions to grade-level expectations, following Zambias official curriculum framework. As measures of foundational skills, none of the included items exceed materials beyond grade 3. In 2025, after our study, Zambia's curriculum for public primary schools will change; here, we are referring to the curriculum in place during the study period. We thank expert reviewers at Zambia's Ministry of Education for their evaluation of the assessments and confirmation that they align with Zambia's "Literacy and Numeracy Education Framework."

the latent ability distributions to differ across the treatment and control groups. To link ability estimates across the baseline and endline assessments, when estimating students' baseline ability levels, we constrain the item characteristics of repeat items ("anchor items") to match those estimated with the endline data only.[29] We use Expected A Posteriori (EAP) estimation to predict one overall, continuous score per subject and standardize this score (with a mean of zero for the control group, as per the test score distribution at endline).

Our instruments measured students' literacy and math ability with high levels of precision. At endline, the average conditional reliability of the literacy and mathematics measures is 0.93 and 0.90, respectively (see Appendix Table D1).[30] Precision is not just high overall but for a wide range of student ability, with reliability estimates remaining either close to or above 0.8, even two standard deviations below and above the mean (see Appendix Figure D1).

For completeness, and to facilitate the reproducibility of our results, we provide the full list of test questions administered at endline, their item IDs, domain-wise mapping, and item parameters in Appendix Tables D2 and D3.

D.1.2 Secondary outcomes

Mathematics skills explicitly targeted by the program. In mathematics, we also generate one continuous, standardized score for those skills targeted by the program (i.e., number recognition and procedural arithmetic) as opposed to the remaining content domains captured by the assessments (e.g., geometry). We estimate each student's score using item response theory and a two-parameter logistic model. We standardize scores with respect to the control group at endline. For completeness, we also report on the proportion of students who have mastered discrete levels of ability, as per the ASER tests (we focus on whether students can at least do two-digit subtraction with borrowing). Using endline data, the average conditional reliability of the measure of skills explicitly targeted by the program is 0.80.

[29] We free up the item parameters of those anchor items exhibiting differential item functioning (DIF) across the baseline and endline rounds.

[30] Here and elsewhere, to report on the reliability of our measures, we calculate the average conditional reliability as

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\mathrm{Var}(\hat{\theta})}{\mathrm{Var}(\hat{\theta})+\mathrm{SE}(\hat{\theta}_i)^2}\right],$$

where N is the sample size, $\mathrm{SE}(\hat{\theta}_i)$ is the predicted standard error of the ability estimate for individual i, and $\mathrm{Var}(\hat{\theta})$ is the variance of the predicted ability scores.

Students' creativity. We measure students' creativity using an adaptation of the Torrance test of creative thinking (Alabbasi et al., 2022) at endline. The creativity test's first item asked students to name all the different things they may do with a pencil. Enumerators counted the number of unusual ideas students listed.[31] Its second item asked students to imagine that they could walk on air or fly and list any problems this might create. Enumerators counted the number of problems students listed. Items three to five asked students to complete three incomplete drawings with a pencil. Enumerators counted the number of unique elements in a child's drawing and the number of drawings students completed. We estimate each student's score with a single-component Poisson principal component analysis (PCA). We standardize scores with respect to the control group at endline. Using endline data, the average conditional reliability of the creativity measure is 0.78.

Students' socio-emotional skills. To measure students' socio-emotional skills, we combine and adapt the Perceiving AI-Generated Emotions assessment (PAGE; Weidmann and Xu, 2025) with the subdomains related to socio-emotional skills from the International Development and Early Learning Assessment (IDELA; Pisani et al., 2018). The PAGE asks students to identify another child's emotion when shown an image of that child. The assessment includes 12 items, all of which consist of AI-generated images of Zambian children. The IDELA includes four questions related to emotional awareness and emotional regulation; it also includes three questions related to empathy and perspective-taking. We estimate each child's score using item response theory and a two-parameter logistic model. We standardize scores with respect to the control group at endline. Using endline data, the average conditional reliability of the measure of socio-emotional skill is 0.58.

Students' working memory. To measure students' working memory, we developed and administered a tablet-based, (forward) visuospatial recall test at endline. Students were presented with a screen where, in a random sequence, up to nine gray boxes lit up in color. They were then asked to remember what they had seen and tab on the boxes in the order in which they had lit up. Starting with two boxes only, the sequence of boxes extended to nine boxes (in increments of one).[32] The tablet recorded, for each box, whether the student picked the correct position in the given sequence. We estimate each child's score using item response theory and a two-parameter logistic model. We standardize scores with respect

[31] We trained enumerators on a protocol that distinguishes "usual" use cases (e.g., writing a letter) from unusual ideas (e.g., using the pencil as a magic wand).

[32] The task is similar to a forward digit-span task. We decided against a digit-span test and favored the given visio-spatial task out of concerns that a digit-span task could conflate students' working memory with their single-digit number recognition skills.

to the control group at endline. Using endline data, the average conditional reliability of the measure of working memory is 0.87.

D.1.3 Potential mechanisms

Student attitudes towards school, literacy, and mathematics. To measure students' attitudes towards school, literacy, and mathematics at endline, we adapted related questions from the Progress in International Reading Literacy Study (PIRLS). Using a four-point answer format ranging from "agree a lot" to "disagree a lot," students noted their level of agreement with eight questions about reading and literacy, nine questions about mathematics, and five questions about their school. Questions were positive (e.g. "I enjoy learning mathematics") or negative (e.g., "mathematics is boring"); for negative statements, we recode all ratings such that positive values reflect a desirable outcome (e.g., "not boring"). We estimate each student's score using item response theory and a graded response model. We standardize scores with respect to the control group at endline. Using endline data, the average conditional reliability of the measure of student attitudes is 0.85.

Student's study habits at home. To measure at endline whether students studied literacy or mathematics outside of school, we inquired about whether, for each of the two subjects, they had studied at home the previous day or done any homework. We focus on a binary variable indicating whether, for at least one of the subjects, the student responded in the affirmative. At endline, 47.9 percent of the students in the control group stated that they had studied or done homework in at least one of the two subjects.

Teacher collaboration and feedback. To measure the extent to which teachers collaborate with their colleagues and receive feedback, we administered one-on-one interviews with them at baseline and during process monitoring.[33] To construct a summary measure across these indicators, we rely on item response theory and a partial credit model. We standardize scores with respect to the control group at follow-up. Using endline data, the average conditional reliability of the summary measure of teachers' participation in and exposure to continuous professional development is 0.77.

[33]Earlier, we referred to this measure as "participation in and exposure to continuous professional development activities". To avoid confusion with our measure of program uptake in the CPD group, we renamed this indicator.

D.1.4 Take-up and implementation

Program take-up and implementation quality. To measure children's exposure to the *Catch Up* program, during the endline child interviews, we recorded administrative data on whether each student had attended any *Catch Up* class the previous school day (in either literacy or mathematics).[34] To account for absenteeism and dropouts, we recorded administrative data on whether each student was present in school the previous school day. Our *main* measure of program exposure multiplies these two binary indicator variables. We will characterize implementation quality with additional *secondary* descriptive indicators (including, for example, whether a child's learning level as diagnosed by her teacher matches the learning levels diagnosed with our independent endline assessments). To measure teachers' participation in the continuous professional development (CPD) program, we report whether at least one teacher in a school participated in the program's mastery challenges, and whether each teacher had ever participated in the CPD program over WhatsApp.[35]

D.1.5 Other measures

Teachers' ability to accurately diagnose their students' learning levels (O1). At baseline and follow-up, we asked teachers to estimate the proportion of their students who can solve a specific question from the study's student assessments (solving a subtraction problem, with carrying, and reading a paragraph). We then compared the teacher's estimate with the observed proportion in her school, as per the student assessments. At baseline, teachers overestimated their students' ability to solve a subtraction question by 47 percentage points and their ability to read a paragraph by 27 percentage points.

Teachers' perceived locus of control. To measure teachers' perceived ability to affect student learning (i.e., their "locus of control"), we administered one-on-one interviews with them at baseline and during process monitoring. Each teacher stated whether they agreed or disagreed with five statements. For example, one statement is "There is little I can do to help a student's learning." For each teacher, we construct a summary measure across these indicators, using item response theory and a two-parameter logistic model. We standardize scores with respect to the control group at follow-up. Using baseline data, the average conditional reliability of the summary measure of teachers' locus of control is 0.51, suggesting that there is substantial noise in the measure.

[34] Initially, we intended to collect direct child reports on whether they had attended a Catch Up class. However, due to a coding error, we did not include this question in the child interviews.

[35] Since teachers are asked to collaborate and participate in the program together, we may see just one teacher per school submit a solution to the CPD program's mastery challenges.

D.2 Measurement in the event study

Our event-study results rely on restricted data from Zambia's Examinations Council (ECZ) covering the universe of grade-7 primary school leaving examinations administered in public schools from 2014 to 2025. This dataset contains center-level statistics (means, standard deviations, and student counts), disaggregated by subject and sex, from all public examination centers nationwide. The data includes location identifiers that allow us to link each center to the timing of program rollout.

To better understand the content domains, cognitive domains, and curricular grade levels captured by these exams, we conducted workshops in collaboration with local content-matter experts and ECZ, uniquely mapping all test questions administered across the twelve years. Our mapping of content domains follows the Global Proficiency Framework (GPF) domains: Algebra, Geometry, Measurement, Number & operations, and Statistics & probability; cognitive domains follow the TIMSS trichotomy of Knowing, Applying, and Reasoning. Curricular grade levels refer to the Zambian national mathematics curriculum. Across the years, we calculated the percentage of test questions mapping to each domain and grade level; we also calculated the percentage of test questions mapping to the procedural number and operations skills covered in the first three grades of the curriculum (capturing the proximal dimension of foundational skills specifically targeted by the intervention).

Figure D2 summarizes the results, showing that the grade-7 math exams go far beyond the scope of foundational skills targeted by the intervention. While about two-thirds of questions relate to the number and operations domain (64.9 percent), only about one-third of test questions (37.1 percent) relate to procedural skills, and about 9 out of 10 test questions capture curricular material from above grade 3 (88.3 percent; foundational skills are usually covered in the first three grades of public-school curricula). Narrowing in on the number and operations domain, only about one quarter of all test questions relate to procedural number and operations skills (25.6 percent), less than 10 percent relate to foundational number and operations skills up to grade 3 (9.3 percent), and only 3.6 percent of test questions cover the targeted subdomain of procedural, foundational number and operations skills. Effects on this exam would therefore reflect either broad transfer to higher-order and higher-grade content, spillovers to non-targeted domains, or both.

Table D1: *Average reliability of outcome measures*

	Average conditional reliability (1)
Literacy	0.932
Math	0.898
Targeted math skills	0.795
Working memory	0.871
Creativity	0.782
Socio-emotional skills	0.579
Attitudes	0.853
Teacher collaboration and feedback	0.767

Notes. This table reports the average conditional reliability of the outcome measures shown in Table 2 and Table 3. Average conditional reliability is calculated as

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\mathrm{Var}(\hat{\theta})}{\mathrm{Var}(\hat{\theta})+\mathrm{SE}(\hat{\theta}_i)^2}\right],$$

where N is the sample size, $\mathrm{SE}(\hat{\theta}_i)$ is the predicted standard error of the ability estimate for individual i, and $\mathrm{Var}(\hat{\theta})$ is the variance of the predicted ability scores. This measure summarizes the average precision of measurement for the study sample at endline.

Table D2: *Item characteristics: Literacy*

Item ID	Domain	Type	Anchor	Discrimination	Difficulty
s10	Vocabulary	2PL	Yes	0.556	-5.903
se21-se26	Listening comprehension	2PL	No	0.550	-4.817
se12	Listening comprehension	2PL	No	0.719	-3.457
se11	Listening comprehension	2PL	No	0.926	-2.866
se7	Phonics	2PL	No	0.746	-2.749
s6	Phonics	2PL	No	1.051	-2.384
se13	Listening comprehension	2PL	No	0.704	-2.369
s7	Phonics	2PL	Yes	1.085	-2.339
s8	Phonics	2PL	Yes	1.075	-2.097
se3	Phonemic awareness	2PL	No	0.967	-1.725
s1	Phonemic awareness	2PL	Yes	1.162	-1.581
s5	Phonics	2PL	Yes	0.803	-1.579
s16	Writing	2PL	Yes	1.638	-1.469
s15	Writing	2PL	Yes	1.681	-1.402
s2	Phonemic awareness	2PL	Yes	1.050	-1.033
se1	Phonemic awareness	2PL	No	1.039	-0.861
s9	Vocabulary	2PL	Yes	0.669	-0.819
se2	Phonemic awareness	2PL	No	0.982	-0.737
se19	Reading	2PL	No	1.903	-0.702
ASER: Beginner | Letter	Reading	GRM	Yes	3.194	-0.700
se18	Reading	2PL	No	1.179	-0.554
s14	Writing	2PL	Yes	2.574	-0.509
ASER: Letter | Word	Reading	GRM	Yes	3.194	-0.412
se17: Not able | String of words	Reading	GRM	No	3.467	-0.399
se16_8	Writing	2PL	No	2.463	-0.398
se28	Reading	2PL	No	1.665	-0.321
se27	Reading	2PL	No	1.684	-0.185
se5	Phonemic awareness	2PL	No	1.198	-0.092
se16_1	Writing	2PL	No	3.873	-0.091
s3	Phonemic awareness	2PL	Yes	0.883	-0.033
se20	Reading	2PL	No	2.089	0.033
se16_5	Writing	2PL	No	3.672	0.086
se16_7	Writing	2PL	No	3.689	0.095
s4	Phonemic awareness	2PL	No	1.199	0.098
se16_3	Writing	2PL	No	3.522	0.202
se16_2	Writing	2PL	No	3.256	0.244
se16_4	Writing	2PL	No	3.313	0.398
se16_9	Writing	2PL	No	1.510	0.403
se16_6	Writing	2PL	No	2.744	0.512
ASER: Word | Paragraph	Reading	GRM	Yes	3.194	0.583
se17: String of words | Fluent	Reading	GRM	No	3.467	0.709
ASER: Paragraph | Story	Reading	GRM	Yes	3.194	0.753
se16_10	Writing	2PL	No	2.021	1.340
se16_12	Writing	2PL	No	2.942	1.462
se16_11	Writing	2PL	No	2.771	1.503

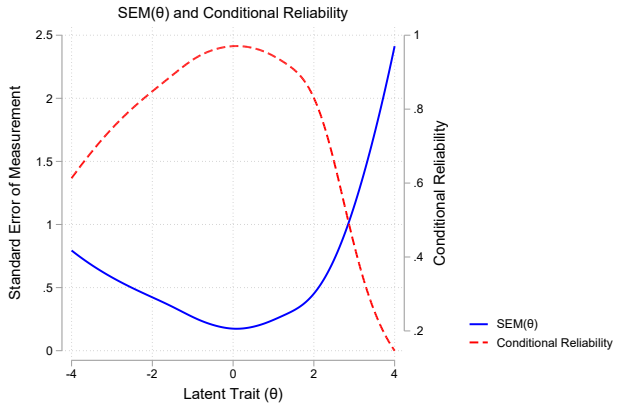
Notes. This table provides item characteristics from a hybrid item response theory (IRT) model, using a two-parameter logistic (2PL) model for binary items and a graded response model (GRM) for ordinal items. Item names are shown in the left-most columns and refer to study-internal question IDs. For ordinal items, a vertical bar in the item name indicates the estimated threshold between response categories. "se21-se26" refers to an indicator capturing whether a student was able to follow a set of six verbal instructions. Rows are sorted in ascending order by difficulty/threshold. The table indicates each item's literacy domain. All questions assess foundational skills expected to be covered in grades 1-3, as per Zambia's curricular expectations. "Anchor" refers to items administered at both baseline and endline; this table shows endline items only. For binary items, discrimination and difficulty parameters refer to the 2PL item parameters; for ordinal items, they refer to the discrimination and threshold parameters, respectively.

Table D3: *Item characteristics: Mathematics*

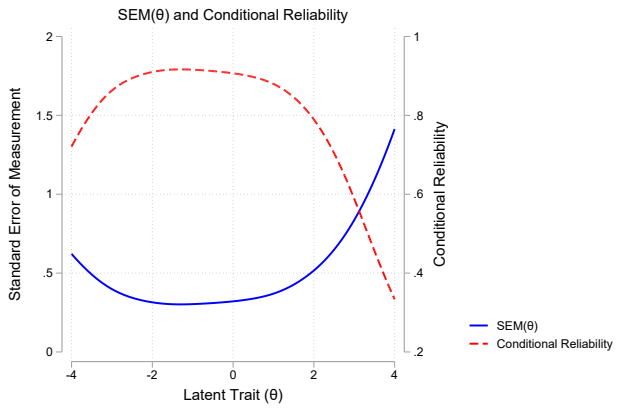
Item ID	Targeted domain	Content domain	Cognitive domain	Type	Anchor	Discrimination	Difficulty
q103	No	Geometric shapes and measures	Applied	2PL	No	0.117	-21.684
q7002	No	Geometric shapes and measures	Applied	2PL	No	0.222	-8.645
ASER: Beginner | One digit	Yes	Numbers	Procedural	GRM	Yes	1.525	-4.092
q101_2	No	Numbers	Procedural	2PL	No	1.520	-2.892
q101_1	No	Numbers	Procedural	2PL	No	2.053	-2.821
q41186	No	Data display	Procedural	2PL	Yes	0.844	-2.818
q101_3	No	Numbers	Procedural	2PL	No	1.797	-2.516
q101_5	No	Numbers	Procedural	2PL	No	1.633	-2.122
ASER: One digit | Two digit	Yes	Numbers	Procedural	GRM	Yes	1.525	-2.052
q100	No	Arithmetic	Applied	2PL	No	1.260	-2.051
q101_4	No	Numbers	Procedural	2PL	No	1.748	-2.023
q7026	No	Numbers	Applied	2PL	Yes	1.089	-1.631
q101_7	No	Numbers	Procedural	2PL	No	1.451	-1.578
q101_6	No	Numbers	Procedural	2PL	No	1.970	-1.547
q101_10	No	Numbers	Procedural	2PL	No	1.539	-1.425
q107	No	Arithmetic	Applied	2PL	No	1.741	-1.396
q101_8	No	Numbers	Procedural	2PL	No	1.342	-1.391
q22	No	Arithmetic	Applied	2PL	Yes	1.278	-1.366
q7008	No	Geometric shapes and measures	Procedural	2PL	No	1.296	-1.289
q1126	No	Geometric shapes and measures	Procedural	2PL	Yes	0.672	-1.265
q101_9	No	Numbers	Procedural	2PL	No	1.452	-1.214
q300	No	Arithmetic	Applied	2PL	No	1.203	-1.095
q105	No	Data display	Procedural	2PL	No	0.898	-0.981
q119	Yes	Arithmetic	Procedural	2PL	No	1.793	-0.729
q800	Yes	Arithmetic	Procedural	2PL	No	1.522	-0.375
q8	Yes	Numbers	Procedural	2PL	Yes	1.748	-0.341
q104	No	Geometric shapes and measures	Applied	2PL	No	0.594	-0.079
ASER: No addition | Addition	Yes	Arithmetic	Procedural	GRM	Yes	1.703	-0.055
q7027	No	Numbers	Procedural	2PL	Yes	1.744	0.061
ASER: Two digit | Three digit	Yes	Numbers	Procedural	GRM	Yes	1.525	0.082
q7006	No	Data display	Applied	2PL	Yes	1.004	0.240
ASER: Addition | Subtraction	Yes	Arithmetic	Procedural	GRM	Yes	1.703	0.405
q7021	No	Arithmetic	Applied	2PL	Yes	1.446	0.406
ASER: Three digit | Four digit	Yes	Numbers	Procedural	GRM	Yes	1.525	0.593
q41	No	Numbers	Applied	2PL	Yes	1.482	0.712
q3036	No	Geometric shapes and measures	Procedural	2PL	Yes	0.831	0.726
q900	No	Data display	Applied	2PL	No	0.958	0.765
q7022	No	Arithmetic	Applied	2PL	No	1.551	0.799
q10	Yes	Arithmetic	Procedural	2PL	No	1.917	0.928
ASER: Subtraction | Multiplication	Yes	Arithmetic	Procedural	GRM	Yes	1.703	1.079
q106	No	Data display	Procedural	2PL	No	1.407	1.212
q108	No	Numbers	Applied	2PL	No	0.837	1.462
q7030	No	Data display	Applied	2PL	Yes	0.712	1.876
ASER: Multiplication | Division	Yes	Arithmetic	Procedural	GRM	Yes	1.703	1.909

Notes. This table provides item characteristics from a hybrid item response theory (IRT) model, using a two-parameter logistic (2PL) model for binary items and a graded response model (GRM) for ordinal items. Item names are shown in the left-most columns and refer to study-internal question IDs. For ordinal items, a vertical bar in the item name indicates the estimated threshold between response categories. Rows are sorted in ascending order by difficulty/threshold. The table indicates whether items pertain to the pre-registered subdomain of targeted skills (number recognition and procedural arithmetic) versus other domains, as well as their mapping to content and cognitive domains. All questions assess foundational skills expected to be covered in grades 1-3, as per Zambia's curricular expectations. "Anchor" refers to items administered at both baseline and endline; this table shows endline items only. For binary items, discrimination and difficulty parameters refer to the 2PL item parameters; for ordinal items, they refer to the discrimination and threshold parameters, respectively.

Figure D1: *Test reliability across varying ability levels*



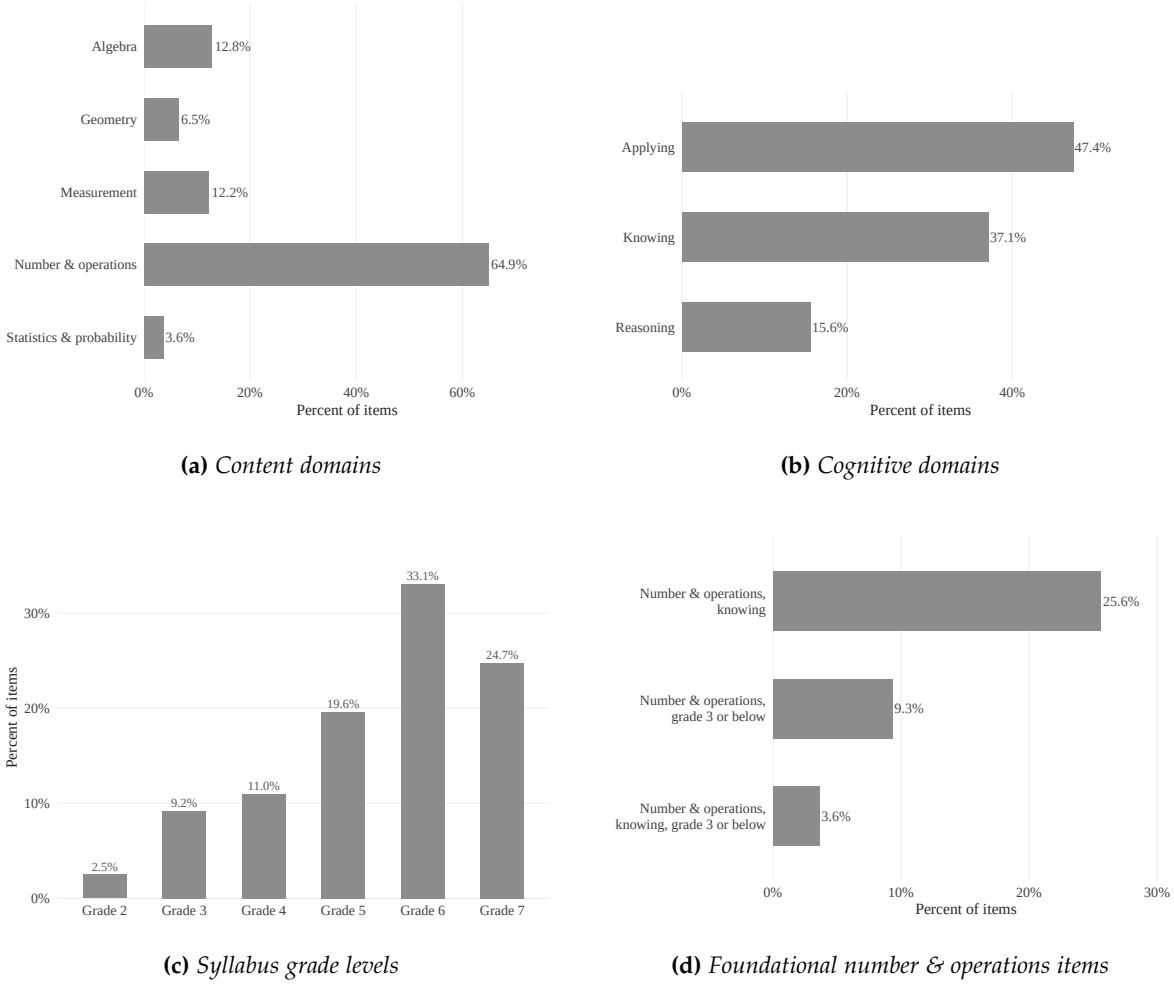
(a) *Literacy*



(b) *Mathematics*

Note: This figure reports the reliability of ability estimates across the latent ability scale θ in both subjects at endline. Solid lines report the standard error of measurement implied by the estimated item response theory (IRT) models, while dashed lines report the corresponding reliability conditional on θ.

Figure D2: *Item classification of ECZ grade-7 math exams (2014-2025)*



(a) *Content domains*

(b) *Cognitive domains*

(c) *Syllabus grade levels*

(d) *Foundational number & operations items*

Note: This figure summarizes the classification of items on the Examination Council of Zambia (ECZ) grade-7 mathematics examination, averaged across the twelve administrations from 2014-2025. Each year, tests include 60 questions. We mapped all of these items in collaboration with local content-matter experts and ECZ staff. Content domains follow the Global Proficiency Framework (GPF) domains: Algebra, Geometry, Measurement, Number & operations, and Statistics & probability. Cognitive domains follow the TIMSS trichotomy of Knowing, Applying, and Reasoning. Curricular grade levels refer to the Zambian national mathematics curriculum. Panel (d) reports three overlapping subsets of Number & operations items: those classified as "Knowing" (i.e., procedural items), those targeting grade 3 or below (i.e., foundational items), and their intersection. The latter relates to the proximal dimension of procedural, foundational number recognition and arithmetic skills specifically targeted by the intervention.

References

Alabbasi, A.M.A., Paek, S.H., Kim, D., Cramond, B., 2022. What do educators need to know about the Torrance Tests of Creative Thinking: A comprehensive review. Frontiers in Psychology 13. doi:10.3389/fpsyg.2022.1000385.

Pisani, L., Borisova, I., Dowd, A.J., 2018. Developing and validating the International Development and Early Learning Assessment (IDELA). International Journal of Educational Research 91, 1–15. doi:10.1016/j.ijer.2018.06.007.

Weidmann, B., Xu, Y., 2025. Measuring Emotion Perception Ability Using AI-Generated Stimuli: Development and Validation of the PAGE Test. Journal of Intelligence 13. doi:10.3390/jintelligence13090116.

Appendix E: Program costs

This Appendix provides a summary of the program costs for implementing the Teaching at the Right Level (TaRL) program in Zambia, which is locally referred to as the *Catch Up* program (CU). The J-PAL program costing template was adapted for this exercise (Glandon et al., 2023). We focus on the time period and locations of program implementation of the randomized trial.

The purpose of collecting and categorizing these costs is to identify all resources required to implement the CU and CU-with-added-CPD models, excluding the costs associated with evaluating the program's impact (research costs). This information may assist NGOs, governments, and other policymakers in estimating the cost of replicating or scaling up similar programs.

The program whose costs are captured was implemented in 1,261 CU schools, out of which 1,170 received the standard CU program and 91 received CU with the added continuous professional development component (CPD). These schools were distributed across Western Province, Central Province, and Itezhi Tezhi District in Southern Province. Specifically, CU was offered to 542 schools in Central Province; 680 schools in Western Province, and 39 schools in the Itezhi-Tezhi District. To establish the number of students enrolled in grades 3 to 5, we use enrollment numbers from the Zambian Education Management Information System (EMIS), as of 2020.[36]

Below, we provide additional details regarding the categorization of costs covered, cost categories common to other projects yet excluded in our case, and our approach to inflation adjustment. Table E1 summarizes the results.

E.1 Costs included

The cost data covers the following key areas.

Program administration. This includes staff salaries and benefits based on their level of effort within the programs, as well as office operation costs.

To apportion the general costs incurred by VVOB, using Catch Up's enrollment data, we determined that 15.36% of all CU schools in the country (across the 2023 and 2024 school years) are located in the study zones. We then estimated that Catch Up accounts for 80%

[36]Conservatively, we include students in government-run schools only; we exclude additional students in community schools, even though the program supports some of these schools. We exclude community schools because we lack zone-level location data for these schools and because not all community schools are covered in the EMIS.

of VVOB's effort in the country, with the remaining 20% allocated to the NGO's remaining program activities. To calculate a flat percentage for general costs, we multiplied these two figures: 15.36% $\times$ 80% = 12.29%. Accordingly, 12.29% was used to apportion VVOB's administrative staff costs (HR, Finance, leadership) as well as office rent and other office expenses.

For TaRL Africa, since both program staff and administrative staff (HR, Finance, and leadership) support multiple countries operating at different scales, costs were apportioned based on each staff member's estimated level of support to the Zambian program.

User training Costs associated with training users involved in the program (costs of training teachers, mentors, and Ministry of Education staff), including refresher training. These costs include travel, venue, meals, and materials.

Implementation. This primarily includes the development of materials, field visit expenses, and the costs of organizing workshops.

Monitoring costs. Costs incurred due to oversight, monitoring, or tracking of the program recipients, such as schools and teachers, and their progress during the intervention. Monitoring activities are conducted for both types of CU schools, with and without the CPD component. When monitoring staff visit schools implementing CU with the added CPD component, they focus on both CU and CPD; since these visits are already accounted for under CU monitoring, there was no need to allocate additional monitoring costs for CPD.

E.2 Costs not included

Some costs, common to other projects, were not included in this case.

Targeting costs. Expenses related to identifying and raising awareness among potential beneficiaries were not part of this intervention. This was not included because the CU program has a clearly defined set of beneficiaries (schools, teachers, and students).

User costs. Costs directly incurred by end users such as schools, mentors, and government officials. This includes the opportunity costs of delivering CU. Schools may occasionally print or photocopy assessment forms, but this is funded through their

government-provided capitation grants, which are intended for general school operations (and remain unchanged with the program). The cost of such photocopying is minimal and difficult to track.

Averted costs. No notable costs were avoided as a result of the intervention.

E.3 Inflation adjustment

Adjustment for inflation is based on the U.S. Bureau of Economic Analysis GDP deflator and the U.S. Bureau of Labor Statistics inflation calculator. Accordingly, 1 USD in July 2022 is equivalent to 1.04 USD in December 2023 and 1.07 USD in December 2024.

Table E1: *Program costs*

	CU (1)	CU CPD (2)
Program administration	350,089	85,253
User training	1,047,874	13,683
Implementation costs	138,180	48,742
User costs	0	0
Monitoring costs	122,310	0
Total	1,658,453	147,678
Cost per student (EMIS)	9.63	10.34

Notes. This table reports on costs for the two program variants investigated in the two years of the randomized trial (in inflation-adjusted US dollars, as of July 2022, excluding research costs). "Program administration" includes the costs of all full-time staff who worked throughout all phases of the intervention and implementation, and other costs related to program administration. It also includes overhead costs (such as an apportioned cost for office rent). "User training" refers to costs incurred to train participants or beneficiaries (such as teachers and headteachers). "Implementation costs" include the costs of program materials, program staff field travel, and workshops (such as costs for printing materials). Under "user costs", we exclude minor costs related to any costs teachers incurred when re-printing teaching materials and their opportunity cost of participating in program activities. "Monitoring costs" refers to costs incurred for oversight, monitoring, or tracking the program. "EMIS" refers to enrollment counts from Zambia's Education Management Information System (conservatively, counts include students in government-run schools only; they exclude students in community schools, even though the program supports some of these schools).

References

Glandon, D., Fishman, S., Tulloch, C., Bhula, R., Morgan, G., Hirji, S., Brown, L., 2023. The State of Cost-Effectiveness Guidance: Ten Best Resources for CEA in Impact Evaluations. Journal of Development Effectiveness 15, 5–16. doi:10.1080/19439342.2022.2034916.

Appendix F: Amendments to the pre-analysis plan

04 April 2025

On 04 April 2025, we introduced the following changes to the pre-analysis plan. These changes are reactions to an independent review of the plan conducted by the study's advisory committee. At this point, we only had access to follow-up data for the control group.

Selection of control variables. Following Cilliers et al. (2024), our analytical strategy and estimating equations now use the post-double selection (PDS) Lasso (Belloni et al., 2014) yet pre-specify the inclusion of both randomization strata fixed effects and baseline measures of students' literacy and mathematics scores as control variables. We now also specify the control variable input set for the PDS algorithm.

Concerns regarding the study's instrumental variable approach. We added caveats concerning our ability to test hypotheses P1*, P2*, S1*, S2*, S3*, S4*, S5*, S6*, S7*, S8*, M1*, and M2* of the pre-analysis plan. We also stressed their exploratory nature more and provided a clearer definition of program exposure. In doing so, our subsequent review of the control group dataset revealed that, due to a coding error, we do not have access to student self-reports of whether they attended a *Catch Up* class the day prior to their interview. To report on students' exposure to the program, we now rely on the study's digitization of attendance data instead.

14 April 2025

After discussions at the Jacobs Foundation / NYU Steinhardt Education Policy Group Agenda Setting Meeting on learning variability in low- and middle-income countries, we added the following clarification to our section on a machine learning-based exploration of heterogeneous effects. At this point, we only had access to follow-up data for the control group.

We are particularly interested in exploring heterogeneous treatment effects by within-school learning variability as measured by schools' Gini coefficient of learning outcomes and subgroup effects for students whose learning, within a given school's distribution of learning outcomes, lagged farthest behind at baseline.

Changes introduced after accessing the full endline data

After consulting with other researchers, we also introduced the following additional changes. At this point, we already had access to complete follow-up data; however, as can be seen below, these changes are not motivated by accessing that data.

IRT models with ordinal items. We had initially planned to dichotomize each ordinal response level—coding whether a respondent had reached at least that level—and then estimate a two-parameter logistic item response theory (IRT) model on these binary indicators. However, this approach violates the IRT assumption of local independence because the dichotomized indicators from the same item are deterministically related. We therefore switched to a hybrid IRT specification that uses a two-parameter logistic model for binary items and a graded response model for ordinal items. We also considered partial credit models and had pre-specified a partial credit model for the index of teacher collaboration and feedback; however, due to convergence issues, we adopted the graded response model for all ordinal items.

Index of teacher collaboration and feedback. Our pre-analysis plan referred to the index of teacher collaboration and feedback as "continuous professional development (team-based problem-solving among teachers, verbal encouragement and discussions, and teachers participation in practical demonstrations of teaching methods)." To avoid confusion between (a) this measure of a potential mediator and (b) our other measures of teachers' take-up of the CPD intervention (e.g., their participation in mastery challenges), we removed the term "continuous professional development" and now refer to the index as "teacher collaboration and feedback".

References

Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on Treatment Effects after Selection among High-Dimensional Controls. The Review of Economic Studies 81, 608–650. doi:10.1093/restud/rdt044.

Cilliers, J., Elashmawy, N., McKenzie, D., 2024. Using Post-Double Selection Lasso in Field Experiments. Working Paper 10931. The World Bank. Washington, D.C. URL: https://openknowledge.worldbank.org/entities/publication/0cde089d-33ba-4f51-8c03-b25b5114d41a.