Pre-Results Review: Stage 1 Submission

Targeted Instruction at Scale*

Andreas de Barros† Theresa Lubozha‡

January 27, 2025

Abstract

Targeted instruction is a promising approach to tackle the low learning levels that plague many developing countries. However, little is known about how to promote this strategy at scale and maintain its effectiveness in government-run schools that serve students in high-poverty, low-resource settings. This cluster-randomized trial measures the impact of the "Teaching at the Right Level" program on foundational literacy and mathematics skills in Zambia's public primary schools. In the program, teachers group children based on their learning needs and pace and provide tailored remedial instruction during additional classes. The study also investigates the effectiveness of combining the program with a continuous professional development initiative for teachers. The study offers causal evidence on the role of teachers in promoting the effectiveness of an education intervention as it is scaled up by a national government.

Keywords: at-scale experiment; scaling; teaching at the right level; teacher effectiveness.

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

---

*This document follows the "reporting checklist" of the Journal of Development Economics (JDE) pre-results review process (Stage 1). The study was registered with the AEA Trial Registry (RCT 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. We 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 Centre for Promotion of Literacy in Sub-Saharan Africa, Innovations for Poverty Action, and Palm Associates assisted with data collection. We thank Prince Muraguri and Victor Olajide, who provided 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, as well as 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 Kodjo Aflagah, Abhijit Banerjee, Rukmini Banerji, Arja Dayal, Caroline Elliot, Maimuna Ginwalla, Rachel Glennerster, Junita Henry, Chavi Jain, Miranda Moolenaar, Shadrick Mwaba, Daniele Ressler, Sharon Schroen, 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 conflicting interests to declare and maintains the final editing rights among the two authors. All views expressed are those of the authors and not of any of the institutions with which they are affiliated.

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

‡Research and Learning Advisor, VVOB – Education for Development.

Timeline: The expected date of completion is April 30, 2025. Data cleaning for the endline is currently underway. The authors have no access to that data but have received a preliminary dataset for the trial's control group only. This Registered Report makes use of that control group data (e.g., to present attrition rates and report on the measurement properties of the study's outcome measures). The authors will not receive access to the complete endline dataset until the peer-review process at the *Journal of Development Economics* has been completed.

1 Introduction

1.1 Background and relevance of the study

There is growing consensus among research and policy circles on what works to address the learning crisis that plagues many developing countries (de Barros and Ganimian, 2023; The World Bank, 2017). Based on a systematic review of the evidence, the Global Education Evidence Advisory Panel identified "great buy" interventions, which are highly cost-effective and are supported by a strong body of evidence (GEEAP, 2023; Angrist et al., 2024). Among these policy guidelines, targeting instruction by learning level, not grade, features as a top recommendation, and adaptations of the "Teaching at the Right Level" program (TaRL) initially developed by the Indian NGO Pratham are currently being scaled by Ministries of Education in 15 developing countries.

However, consistently, high-efficacy educational interventions are found to be less effective (or even detrimental) once substantial external supports are removed, responsibilities are transferred from a non-governmental organization to the government, and implementation occurs during regular school functioning (Bold et al., 2018; Vivalt, 2020). This observation holds for TaRL, which can be effective if implemented with high fidelity by volunteers or during summer camps (Banerjee et al., 2017). In contrast, in India, only one study found positive effects of the program on student learning if implemented by teachers during regular classes (in Haryana), and only in one subject (literacy, not math); other evaluations documented null findings (in Bihar and Uttarakhand).[1] In Ghana, similarly, attempts to scale the program with public school teachers were unsuccessful.[2] Thus, it remains an open question how educational interventions, including TaRL, can maintain their effectiveness if they are implemented under government ownership at scale.

This study centers on the role of teachers and the potential of promoting the effectiveness of targeted instruction programs through continuous professional development. Building on a four-year research-practice partnership in Zambia, we present a large-scale

[1]The Haryana evaluation started with a program focused on mathematics and Hindi, in July 2012. In November 2012, the implementing organization Pratham conducted a midline assessment and "decided, following their midline assessment, to conduct a more concerted and intensive push for implementation of the Hindi curriculum over the math curriculum" (Duflo et al., 2015, 18). This change went into effect with a refresher teacher training in January 2013. The study's endline assessment started the following month, in February 2013. Despite this change in the program—late in the implementation period and endogenous to observed implementation fidelity—the Haryana evaluation is sometimes presented as a study of a program focused on Hindi only, not mathematics (Banerjee et al., 2017). In Haryana, the program did not expand after the evaluation, as a "lack of government buy-in and interest derailed (...) scale-up plans in the state" (Menon and Leach, 2019, 10).

[2]Duflo et al. (2024) document how, during spotchecks in Ghana, only 5.6 percent of the teachers assigned to the TaRL program taught to their intended group of students. Consistent with this lack of program take-up, the authors report precisely estimated null effects on student learning at the end of the study period.

cluster-randomized trial that investigates, through one experimental arm, the effectiveness of the country's national TaRL program (which is locally known as "*Catch Up*"). With a second experimental arm, the trial measures the impact of supporting the program's TaRL teachers through additional continuous professional development (CPD) opportunities in communities of practice. The randomized experiment involves 8,025 students attending 273 public primary schools. In secondary analyses, we set out to explore whether subgroups of students and teachers are affected differentially. In addition, we use fine-grained data on intermediate outcomes and mechanisms to pinpoint where, in the case of a null finding, the program's Theory of Change broke down.

The study's first and most direct contribution is to the evidence on the impact of targeted, differentiated instruction in developing countries. Prior research has demonstrated that the effectiveness of TaRL interventions correlates with program take-up and teachers' implementation fidelity (Angrist and Meager, 2023). Additionally, other studies have found that many public-school teachers tend to overestimate their students' abilities, underestimate the variability in student performance, and primarily engage with higher-achieving students ("teaching to the top") (Djaker et al., 2023). While these findings *motivate* efforts to enhance program sustainability through additional teacher-centered supports, assessing whether such supports *indeed causally improve* program outcomes will yield several valuable insights. First, it will clarify the role of teacher development within the program's broader Theory of Change, offering a clearer understanding of its relative impact on student outcomes. Second, by conducting a cost-effectiveness analysis, this study will provide policymakers with information on whether additional teacher-focused program components are a worthwhile investment of their limited resources. Third, more broadly, it may help policymakers determine whether to focus on the intensive margin by prioritizing the strengthening of implementer capacity.

Moreover, in developing countries, research on the factors that effectively drive teachers to change their instructional behaviors is limited, and it is unclear whether related theoretical frameworks developed in rich, Western countries "travel." Specifically, while Desimone's (2009) framework, which identifies five key program features expected to drive instructional change in the United States, is widely recognized, these elements may not easily apply to other contexts. Aligned with this framework, recent literature from the United States emphasizes the value of collective participation, such as collaborative conversations among teachers (Horn et al., 2017), pedagogically productive teacher talk (Lefstein et al., 2020), and inquiry-driven on-site problem-solving through peer facilitation (Gallimore et al., 2009).[3] However, a recent review of teacher professional development programs in developing

[3]Similarly, the TALIS 2018 survey, which covers predominantly developed countries, highlights the importance of school-embedded collaboration among teachers (OECD, 2019).

countries concluded that teacher participation in discussions and collaboration did not predict improvements in instruction or student learning (Popova et al., 2022). This finding underscores the need to examine the generalizability of Desimone's model across widely different contexts, which this study aims to address.

Lastly, our research also contributes to the literature on strengthening state capacity in education. Successfully scaling educational reforms often requires governments to improve their ability to manage, adapt, and effectively implement programs (Ganimian et al., 2024). Many impactful educational interventions have relied on non-governmental inputs, such as NGO-run summer camps (Banerjee et al., 2017), external organizations hiring additional para-teachers (Eble et al., 2021), or governments "outsourcing" public school management to private providers (Romero et al., 2020). Other interventions have involved external partners providing "packaged" solutions of many components (Maruyama and Igei, 2024). By contrast, efforts to improve government-provided education using only public sector resources have often struggled to enhance student learning (de Barros et al., 2024).[4] In this study, we examine a government-implemented program that operates solely with existing teachers and inspectors without added personnel. This approach offers rare evidence on how governments can implement large-scale reforms to boost public sector efficiency and improve student learning outcomes independently of external actors.

1.2 Research questions

The study's population of interest consists of Zambian public primary school students. We focus on grade-3 students and investigate their development over two years. We seek to answer the following primary research questions.

1. Does assignment to Zambia's national *Catch Up* program impact students' foundational skills in literacy and mathematics?

2. Does assignment to the continuous professional development (CPD) program for teachers impact the program's effectiveness in improving students' foundational skills in literacy and mathematics?

The study will also report on the per-student cost of both program variants and offer a cost-effectiveness analysis related to impacts on child learning.

The study seeks to answer the following secondary research questions.

[4]Private-public partnerships and government collaborations with NGOs are not at all immune to the challenges of scaling up promising interventions within the government system; for an unsuccessful attempt to scale a well-known early-childhood intervention within India's childcare system, see Arteaga et al. (2024).

1. Does a school's assignment to Zambia's national *Catch Up* program impact the subset of foundational mathematics skills explicitly targeted by the program?
2. Does a school's assignment to Zambia's national *Catch Up* program impact the foundational mathematics and literacy skills among two subgroups of students: girls and students in the lowest quartile of baseline learning levels?
3. Does a school's assignment to Zambia's national *Catch Up* program impact three sets of additional skills among students: creativity, socio-emotional skills, and working memory?

To better understand program mechanisms and impacts on potential mediators, the study will explore the following research questions.

1. Does a school's assignment to Zambia's national *Catch Up* program impact students' attitudes towards school and their attitudes toward math and literacy instruction?
2. Does a school's assignment to Zambia's national *Catch Up* program impact students' study habits at home?
3. Does assignment to the continuous professional development (CPD) program for teachers impact their continuous professional development (team-based problem-solving among teachers, verbal encouragement and discussions, and teachers' participation in practical demonstrations of teaching methods)?

To better understand program take-up and implementation fidelity, the study will also explore the following research questions.

1. To what extent is the *Catch Up* program implemented well and taken up as intended?
2. To what extent is the additional continuous professional development (CPD) program component implemented well and taken up as intended?
3. Does assignment to the continuous professional development (CPD) program for teachers impact program take-up and implementation quality?

Lastly, due to data limitations, we de-prioritized the following, additional research questions.

1. Does a school's assignment to Zambia's national *Catch Up* program impact teachers' ability to accurately diagnose their students' learning levels?

2. Does a school's assignment to Zambia's national *Catch Up* program impact teachers' perceived ability to affect student learning (i.e., their "locus of control")?

While our primary focus is on the above research questions concerning the effect of *assignment* to the *Catch Up* program, we will also investigate the effect of *exposure* to the program.

2 Research design

2.1 Basic methodological framework

This is a cluster-randomized trial with a waitlist design. We randomly assigned 91 of 182 zones and all public primary schools in those zones to receive the *Catch Up* program.[5] The remaining zones and their public primary schools were assigned to continue with business as usual; they may receive the program once the study is complete. 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 (the intent-to-treat, or "ITT" effect) and to compare this effect across the two program variants (with vs. without the CPD program component). Moreover, data on students' absenteeism and attendance of the program's remedial classes allows us to estimate local average treatment effects (LATE) for the students assigned to the program and exposed to it.

2.2 Hypotheses

Our pre-specified hypotheses regarding each of the primary research questions presented in section 1.2 are as follows.

P1. Assignment to Zambia's national *Catch Up* program impacts students' foundational skills in mathematics.

P2. Assignment to Zambia's national *Catch Up* program impacts students' foundational skills in literacy.

[5]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.

P3. Assignment to the continuous professional development (CPD) program for teachers impacts the program's effectiveness in improving students' foundational skills in mathematics.

P4. Assignment to the continuous professional development (CPD) program for teachers impacts the program's effectiveness in improving students' foundational skills in literacy.

Our pre-specified hypotheses regarding each of the secondary research questions are as follows.

S1. Assignment to Zambia's national *Catch Up* program impacts the subset of mathematics skills explicitly targeted by the program.

S2. Assignment to Zambia's national *Catch Up* program impacts the foundational literacy skills of students in the lowest quartile of baseline literacy levels.

S3. Assignment to Zambia's national *Catch Up* program impacts the foundational mathematics skills of students in the lowest quartile of baseline mathematics levels.

S4. Assignment to Zambia's national *Catch Up* program impacts students' creativity.

S5. Assignment to Zambia's national *Catch Up* program impacts students' socio-emotional skills.

S6. Assignment to Zambia's national *Catch Up* program impacts the foundational literacy skills of girls.

S7. Assignment to Zambia's national *Catch Up* program impacts the foundational mathematics skills of girls.

S8. Assignment to Zambia's national *Catch Up* program impacts students' working memory.

To better understand program mechanisms and impacts on potential mediators, the study will explore the following research questions.

M1. Assignment to Zambia's national *Catch Up* program impacts students' attitudes towards school and their attitudes toward mathematics and literacy.

M2. Assignment to Zambia's national *Catch Up* program impacts students' study habits at home.

M3. Assignment to the continuous professional development (CPD) program for teachers impacts their participation in and exposure to continuous professional development (team-based problem-solving among teachers, verbal encouragement and discussions, and teachers' participation in practical demonstrations of teaching methods).

To better understand program take-up and implementation fidelity, the study will also explore the following research question.

T1. Assignment to the continuous professional development (CPD) program for teachers impacts program take-up and implementation quality.

Our priority order of hypothesis tests is as per the above list (P1, followed by P2, ..., S1, ... M1, ..., T1). The study's abstract will report on hypothesis tests P1, P2, P3, P4, and S1.

If we do not find support for hypotheses P3 and P4, we see limited benefit in distinguishing the two program variants in the remainder of our analyses and will pool the two treatment arms (improving the statistical power of our hypothesis tests).[6] If we find support for hypotheses P3 or P4, favoring one of the two treatment groups, we will focus hypothesis tests S1 to M2 (and S1* to M2*, explained below) on comparisons between the more effective program variant and the control group.

Using machine learning-based methods, we will also explore the potential heterogeneity of program effects on students' literacy and mathematics skills across a large vector of student, teacher, and school background characteristics.

For completeness, in supplemental materials, we will also report on the program effects on various subskills of student learning (e.g., by content domains and cognitive domains). We will also report on program effects among the subgroups of students not pre-registered here (i.e., effects among boys and effects among the quartile of top-performing students), and on the difference in program effects across subgroups (e.g., the difference in effects among girls vs. boys).

We also considered two additional hypotheses for which we only have limited data. We de-prioritize these other research questions and may report on related results as additional information, only.

O1. Assignment to Zambia's national *Catch Up* program impacts teachers' ability to accurately diagnose their students' learning levels.

[6] We would report the variant-specific results in appendix materials.

O2. Assignment to Zambia's national *Catch Up* program impacts teachers' perceived ability to affect student learning (i.e., their "locus of control").

While our focus is on the above hypotheses concerning the effect of *assignment* to the program, we will also test corresponding hypotheses regarding the effect of *exposure* to the program. Specifically, we will test the following hypotheses: P1*, P2*, S1*, S2*, S3*, S4*, S5*, S6*, S7*, S8*, M1*, and M2*. Here, "*" denotes the change from assignment to exposure, and all else is analog to the above. For example, hypothesis P1* states "Exposure to Zambia's national *Catch Up* program impacts students' foundational skills in mathematics."

2.3 Outcome variables

This research rests on primary data collected across four data collection rounds, additional administrative data, and detailed data on program costs. All primary data is collected with independent research teams, not by the Ministry or NGOs. Data collection included a "baseline" and "endline" in all 273 schools sampled for the study (December 2022; December 2024), one round of unannounced school visits in a random subset of 72 program schools (split evenly across the two treatment arms; October 2023), and one round of unannounced visits to all 273 study 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).[7] The administrative data includes rich backend data on all teacher interactions with the CPD communication system and records on training attendance. Below, we list the study's outcome variables (along with their corresponding hypothesis tests in parentheses).

Foundational skills in literacy and mathematics (P1-P4). 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).

[7] 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).

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.[8] In mathematics, the assessments capture four content domains (basic arithmetic; data display; geometric shapes and measurement; and number sense). They also capture two cognitive domains that cut across the content domains (applied or higher-order thinking skills; procedural or lower-order thinking skills). In literacy, the assessments capture seven domains (phonemic awareness; phonics; vocabulary; listening comprehension; writing; reading fluency; reading comprehension). The blueprint also maps the test questions to grade-level expectations, following Zambia's official curriculum framework.[9]

The assessments recorded students' responses to all test questions (or test "items") for both test components. We use a two-parameter logistic (2PL) item response theory (IRT) model to aggregate these responses and generate continuous estimates of student ability. We generate one overall score per subject and standardize the score (with a mean of zero for the control group, as per the test score distribution at endline). Using baseline data, the average conditional reliability of the literacy and mathematics measures is 0.92 and 0.83, respectively.[10]

Mathematics skills explicitly targeted by the program (S1). 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 baseline data,

[8]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).

[9]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."

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

$$\frac{1}{N}\sum\left[1-\frac{\mathrm{SE}(\hat{\theta})^2}{\mathrm{Var}(\hat{\theta})}\right],$$

where N is the sample size, $\mathrm{SE}(\hat{\theta})$ is the predicted standard error of the ability estimate, and $\mathrm{Var}(\hat{\theta})$ is the predicted ability variance. Unlike marginal reliability, which uses the theoretical θ distribution, this measure reflects the predicted test performance for our specific sample.

the average conditional reliability of the measure of skills explicitly targeted by the program is 0.81.

Students' creativity (S4). 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.[11] 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 for the control group, the average conditional reliability of the creativity measure is 0.82.

Students' socio-emotional skills (S5). To measure students' socio-emotional skills, we combine and adapt the Perceiving AI-Generated Emotions assessment (PAGE; Weidmann and Xu, 2024) 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 for the control group, the average conditional reliability of the measure of socio-emotional skill is 0.61.

Students' working memory (S8). 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

[11] 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).

extended to nine boxes (in increments of one).[12] 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 to the control group at endline. Using endline data for the control group, the average conditional reliability of the measure of working memory is 0.87.

Student attitudes towards school, literacy, and mathematics (M1). 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 for the control group, the average conditional reliability of the measure of student attitudes is 0.86.

Student's study habits at home (M2). 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.

Teachers' participation in and exposure to continuous professional development (M3). To measure teachers' participation in and exposure to continuous professional development activities, we administered one-on-one interviews with them at baseline and during process monitoring. Appendix Figure A4 reports on the baseline results. 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 baseline data, the average conditional reliability of the summary measure of teachers' participation in and exposure to continuous professional development is 0.78.

[12]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.

Program take-up and implementation quality (T1). To measure children's exposure to the *Catch Up* program, during the endline child interviews, we recorded whether each student had attended any remedial class the previous school day (by subject). We also recorded the program name (*Catch Up* or other). To account for absenteeism and dropouts, we use data on whether each student was present in school during the first visit to the school at endline. Our *main* measure of program take-up multiplies the proportion of children exposed to a *Catch Up* class by the proportion of children present in school. 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.[13]

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 (O2). 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.

[13]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.

2.4 Intervention

2.4.1 The Teaching at the Right Level program in Zambia's public primary schools

Public provision of education, as governed by the Ministry of Education, is the most common type 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.[14] 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 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 2016-17 academic year (Lipovsek et al., 2023). The Zambian TaRL program focuses on grades three to five, and it is locally known as *Catch Up*. It currently covers nine (out of ten) provinces, reaching 5,700 schools and impacting about 855,000 learners per year. 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, number sense, and procedural arithmetic (de-emphasizing geometric shapes, measurement, and data display). For literacy, there are no notable differences. All of the program implementers are regular teachers and inspectors commonly assigned to the public primary schools, with no additional staff being assigned to schools and all NGO involvement being limited to technical assistance.

2.4.2 Supporting TaRL teachers with continuous professional development

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., 2025). 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

[14]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 will introduce English as the language of instruction.

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 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. Take-up of the CPD program has been promising. Among the schools assigned to the CPD group, 96 percent have had at least one teacher participating in a mastery challenge, and 63 percent of teachers have participated in the CPD program over WhatsApp.

2.5 Sample

Our sampling strategy followed a three-step process. First, we identified a convenience 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). With access to micro data 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 sub-sampled 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). We successfully surveyed 8,025 students (4,091 girls and 3,934 boys), or about 29 students per school.[15] These students represent the study's sample we have been tracking over the past two years.

Figure 1 reports on the study's minimal detectable effects (given the analytical strategy described below, power of 0.8, and a statistical significance level of 0.05). Three assumptions go into these calculations. Firstly, regarding attrition, our preliminary endline data for the control group suggest only 12.6 percent of the students could not be tracked. Given the preliminary nature of the data and since there may be differential attrition, we (conservatively) show calculations for a worst-case scenario of 20 percent attrition. Secondly, we do not know yet the intra-cluster correlation of students' growth from baseline to endline, conditional on covariates. We use the intracluster correlation of baseline scores, controlling for randomization stratum fixed effects and student demographics. Finally, we do not know yet how much endline variance can be explained with baseline covariates. We think an R-squared of 0.45 is reasonable but show a range of possible values. Our results indicate the study is well-powered to detect even small effects of approximately 0.12 standard deviations and between 2-3 percentage-point improvements over the baseline levels of students who can read a paragraph or do two-digit subtraction. These effects are in line with those of successful large-scale education programs in less-developed countries (Evans and Yuan, 2022), and they reflect the Ministry's minimal expectations for the program's success.

2.6 Randomization

We created three experimental groups of schools for the study. 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).[16] 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.[17] After that, in the

[15]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).

[16]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).

[17]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.

zones assigned to the program, we randomly assigned one school to receive the program together 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. Appendix Figure A2 provides a map of the study's sample of schools, along with their assignment to the three experimental groups. Table 1 and Appendix Figure A3 confirm randomization successfully led to three balanced groups of schools and students whose differences in observable characteristics do not exceed what can be expected by chance.

2.7 Theory of Change

In this section, we discuss what we expect to find for each of the research questions presented in section 1.2. Specifically, we tie together the hypotheses that we introduced in section 2.2.

2.7.1 The Teaching at the Right Level program in Zambia's public primary schools

The 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 2.3).

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 2.4). 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 2.3 (for a description of the program, see section 2.4).

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 2.3).

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.

2.7.2 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 2.4). 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.[18]

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 2.3).

[18]This includes data on whether messages were received and opened by their recipients.

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.

2.8 Variations from the intended sample, and non-compliance

We do not expect attrition to exceed the (conservative) 20 percent that we have already factored into our statistical power calculations (see section 2.5), based on recent studies with similar characteristics conducted by the Regional Office for Africa of the Abdul Latif Jameel Poverty Action Lab (J-PAL), through which we are conducting our study.

We may encounter differential attrition at endline if students in the treatment group have changed their propensity to come to school. We minimized this possibility, however, by conducting home visits at endline, in addition to conducting school visits. We discuss how we will address differential attrition in section 3.2.2.

We believe cross-over or "contamination" across experimental groups is highly unlikely. We randomized zones in the country to receive (or not receive) the program, which limits contamination from one school to another.[19] We also closely worked with the Ministry that oversees teacher training and program implementation. For the continuous professional development component, teachers in the control group have no way to self-select into that program or join its WhatsApp groups.

Non-compliance with the intervention is possible but of inherent interest to the study, given that it is an effectiveness trial of a national program. Take-up of the added continuous professional development (CPD) program component has been promising. Among the schools assigned to the CPD group, 96 percent have had at least one teacher participating in a mastery challenge, and 63 percent of teachers have participated in the CPD program over WhatsApp. However, here again, non-compliance would be of interest, given that the added program component builds on an earlier mixed-methods study and a pilot.

2.9 Data collection and processing

We adhere to J-PAL Africa's strict data collection protocols, including high-frequency checks for electronic forms, spot-checks and accompaniments, and weekly monitoring and debriefs for enumerators (see Glennerster, 2017; J-PAL, 2017).

[19] Schools in a zone might coordinate through their zone's support officer (the "Zone Inset Coordinator"), but collaboration above the zone level is highly unlikely.

3 Empirical analysis

3.1 Statistical model

Our identification strategy rests on the study's random assignment of schools to the three experimental groups. We will exploit this random assignment to estimate the causal effects of being assigned to the interventions through linear regressions, with the following specification.

$$Y_{isr}^{t} = \alpha_r + \beta_1^t T_{sr} + \beta_2^t D_{sr} + \boldsymbol{\delta'} \boldsymbol{X}_{isr}^{t=0} + \epsilon_{isr}^{t} \tag{1}$$

Here, Y_{isr}^{t} is the outcome of interest for student i in school s, and randomization stratum r, at time t. In our primary analyses, Y_{isr}^{t} represents test scores. The α_r terms are strata fixed effects, T_{sr} is the treatment dummy for the regular *Catch Up* program, D_{sr} is a dummy indicating a school's random assignment to the program with continuous professional development, and ϵ_{isr}^{t} is the residual. To increase precision, all specifications include $\boldsymbol{X}_{isr}^{t=0}$ as covariates. Measured at baseline ($t = 0$), $\boldsymbol{X}_{isr}^{t=0}$ is a vector of baseline controls selected by a Lasso procedure on student, teacher, and school characteristics. If selected by the Lasso procedure, this vector may include $Y_{isr}^{t=0}$ (a student's initial outcome of interest).

Concerning primary research question one (hypothesis tests P1 and P2), the coefficient of interest β_1^t reflects the interventions' intent-to-treat (ITT) effects for follow-up round t (where t reflects the endline data collection round). Concerning primary research question two (hypothesis tests P3 and P4), to assess whether the program effects differ, we will test the equality of coefficients β_1^t and β_2^t and compute their difference. If these hypothesis tests do not yield evidence in support of hypotheses P3 and P4 (at the conventional five-percent level of statistical significance), we will pool the two treatment groups in all other hypothesis tests (including P1 and P2). In our analyses of treatment effects among student subgroups (S2, S3, S6, S7), we will interact treatment with the respective subgroup indicators. In our exploratory analyses, we will employ a machine learning-based method (causal forests) to investigate heterogeneous effects among subgroups of schools, teachers, and students (following Carlana et al., 2022; Athey and Imbens, 2016).

In our additional analyses concerning program exposure, we will present local average treatment effects (LATE) estimates of the impact of actually receiving the program. We will estimate the LATE effects with the following equation.

$$Y_{isr}^{t} = \alpha_r + \mu^t A_{sr} + \boldsymbol{\delta'} \boldsymbol{X}_{isr}^{t=0} + \epsilon_{isr}^{t} \tag{2}$$

Here, A_{sr} is the main indicator of program exposure presented above, and everything else is defined as before. Since exposure may be endogenous, we instrument A_{sr} with our assignment indicators T_{sr} and D_{sr}.

3.2 Statistical methods

3.2.1 Estimation

We will estimate equation (1) using ordinary least-squares (OLS) regressions and equation (2) using instrumental-variable (IV) regressions. We cluster standard errors at the zone level (de Chaisemartin and Ramirez-Cuellar, 2020; Abadie et al., 2022).

3.2.2 Rules for handling missing values

We expect to encounter two types of missing data: attrition (e.g., students not participating in the endline assessment) or missing values (e.g., students participating in the endline, but not answering specific questions therein).

Missing values due to attrition. We will address the first type of missingness as follows. First, we will document the overall attrition rate. Then, we will investigate whether attrition is systematically related to intervention assignment by fitting a version of equation (1) that replaces the outcome variable with an indicator variable for not participating in the endline. Next, if we find differential attrition, following Barrera-Osorio et al. (2024) we will exploit tracking information and the number of days needed to survey a respondent to model their propensity to attrit (Behaghel et al., 2014; Molina Millán and Macours, 2017).[20]

Missing values due to non-response. We will address the second type of missingness as follows. For missing responses on outcome variables, we will scale responses using item-response theory (IRT) models that account for missing values by using concurrent calibration via marginal maximum likelihood estimation (Kolen and Brennan, 2004), given that non-response on specific questions is akin to missingness in any non-equivalent anchor test (NEAT) design in which not all respondents are administered the same questions. To maintain the largest possible sample size, we will exclude covariates with missing values.

[20]This approach improves upon conventional inverse-probability weighting (IPW) and Lee (2009) bounds estimations.

3.2.3 Definition and rules for handling outliers

We do not expect to encounter outliers because all of our outcome variables are measured on pre-determined scales. Therefore, we will not seek to identify outliers or winsorize results.

3.3 Multiple hypothesis testing

Accounting for a pre-registered hierarchy of primary, secondary, and exploratory research hypotheses, we will adjust for multiple hypothesis testing by computing the sharpened false discovery rate adjusted q-values. We prioritize the study's statistical tests in the given order and will not adjust p-values for our main tests of interest (P1-P4). In turn, we will calculate q-values within two families of tests: tests related to secondary research questions (S) and tests related to potential mediators and mechanisms (M).[21] We will not adjust p-values for our analyses of program take-up and implementation fidelity (T) and other, de-prioritized research questions (O).

[21] For example, S1 adds one test to our tests of primary research hypotheses P, and requires adjustment (but we ignore tests S2-S8, which are of lower priority). S2 adds another test of a secondary hypothesis that also requires adjustment (but we ignore tests S3-S8). Etc. Similarly, M1 adds one test to our tests of primary research hypotheses P, requiring adjustment, P2 adds another test that requires adjustment, etc.

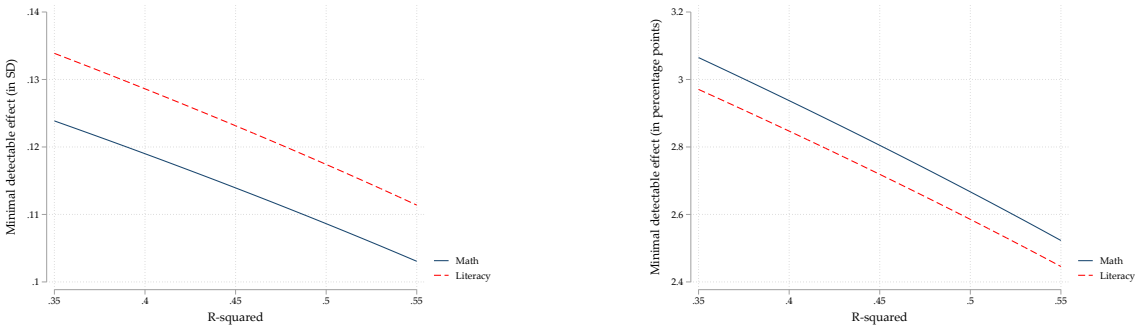
Tables and Figures

Table 1: *Balance table for student and school characteristics*

	Number of observations			Mean			Differences		
	Control	CU	CPD	Control	CU	CPD	CU vs Control	CU vs CPD	CPD vs Control
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Main outcomes									
Math score (IRT, std.)	2776	2707	2542	-0.00	-0.02	-0.04	-0.04	0.03	-0.07
				[1.00]	[1.00]	[1.00]	(0.06)	(0.05)	(0.06)
Literacy score (IRT, std.)	2776	2707	2542	-0.00	-0.07	0.02	-0.06	-0.05	-0.01
				[1.02]	[0.98]	[1.00]	(0.06)	(0.05)	(0.06)
Math score, ASER-related (IRT, std.)	2776	2707	2542	-0.00	-0.02	-0.02	-0.03	0.03	-0.06
				[1.00]	[1.01]	[1.00]	(0.06)	(0.05)	(0.05)
Math score, not ASER-related (IRT, std.)	2776	2707	2542	0.00	-0.02	-0.05	-0.04	0.03	-0.07
				[1.02]	[1.00]	[0.99]	(0.05)	(0.04)	(0.05)
Panel B: Sub-domains, math									
Percent correct, arithmetic	2776	2707	2542	0.40	0.39	0.39	-0.01	0.00	-0.01
				[0.24]	[0.24]	[0.24]	(0.01)	(0.01)	(0.01)
Percent correct, data	2776	2707	2542	0.49	0.48	0.47	-0.00	0.01	-0.02*
				[0.25]	[0.24]	[0.24]	(0.01)	(0.01)	(0.01)
Percent correct, geometry and shapes	2776	2707	2542	0.62	0.63	0.62	0.00	0.01	-0.01
				[0.19]	[0.19]	[0.18]	(0.01)	(0.01)	(0.01)
Percent correct, number sense	2776	2707	2542	0.29	0.28	0.28	-0.01	-0.00	-0.01
				[0.25]	[0.25]	[0.25]	(0.01)	(0.01)	(0.01)
Percent correct, applied math	2776	2707	2542	0.41	0.41	0.40	-0.01	0.00	-0.01
				[0.19]	[0.19]	[0.18]	(0.01)	(0.01)	(0.01)
Percent correct, procedural math	2776	2707	2542	0.48	0.48	0.48	-0.01	0.01	-0.01
				[0.19]	[0.19]	[0.19]	(0.01)	(0.01)	(0.01)
Percent correct, math (grade 1)	2776	2707	2542	0.64	0.64	0.63	-0.01	0.01	-0.01*
				[0.19]	[0.18]	[0.18]	(0.01)	(0.01)	(0.01)
Percent correct, math (grades 2-3)	2776	2707	2542	0.33	0.33	0.32	-0.01	0.01	-0.01
				[0.19]	[0.19]	[0.19]	(0.01)	(0.01)	(0.01)
Panel C: Sub-domains, literacy									
Percent correct, phonemic awareness	2776	2707	2542	0.58	0.59	0.58	0.01	0.01	-0.01
				[0.31]	[0.31]	[0.31]	(0.01)	(0.01)	(0.01)
Percent correct, phonics	2776	2707	2542	0.49	0.48	0.49	-0.00	0.00	-0.01
				[0.36]	[0.36]	[0.36]	(0.02)	(0.02)	(0.02)
Percent correct, vocabulary	2776	2707	2542	0.69	0.69	0.69	0.00	0.01	-0.01
				[0.29]	[0.29]	[0.29]	(0.01)	(0.01)	(0.01)
Percent correct, listening comprehension	2776	2707	2542	0.80	0.80	0.80	0.01	0.01	-0.00
				[0.30]	[0.30]	[0.31]	(0.01)	(0.01)	(0.01)
Percent correct, writing	2776	2707	2542	0.36	0.36	0.39	0.00	-0.01	0.01
				[0.40]	[0.40]	[0.40]	(0.02)	(0.02)	(0.02)
Percent correct, reading with comprehension	2776	2707	2542	0.14	0.12	0.14	-0.02	-0.01	-0.01
				[0.29]	[0.27]	[0.28]	(0.01)	(0.01)	(0.01)
Percent correct, literacy (grade 1)	2776	2707	2542	0.53	0.53	0.53	0.00	0.00	-0.00
				[0.23]	[0.22]	[0.23]	(0.01)	(0.01)	(0.01)
Percent correct, literacy (grades 2-3)	2776	2707	2542	0.39	0.39	0.40	-0.01	-0.00	-0.00
				[0.22]	[0.21]	[0.22]	(0.01)	(0.01)	(0.01)
Panel D: ASER, math									
ASER math at any number	2776	2707	2542	0.96	0.95	0.96	-0.01	-0.00	-0.01
				[0.20]	[0.21]	[0.20]	(0.01)	(0.01)	(0.01)
ASER math at two-digit or better	2776	2707	2542	0.63	0.63	0.63	0.00	0.01	-0.01
				[0.48]	[0.48]	[0.48]	(0.02)	(0.02)	(0.02)
ASER math at three-digit or better	2776	2707	2542	0.15	0.17	0.16	0.02	0.02	0.00
				[0.36]	[0.37]	[0.37]	(0.02)	(0.02)	(0.02)
ASER math at addition or better	2776	2707	2542	0.15	0.14	0.14	-0.00	0.01	-0.01
				[0.35]	[0.35]	[0.35]	(0.01)	(0.02)	(0.02)
ASER math at subtraction or better	2776	2707	2542	0.09	0.09	0.09	-0.00	0.01	-0.01
				[0.29]	[0.29]	[0.29]	(0.01)	(0.01)	(0.01)
ASER math at multiplication or division	2776	2707	2542	0.04	0.04	0.03	-0.01	0.00	-0.01
				[0.20]	[0.19]	[0.18]	(0.01)	(0.01)	(0.01)
Panel E: ASER, literacy									
ASER literacy at letter or better	2776	2707	2542	0.44	0.41	0.45	-0.03	-0.02	-0.00
				[0.50]	[0.49]	[0.50]	(0.03)	(0.02)	(0.02)
ASER literacy at word or better	2776	2707	2542	0.35	0.32	0.35	-0.03	-0.02	-0.01
				[0.48]	[0.47]	[0.48]	(0.02)	(0.02)	(0.02)
ASER literacy at para or better	2776	2707	2542	0.12	0.11	0.13	-0.02	-0.02	-0.00
				[0.33]	[0.31]	[0.33]	(0.01)	(0.01)	(0.01)
ASER literacy at story	2776	2707	2542	0.09	0.07	0.09	-0.02**	-0.02	-0.01
				[0.29]	[0.26]	[0.29]	(0.01)	(0.01)	(0.01)
Panel F: Child characteristics									
Student is female	2776	2707	2542	0.51	0.52	0.50	0.01	0.02	-0.01
				[0.50]	[0.50]	[0.50]	(0.01)	(0.01)	(0.01)
Asset index (ICW, std.)	2776	2707	2542	-0.00	-0.02	-0.03	-0.03	0.00	-0.03
				[1.00]	[1.02]	[1.02]	(0.06)	(0.05)	(0.06)
Home language different from schools' language	2776	2707	2542	0.44	0.48	0.42	0.03	0.04*	-0.01
				[0.50]	[0.50]	[0.49]	(0.03)	(0.02)	(0.03)
Best friend in school, attends the same grade	2776	2707	2542	0.69	0.70	0.67	0.01	0.03	-0.02
				[0.46]	[0.46]	[0.47]	(0.02)	(0.02)	(0.02)
Panel G: School characteristics									
No. of grade-3 girls	91	91	91	33.95	30.77	26.10	-3.18	4.67	-7.85**
				[27.49]	[30.05]	[17.45]	(3.51)	(3.51)	(3.51)
No. of grade-3 boys	91	91	90	33.37	30.37	28.11	-3.00	2.33	-5.33
				[26.56]	[29.47]	[18.21]	(3.45)	(3.46)	(3.46)
Proportion present (/not dropped out)	91	91	91	0.75	0.75	0.74	-0.00	0.02	-0.02
				[0.14]	[0.16]	[0.16]	(0.02)	(0.02)	(0.02)

Notes. This table provides descriptive statistics for the study sample, by treatment status. Standard deviations in brackets; standard errors in parentheses (standard errors for student-level data are clustered at the zone level). Continuous test scores are aggregated with a two-parameter logistic item response theory (IRT) model. The asset index reflects the inverse-covariance-weighted (ICW) average across eight yes/no questions. Continuous test scores and the asset index are standardized with respect to the control group. Estimations of group differences include randomization strata fixed effects (F.E.s). * significant at 10%; ** significant at 5%; *** significant at 1%.

Figure 1: *Power calculations*



(a) *IRT scores* **(b)** *ASER*

Note: This figure reports on the study's minimal detectable effect sizes for impacts on the continuous test scores, as per a two-parameter logistic (2PL) item response theory (IRT) model (to the left, measured in standard deviations) and for the ASER (to the right, measured in percentage point increases). Both plots assume a worst-case scenario, in which 20 percent of students attrit between baseline and endline. For the ASER, in math, we focus on the percentage of students who can at least do subtraction (which has a baseline level of 9.3 percent); in literacy, we focus on the percentage of students who can at least read a paragraph (which has a baseline level of 11.9 percent). Power= 0.8; $p = 0.05$. We show a range of possible R-squared values; we believe a value of 0.45 or more is reasonable. Calculations of the intra-cluster correlation (ICC) reflect the within-school correlation of residuals from a regression of each of the four ability measures on randomization strata fixed effects and student demographics (these ICCs range from 0.04 to 0.12).

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.

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.

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.

Arteaga, I., de Barros, A., Ganimian, A.J., 2024. The Challenges of Scaling up Effective Child-Rearing Practices Using Technology in Developing Settings: Experimental Evidence From India. Journal of Research on Educational Effectiveness (in press) doi:https://doi.org/10.26300/9fb8-g360. journal of Research on Educational Effectiveness.

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.

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.

Barrera-Osorio, F., de Barros, A., Filmer, D., 2024. Long-term Impacts of Primary School Scholarships: Evidence from Cambodia. Journal of Policy Analysis and Management 43, 10–38. doi:10.1002/pam.22533.

de Barros, A., Fajardo-Gonzalez, J., Glewwe, P., Sankar, A., 2024. The Limitations of Activity-Based Instruction to Improve the Productivity of Schooling. The Economic Journal 134, 959–984. doi:10.1093/ej/uead099.

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. economics of Education Review.

de Barros, A., Henry, J., Mathenge, J., 2025. What Drives Teachers to Change Their Instruction? A Mixed-Methods Study from Zambia. Comparative Education Review (in press) URL: https://de-barros.com/publication/de-barros-what-2021/de%20Barros%20et%20al.%20-%202021%20-%20What%20Drives%20Teachers%20to%20Change.pdf.

Behaghel, L., Crépon, B., Gurgand, M., Le Barbanchon, T., 2014. Please Call Again: Correcting Nonresponse Bias in Treatment Effect Models. The Review of Economics and Statistics 97, 1070–1080. URL: https://doi.org/10.1162/REST_a_00497, doi:10.1162/REST_a_00497.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., Sandefur, J., 2018. Experimental evidence on scaling up education reforms in Kenya. Journal of Public Economics 168, 1–20. URL: https://linkinghub.elsevier.com/retrieve/pii/S0047272718301518, doi:10.1016/j.jpubeco.2018.08.007.

Carlana, M., La Ferrara, E., Pinotti, P., 2022. Goals and Gaps: Educational Careers of Immigrant Children. Econometrica 90, 1–29. doi:10.3982/ECTA17458. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA17458.

de Chaisemartin, C., Ramirez-Cuellar, J., 2020. At What Level Should One Cluster Standard Errors in Paired Experiments, and in Stratified Experiments with Small Strata? arXiv:1906.00288 [econ] URL: http://arxiv.org/abs/1906.00288. arXiv: 1906.00288.

Desimone, L.M., 2009. Improving Impact Studies of Teachers' Professional Development: Toward Better Conceptualizations and Measures. Educational Researcher 38, 181–199. URL: https://doi.org/10.3102/0013189X08331140, doi:10.3102/0013189X08331140. publisher: American Educational Research Association.

Djaker, S., Ganimian, A.J., Sabarwal, S., 2023. Out of Sight, Out of Mind? The Gap between Students' Test Performance and Teachers' Estimations in India and Bangladesh. Working Paper 23-750. Annenberg Institute at Brown University. Providence, RI. URL: https://edworkingpapers.com/ai23-750. publisher: edworkingpapers.com.

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.

Duflo, E., Berry, J., Mukerji, S., Shotland, M., 2015. A wide angle view of learning Evaluation of the CCE and LEP programmes in Haryana, India. Report 22. International Initiative for Impact Evaluation (3ie). New Delhi, India.

URL: https://www.3ieimpact.org/sites/default/files/2017-11/ie_22_evaluation_of_cce_and_lep_in_haryana.pdf.

Eble, A., Frost, C., Camara, A., Bouy, B., Bah, M., Sivaraman, M., Hsieh, P.T.J., Jayanty, C., Brady, T., Gawron, P., Vansteelandt, S., Boone, P., Elbourne, D., 2021. How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in the Gambia. Journal of Development Economics 148, 102539. doi:10.1016/j.jdeveco.2020.102539.

Evans, D.K., Yuan, F., 2022. How Big Are Effect Sizes in International Education Studies? Educational Evaluation and Policy Analysis , 01623737221079646doi:10.3102/01623737221079646.

Gallimore, R., Ermeling, B., Saunders, W., Goldenberg, C., 2009. Moving the Learning of Teaching Closer to Practice: Teacher Education Implications of School-Based Inquiry Teams. The Elementary School Journal 109, 537–553. URL: https://www.journals.uchicago.edu/doi/full/10.1086/597001, doi:10.1086/597001. publisher: The University of Chicago Press.

Ganimian, A.J., Muralidharan, K., Walters, C.R., 2024. Augmenting State Capacity for Child Development: Experimental Evidence from India. Journal of Political Economy 132, 1565–1602. doi:10.1086/728109. publisher: The University of Chicago Press.

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.

Glennerster, R., 2017. The Practicalities of Running Randomized Evaluations: Partnerships, Measurement, Ethics, and Transparency, in: Banerjee, A.V., Duflo, E. (Eds.), Handbook of Economic Field Experiments. Elsevier. volume 1, pp. 175–243. doi:10.1016/bs.hefe.2016.10.002.

Horn, I.S., Garner, B., Kane, B.D., Brasel, J., 2017. A Taxonomy of Instructional Learning Opportunities in Teachers' Workgroup Conversations. Journal of Teacher Education 68, 41–54. URL: https://doi.org/10.1177/0022487116676315, doi:10.1177/0022487116676315. publisher: SAGE Publications Inc.

J-PAL, 2017. J-PAL Research Protocols. URL: https://drive.google.com/file/d/0B97AuBEZpZ9zZDZZbV9abllqSFk/view.

Kolen, M.J., Brennan, R.L., 2004. Test Equating, Scaling, and Linking. 3rd ed., Springer, New York, NY.

Lee, D.S., 2009. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. The Review of Economic Studies 76, 1071–1102. doi:10.1111/j.1467-937X.2009.00536.x.

Lefstein, A., Vedder-Weiss, D., Segal, A., 2020. Relocating Research on Teacher Learning: Toward Pedagogically Productive Talk. Educational Researcher 49, 360–368. URL: https://doi.org/10.3102/0013189X20922998, doi:10.3102/0013189X20922998. publisher: American Educational Research Association.

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:https://doi.org/10.4337/9781802205930.00012.

Maruyama, T., Igei, K., 2024. Community-Wide Support for Primary Students to Improve Foundational Literacy and Numeracy: Empirical Evidence from Madagascar. Economic Development and Cultural Change 72, 1963–1992. doi:10.1086/726178. publisher: The University of Chicago Press.

Menon, R., Leach, B., 2019. Using evidence to improve children's foundational skills: a successful teaching and learning approach expands in India and beyond. Brief. International Initiative for Impact Evaluation (3ie). Delhi, India. URL: https://3ieimpact.org/sites/default/files/2019-11/Haryana-EU-brief.pdf.

Molina Millán, T., Macours, K., 2017. Attrition in Randomized Control Trials: Using Tracking Information to Correct Bias. Discussion Paper 10711. IZA Institute of Labor Economics. Bonn. URL: http://ftp.iza.org/dp10711.pdf.

OECD, 2019. TALIS 2018 Results: Teachers and School Leaders as Lifelong Learners. volume 1. OECD Publishing, Paris. URL: https://www.oecd-ilibrary.org/content/publication/1d0bc92a-en. type: doi:https://doi.org/10.1787/1d0bc92a-en.

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.

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. URL: https://doi.org/10.1093/wbro/lkab006, doi:10.1093/wbro/lkab006.

Romero, M., Sandefur, J., Sandholtz, W.A., 2020. Outsourcing Education: Experimental Evidence from Liberia. American Economic Review 110, 364–400. doi:10.1257/aer.20181478.

The World Bank, 2017. Learning to Realize Education's Promise. World Development Report 2018. The World Bank, Washington, D.C. OCLC: 992735784.

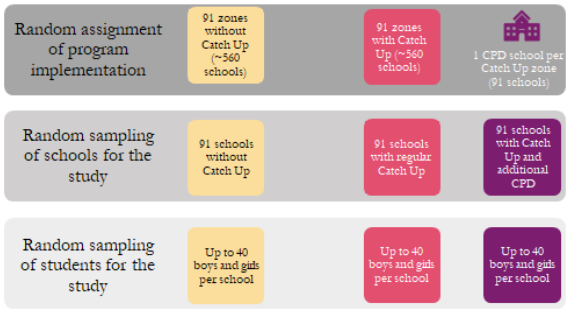
Vivalt, E., 2020. How Much Can We Generalize From Impact Evaluations? Journal of the European Economic Association , 1–45URL: https://academic.oup.com/jeea/advance-article/doi/10.1093/jeea/jvaa019/5908781, doi:10.1093/jeea/jvaa019.

Weidmann, B., Xu, Y., 2024. PAGE: A Modern Measure of Emotion Perception for Teamwork and Management Research. doi:10.48550/ARXIV.2410.03704. version Number: 1.

Appendices

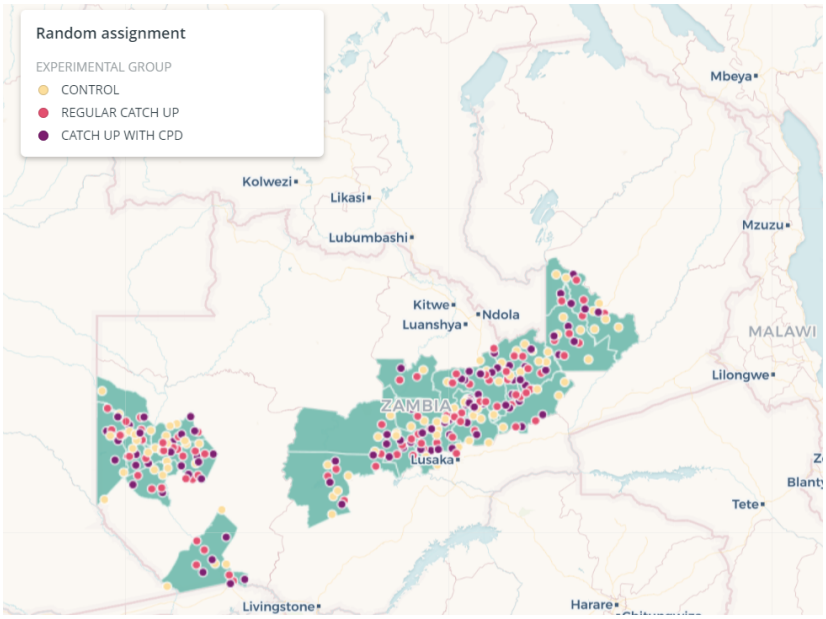
Appendix A: Additional figures and tables

Figure A1: *Sampling and randomization procedure*



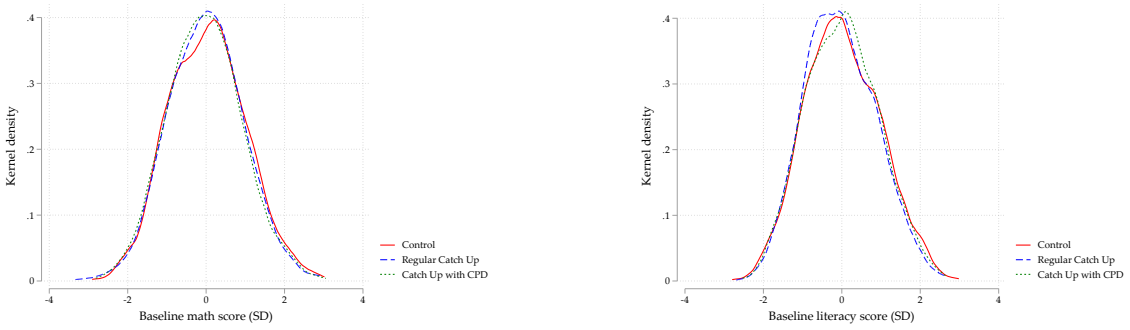
Note: This figure summarizes the study's sampling and randomization procedure. Within the convenience sample of 182 zones, we randomized half the zones to the Catch Up program and the other half to the control group. Within each control zone, we randomly sampled one school for the study. Within each 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 CPD component. Within each sampled school, we randomly sampled up to 40 boys and girls for the study (stratified by gender).

Figure A2: *Geographic scope of the study*



Note: This figure provides a map of the study's schools in Zambia. Each dot reflects a school; experimental groups are shown in different colors, as per the map's legend. Central province, Western province, and Itezhi Tezhi district (in Southern province) are highlighted in green. Across these areas, the study covers a convenience sample of 182 zones (schools of other zones do not show on the map).

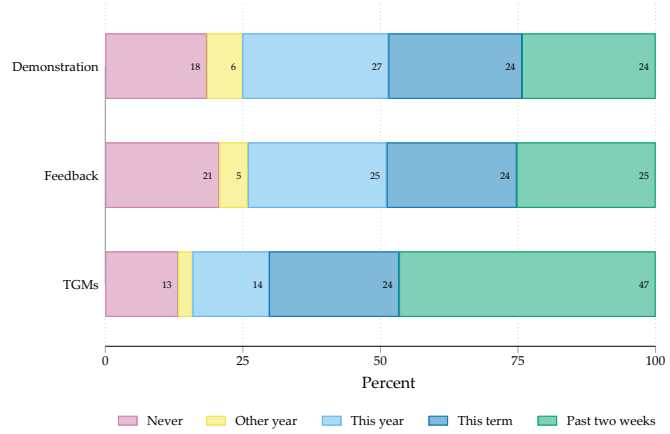
Figure A3: *Balance on baseline test scores*



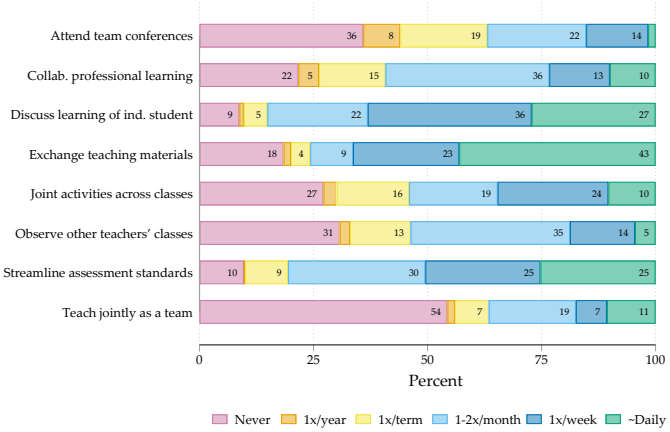
(a) *Mathematics* **(b)** *Literacy*

Note: This figure reports on the sample's balance across the three groups, as per the baseline tests in mathematics and literacy. Test scores are aggregated with a two-parameter logistic (2PL) item response theory (IRT) model, standardized, and centered with respect to the control group. Each panel shows kernel density plots, by treatment status, of residuals from a regression of baseline test scores on strata fixed effects. The left panel reports results for mathematics; the right panel reports results for literacy.

Figure A4: *Teachers' participation in professional development activities and collaboration*



(a) *Professional development activities*



(b) *Peer collaboration*

Note: These figures report on teachers' (self-reported) frequency of participation in professional development activities (top panel) and peer collaboration with other teachers (bottom panel). "Demonstration" refers to teachers' participation in someone else's practical demonstration of how to teach something, "Feedback" to whether they have received one-on-one feedback on their teaching, and "TGMs" to teacher group meetings. Bar labels show percentages (labels are omitted for values below three percent).