A RANDOMIZED EVALUATION OF THE *Catch Up* PROGRAM IN ZAMBIA: BASELINE REPORT

Authors: Andy de Barros (MIT, J-PAL); Theresa Lubozha (VVOB – Education for Development); Prince Muraguri (TaRL Africa); Nico Vromant (VVOB – Education for Development)

Date: April 6, 2023

This document reports on the 2022 baseline data collection of the "Catch Up" evaluation in Zambia. The baseline took from July to November 2022 and covered 273 government primary schools across 182 zones (all sampled schools). The baseline report highlights the following points.

1. *Sample*. 8,025 grade-3 students participated in the baseline. These students represent the study's sample that will be tracked over the following two years. The study includes randomly sampled students from among those present during a school visit. Of the formally enrolled students, 11.8 percent had not been to school in the past four weeks; 23.6 percent of those who had come to school in the past four weeks were absent on the day of the visit.

2. *Data*. We conducted one-on-one assessments with students to measure their foundational skills in mathematics and literacy. To facilitate the tracking of students, we also collected information on the names of students' fathers and mothers. We collected additional background information for all students, including their gender, home language, and socio-economic status. We also conducted interviews with grade-3 teachers.

3. *Student performance*. We find low levels of student performance in mathematics and literacy. For example, only 9.3 percent of students knew how to subtract two-digit numbers, and only 11.9 percent could read a short paragraph. Our findings also highlight performance differences across student subgroups, particularly for literacy—by province, by students' poverty levels, by whether children speak the school's official language at home, and by whether students say that their best friend attends the same school and grade.

4. *Balance at baseline*. There is balance at baseline, with no difference in students' math and literacy performance across the three experimental groups.

5. *Test and item characteristics*. Tests performed well, with satisfactory levels of reliability and no evidence of floor effects. Tests successfully covered a wide range of (content and cognitive) subdomains, capturing those subdomains typically understood to jointly constitute foundational mathematics and literacy.

We are extremely grateful for the excellent collaboration between J-PAL, Teaching at the Right Level Africa, VVOB – Education for Development, and Zambia's Ministry of Education on this round of data collection. We also thank Innovations for Poverty Action for data collection activities. We thank the GPE Knowledge & Innovation Exchange (KIX) on Education for funding.

1 Introduction

Remedial education and differentiated instruction are promising approaches to tackle the low learning levels that plague many low- and middle-income countries. However, little is known about how to promote these strategies at scale. This study evaluates the impact of the "Teaching at the Right Level" program on students' foundational literacy and mathematics skills.[1] The program divides children into groups based on their learning needs and pace and adds extra time during which teachers provide tailored instruction to each group. The program runs in grades 3 to 5 of Zambia's government and community primary schools and is locally known as "Catch Up." The study also investigates the effectiveness of combining the Catch Up program with a continuous professional development program for teachers.

This document reports on the study's baseline. Section 2 summarizes the research design, including the study's sampling strategy, its random assignment of zones and schools to experimental groups, the measurement approach and instruments, and the overall study timeline. Section 3 describes operational details of the baseline, including field logistics, data quality checks, and data cleaning procedures. Section 4 provides the baseline's main findings concerning student performance in mathematics and literacy. Section 5 explores additional questions regarding students who do not speak their school's official language at home, teacher characteristics, and a comparison of students' performance across countries (with students in India and Nigeria). Finally, Section 6 offers additional technical details, covering an analysis of baseline balance, power calculations, and a psychometric analysis of the study's student assessments.

2 Research design

2.1 Sampling

Our sampling strategy followed a three-step process. First, in collaboration with the Ministry and VVOB, we identified a convenience sample of 182 zones.[2] 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.[3]

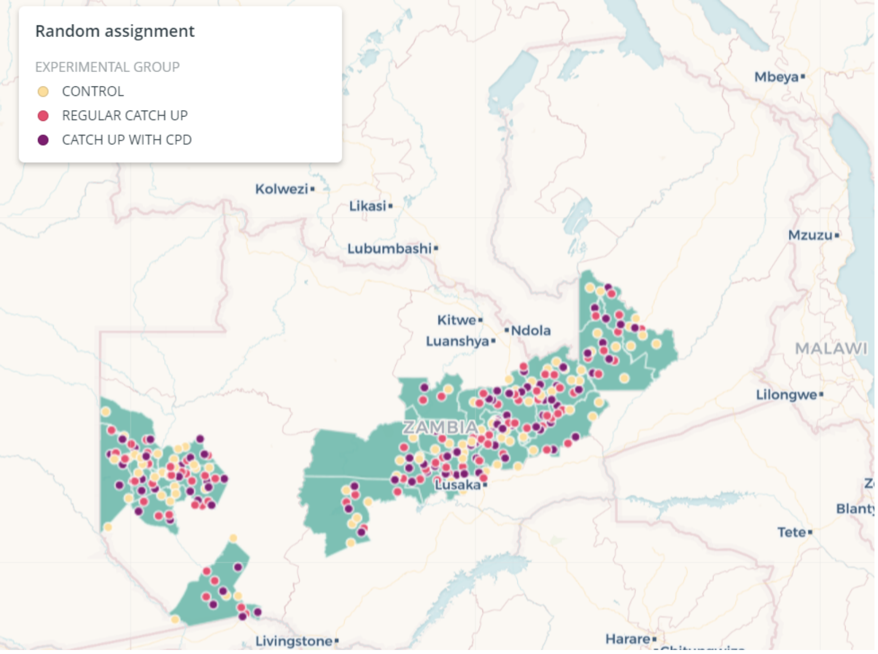
Second, we determined the sample of 273 schools. If a zone gets assigned to receive Catch Up, all publicly-funded schools in that zone will be targeted by the program (government-run schools and government-supported "community" schools). Yet, for the study's data collection activities, we drew a random subsample of government schools (excluding community schools). In collaboration with the Ministry, we first constructed a list of all 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. Sampling gave each government school an equal probability of being selected; we did not have access to accurate enrolment data and, therefore, we could not weight schools' selection probabilities proportionally to their size.

[1] We do not distinguish between the terms foundational "mathematics" and foundational "numeracy." However, for consistency, we refer to foundational "mathematics" throughout this document.

[2] 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] Until 2021, the district in Southern province (Itezhi Tezhi district) belonged to Central province; it is now part of Southern province. In any province-level analyses, we will group this district along with the eleven districts of Central province. This is for simplicity—we do not pass judgment on geographical or political boundaries.

Figure 1: 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).

In the third and final step, during school visits, we randomly sub-sampled third-graders from among those who were present on the day of the visit. We stratified our sampling by gender and selected a maximum of 40 students per school (not all schools had 40 students present). We successfully surveyed 8,025 students (4,089 girls and 3,936 boys), or about 29 students per school.[4] These students represent the study's sample that we will track over the following two years.

2.2 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).[5] After that, in the zones assigned to the program, we randomly assigned one school to receive the program together with an additional continuous professional development (CPD) program (which is not part of the regular Catch Up intervention). In what follows, we refer to these three sets of schools as the "Control", "regular Catch Up", and "Catch Up with CPD" groups, respectively. Figure 1 provides a map of the study's sample of schools, along with their assignment to the three experimental groups. Appendix Figure A.1 summarizes the study's sampling and randomization procedures.

[4] 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).

[5] 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).

We point to two additional technical details regarding our randomization procedure. First, 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. 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.[6] Second, at the zone level and the school level, we repeated our randomization ten times to make sure the groups did not drastically differ in observable characteristics. When we randomized zones, we did not have access to baseline data and had to balance mathematics and language test scores from the grade-7 exams; when we randomized Catch Up schools into receiving the CPD component or not, we were able to use the baseline data to balance student test scores and the proportion of female students.[7]

2.3 Measurement

We measure four sets of variables in the study. First, our main outcome of interest is children's learning of foundational skills in mathematics and literacy. Beyond children's overall learning, in secondary analyses, we also concentrate on sub-skills (e.g., arithmetic vs. geometry in mathematics or reading vs. listening comprehension in literacy). Second, we collect information on child, teacher, and school characteristics. These variables provide descriptive insights regarding the study context, allow for the exploration of heterogeneous treatment effects across subgroups, and increase the statistical precision of impact estimates. Third, we measure potentially mediating variables, such as whether teachers know their students' learning levels. Fourth, we document whether the program is implemented well and taken up as intended. Here, we describe the first three sets of measures in more detail, focusing on those variables measured at baseline; we will only observe the program's implementation and take-up later in the study, during process monitoring.[8]

2.3.1 Measuring students' foundational skills in mathematics and literacy

We measured students' foundational skills in mathematics and literacy 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

[6] 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.

[7] Our randomization strategy closely follows recent methodological discussions around whether researchers should randomize just once or re-randomize multiple times to achieve the greatest level of balance for a given set of baseline covariates. Bruhn and McKenzie (2009) argued for the former; more recently, Banerjee et al. (2020) suggested pre-specifying the randomization strategy and choosing a conservative number of re-randomizations (100 or fewer). We followed Banerjee et al. (2020), conducted ten randomization (at the zone and the school levels), and selected the most balanced of these randomizations. More specifically, for each of the ten randomizations, we recorded the maximum t-statistic (using linear regressions to compare covariates across experimental groups). Then, we selected the randomization with the smallest maximum t-statistic (Bruhn and McKenzie (2009) call this procedure the "min-max method").

[8] In the baseline teacher interviews, we did, however, already include questions about teachers' participation in teacher group meetings and their membership in WhatsApp groups among colleagues. These aspects will be relevant for teachers' take-up of the CPD intervention.

(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.[9] 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.[10] Appendix Tables A.1 and A.2 show the number of test questions per domain and curricular grade levels, for the two subjects,

We administered all student assessments in the official, local language of a given school (Bemba, Lozi, Nyanja, and Tonga).[11] Translations and local adaptations included multiple field pilots, discussions during enumerator training sessions, and an expert review with native speakers who had taught in the given language (one expert per language).

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).[12] In mathematics, we also generate one continuous, standardized score for those skills targeted by the program (i.e., number recognition and procedural arithmetic) and a similar score capturing the remaining domains (e.g., geometry). In addition, we calculate the proportion of correctly answered questions by content domain, cognitive domain, and questions' curricular grade level. Finally, we also report on the proportion of students who have mastered discrete levels of ability, as per the ASER tests. In number recognition and arithmetic, we focus on whether students can at least do two-digit subtraction with borrowing; in literacy, we focus on whether students can at least read a short paragraph of about 30 words (making three mistakes or fewer).[13]

[9] 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).

[10] As per Zambia's official curricular expectations, the assessments covered materials from grades 1 to 3.

[11] We also prepared tests in Kaonde. However, no sampled school used Kaonde as its official language.

[12] Here and elsewhere, we follow Penney (2023) and standardize variables by dividing them by the square root of their within variation.

[13] We also report on the remaining ability levels, as per the ASER tests. In mathematics, they capture whether students can recognize one-digit, two-digit, or three-digit numbers from a list, whether students can add two-digit numbers, whether students can multiply a two-digit number by a one-digit number, or whether students can divide a two-digit number by a one-digit number (with a remainder). For number recognition, to be marked proficient, students need to recognize at least four numbers; for arithmetic, they need to solve two problems correctly. In literacy, the remaining levels reflect whether students can recognize letters, whether they can read words, or whether they can read a short story of about 75 words). To be marked proficient at a given level, the child must be able to read at least four letters, at least four words, or read the story (making three mistakes or fewer).

2.3.2 Measuring background characteristics at the student, teacher, and school levels

Along with the one-on-one assessments with students, we also administered a battery of questions to capture additional background characteristics. More specifically, we recorded students' gender and asked whether the student had a "best friend" in the same school and grade. We also documented the basic composition of the household (e.g., the number of siblings) and whether the student speaks the school's official language at home. Lastly, we construct an index of household assets and socio-economic status by calculating the (standardized) covariance-weighted average of students' responses to eight yes-no questions.[14]

We also attempted to interview all grade-3 mathematics and literacy teachers in the sample of schools.[15] These interviews captured a teacher's gender, level of education, years of teaching experience, whether the teacher has an education degree and teaching license, and the main language they speak at home. We also collected data on whether the teacher owns a phone. To measure their content knowledge, we asked them to solve two division problems and two reading comprehension questions from our child assessments. Finally, we administered the Marlowe-Crowne social desirability scale (which measures respondents' concern with social approval) and a short battery of five questions to measure teachers' locus of control (which measures their belief that they, as opposed to external forces, can impact child learning).

At the school level, we recorded information during a phone call to the school before the visit and in person during the school visit. More specifically, we collected data on student enrollment numbers by grade, the number of teachers (by grade and subject), and levels of student absenteeism. In addition, we recorded each school's geographic coordinates.

2.3.3 Measuring potential mediators

Because of the Catch Up intervention, we expect teachers to develop a better understanding of students' learning levels. During our baseline interviews, we, therefore, asked teachers whether they thought their students were able to correctly answer select math and reading tasks (from the baseline assessment). We asked these questions in general, and individually, for a random subsample of a teacher's students. In addition, we asked teachers to rank a random subsample of students by their ability. We can compare these responses to our baseline assessments for the same students.

Because of the additional continuous professional development (CPD) intervention, we moreover expect teachers to engage in team-based problem solving, discuss their teaching with colleagues, and receive verbal encouragement at work. We also expect them to observe additional demonstrations of teaching practices. We included related questions in our teacher interviews.

2.4 Study timeline

The baseline data collection took from July to November 2022 (when students were about to finish their third grade of primary school). We completed the randomization of zones to the Catch Up

[14] We asked about the household's connection to electricity, whether the household owned a television, a charcoal cooking stove ("Mbaula"), a gas or electric stove, a sofa, an electric iron, or a phone (including any cell phone), and whether anyone had consumed dairy products in the past two weeks.

[15] There are 339 grade-3 mathematics and literacy teachers across the 273 schools. We interviewed 309 (92.2 percent) of them, 23 (6.8 percent) were absent on the day of the school visit, three (0.9 percent) refused, and for the remaining four (1.2 percent), the non-response reason is unknown.

program in June 2022, and schools' random assignment to the continuous professional development component in December 2022. The overall program rollout launched in January 2023. The rollout of the CPD intervention was planned for January 2023, but it is delayed (VVOB and TaRL Africa currently expect to launch the CPD intervention in March 2023). There will be two rounds of process monitoring in all 273 schools, in June and July 2023 (round 1), and one year later in June and July 2024 (round 2). Throughout the study period, we will also collect data on VVOB's training and monitoring activities. The study's endline data collection is planned for October and November 2024 (about two years after the baseline assessment).

3 Baseline implementation

Innovation for Poverty Action (IPA) Zambia collected the baseline data in two phases. Phase one was in term two, and phase two was in term three of the 2022 school year. At the beginning of each phase, the IPA field team conducted a short phone survey with all sampled schools to confirm their existence, enrollment number, and official language. Each phase also started with a week-long training of field staff. IPA managed to cover all 273 schools over this period.

All school visits followed a standard protocol. Upon entering a new district, IPA's team would pay a courtesy call to the district office and get permission to visit schools. The field teams would call schools a day before a visit and ask them to have the third-grade enrollment register ready. For schools that operated grade-3 classes in morning and afternoon shifts, the team also informed the school to ask learners to come in one session. On the day of the visit, IPA enumerators reported to a school an hour before classes started. The supervisor would then ask for the grade-3 enrollment roster, conduct the subsampling of students, assign her team members their respective lists of students, and conduct all teacher interviews. The median duration of student interviews was 37 minutes.[16]

We used high-frequency checks, accompaniments, and spot checks to control the quality of data collection. During daily high-frequency checks, we would analyze the incoming data to detect suspicious patterns. From these checks, the field-team supervisors would receive feedback and requests to accompany those surveyors requiring additional training and supervision. Innovations for Poverty Action was also required to have its supervisors accompany at least 20 percent of child interviews and submit a digital form for each accompaniment.[17] IPA team leaders, TaRL Africa, VVOB, and Ministry of Education staff also randomly selected teams to conduct spot checks.

The field team followed strict protocols to consider ethical concerns and ensure child protection. First, all study procedures are under the oversight of ethics review boards in the United States and Zambia. This required that all Parent Teacher Associations (PTA) could opt out of the study (none of them did). In addition, all surveyors received dedicated training in child protection and data security protocols. Beyond such "minimum must-dos," we trained surveyors to ensure students feel well and in a safe space during their interviews. We also tracked any unexpected events that would have required that the ethics review boards be informed (we did not detect any). Further, we implemented several Covid-19 protection protocols (e.g., testing field staff and distributing face masks and hand sanitizer).

Although they did not notably affect the study's data quality and integrity, there are three diffi-

[16] During the student interviews, the median ASER test took 12 minutes. The median time spent on other test questions was 14 minutes. The remaining time was needed to ask non-assessment questions and to set up the interview.

[17] Effectively, IPA accompanied 14.9 percent of child interviews.

culties we would like to highlight. First, in one school, its staff brewed a local alcoholic beverage on school grounds. Field staff verified that this did not compromise child safety and we reported this case to the Ministry of Education. Second, some schools showed high absenteeism levels due to local events (e.g., funerals), farming, or if the local community was suspicious of the field team. We attempted to prevent such cases with courtesy calls and revisited 13 schools a second time (4.8 percent). Third, Innovations for Poverty Action had delays in the survey implementation, leading to the survey extending into a second school term (instead of one term, as planned originally). All survey operations concluded before the intervention was launched in schools.

4 Main results

4.1 Student performance by subject

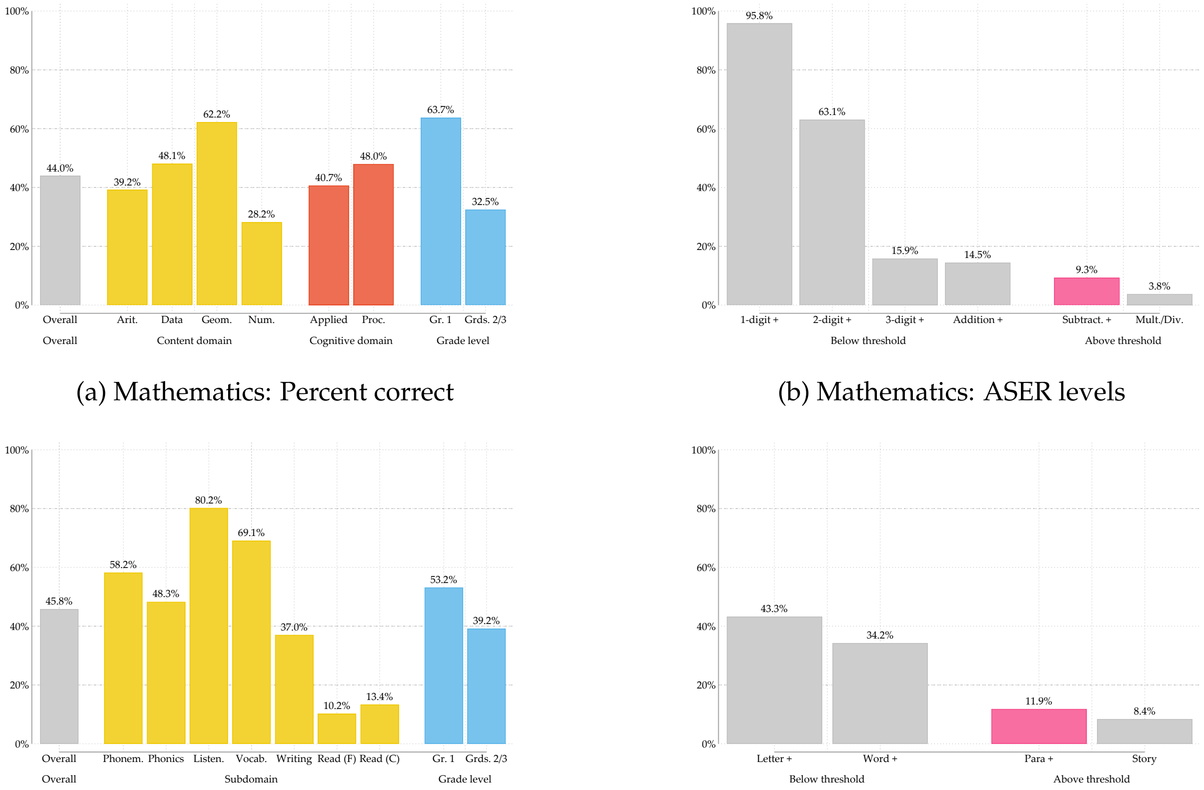
Figure 2 reports on students' performance on the mathematics and literacy assessments. The top panels report results for mathematics; the bottom panels report results for literacy. The left panels show the average percentage of questions students answered correctly; the right panels report the percentage of students who either reach or exceed a given level of mastery, as per the ASER test.[18] As shown on the left, the tests were easy enough to avoid floor effects. On average, students could answer about half of the test questions correctly (44 percent in mathematics and 46 percent in literacy). Yet, these questions required very low levels of skills. As shown on the right, the ASER tests suggest only 9.3 percent of students knew how to subtract two-digit numbers, and only 11.9 percent could read a short paragraph of about 30 words (making three mistakes or fewer).

Figure 2 also displays students' performance by subskill and curricular grade level. In mathematics, students performed better on those content domains related to data display (41 percent of questions answered correctly), and geometry and measurement (62 percent). In contrast, number sense and number recognition, and arithmetic-related questions were harder for students (28 percent and 39 percent, respectively). Moreover, the ASER test suggests students' number recognition skills drop off after two-digit recognition (63 percent could recognize four two-digit numbers, but only 16 percent could recognize four three-digit numbers). As expected, applied questions were harder for students than procedural questions (41 vs. 48 percent). Finally, students were much more familiar with curricular content mapped to grade 1 (64 percent of questions answered correctly) vs. materials mapped to grades 2 and 3 (33 percent).

In literacy, the results by subskill and curricular grade level suggest most students can understand the school's local language—they answered 80 percent of the listening comprehension questions and 69 percent of the vocabulary-related questions correctly. Yet, students solved only about half of the questions that asked them to identify differences in sounds (phonemic awareness, 58 percent) or to connect sounds to letters (48 percent). This finding is corroborated by the ASER test's percentage of students who can read four letters (43 percent). Students solved 37 percent of the tasks that required them to write. Finally, few students were able to solve questions that asked them to read with fluency (10 percent) or to read with comprehension (13 percent). Reassuringly, we once again find very similar results across the two assessment components (i.e., other test questions vs. the ASER assessment): The categorization of students as per the ASER test places 12 percent of students at the "paragraph" level and 8 percent at the "story" level. Lastly, as with

[18] "Overall" imputes a zero for those students who were not administered a given question, due to the increasing difficulty of the test and its adaptive nature (e.g., students who could not read letters are counted as not being able to solve a reading comprehension question).

Figure 2: Student performance



(a) Mathematics: Percent correct

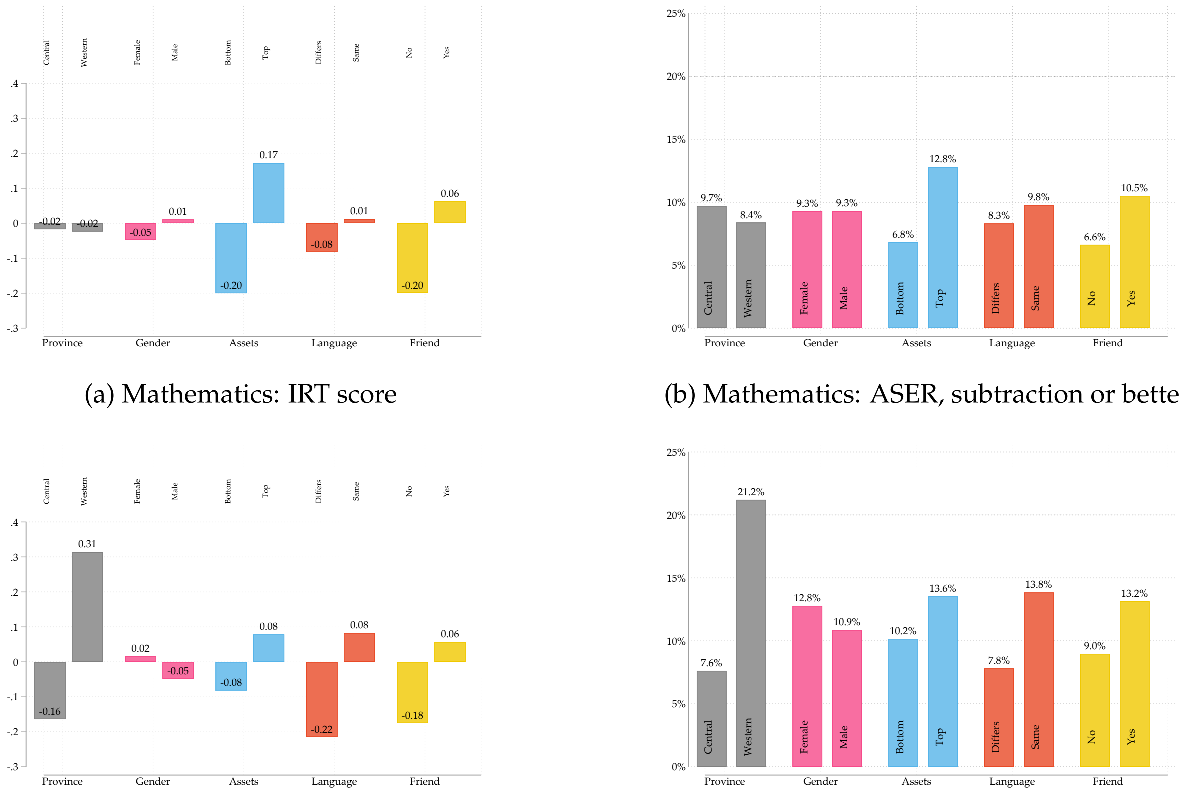
(b) Mathematics: ASER levels

(c) Literacy: Percent correct

(d) Literacy: ASER levels

Note: This figure reports on students' performance on the tests, by sub-skill and curricular expectations. The top panels report results for mathematics; the bottom panels report results for literacy. The left panels show the average percentage of questions students answered correctly; the right panels report the percentage of students who either reach or exceed a given level of mastery, as per the ASER test. "Overall" imputes a zero for those students who were not administered a given question, due to the increasing difficulty of the test and its adaptive nature (e.g., students who could not read letters are counted as not being able to solve a reading comprehension question). "Gr. 1" refers to questions mapped to curricular expectations for grade 1; "Grds. 2/3" refers to questions mapped to curricular expectations for grades two and three. Abbreviations for content domains in mathematics are as follows: "Arit.": Arithmetic; "Geom.": Geometric shapes and measurement; "Num.": Number sense. Abbreviations for cognitive domains in mathematics are as follows: "Applied": Applied, higher-order thinking skills; "Proc.": Procedural, lower-order thinking skills. Abbreviations for subdomains in literacy are as follows: "Phonem.": Phonemic awareness; "Listen.": Listening comprehension; "Vocab.": Vocabulary; "Read (F)": Reading with fluency; "Read (C)": Reading with comprehension. "Below threshold" vs. "above threshold" indicates the ASER levels of main interest; the percentage of students who crossed this threshold is highlighted in pink.

Figure 3: Heterogeneity in student performance



(a) Mathematics: IRT score

(b) Mathematics: ASER, subtraction or better

(c) Literacy: IRT score

(d) Literacy: ASER, paragraph or better

Note: This figure reports on students' performance by province, gender, household assets, whether their home language differs from the school's language, and whether their best friend attends the same school and grade. For simplicity, "Central" includes Central province and Itezhi Tezhi district in Southern province. For household assets, the figure compares students in the top vs. bottom quartile of the study's asset index. The top panels report results for mathematics; the bottom panels report results for literacy. The left panels show aggregate test scores as per a two-parameter logistic (2PL) item response theory (IRT) model (standardized); the right panels report the percentage of students who either reach or exceed the ASER tests' subtraction and paragraph levels, respectively.

mathematics, we see a steep drop in students' ability to solve questions related to grade-1 material (54 percent of questions answered correctly) vs. curricular content mapped to grades 2 and 3 (39 percent).

4.2 Performance across subgroups of students

We also investigate heterogeneity in student performance for five sets of student subgroups: by province, student gender, household wealth, whether students speak the school's official language at home, and whether students say that they have a "best friend" in the same school and grade. In Figure 3, the top panels show each group's performance in mathematics; the bottom panels report respective results for literacy. The left panels show aggregate (and standardized) test scores as per the study's two-parameter logistic (2PL) item response theory (IRT) model; the right panels report the percentage of students who either reach or exceed the ASER tests' subtraction and paragraph levels.

We observe large differences in students' literacy performance, with students in Western province outperforming their peers in Central province and Itezhi Tezhi district (by 0.47 standard deviations overall and a 14-point gap in the percentage of students who can read a paragraph). In math, we do not see a similar difference across provinces. For both subjects, we do not detect large differences by student gender. Students in the top wealth quartile (as per household assets) performed better than those in the bottom quartile. This difference is more pronounced in mathematics (0.37 standard deviation overall and a 6.0-point gap in the percentage of students who can do subtraction), as compared to literacy (0.16 standard deviations and a 3.4-point gap in the percentage of students who can read a paragraph. Whether students speak the school's official language of instruction at home is associated with smaller differences in mathematics (0.09 standard deviations overall; 1.5 percentage points on the ASER) yet large differences in literacy (0.30 standard deviations overall; 6.0 percentage points on the ASER).[19] Lastly, students whose best friend attends the same grade and school outperformed others by 0.26 standard deviations in math and 0.24 standard deviations in literacy (3.9 and 4.2 percentage points on the ASER, respectively).[20]

5 Additional insights

5.1 Literacy performance of students who do not speak the school's language at home

In Section 4.2, we documented substantial differences in the literacy performance across students who do vs. do not speak the school's language of instruction at home. Here, we explore whether these differences can be linked to students' acquisition of specific literacy subskills.[21]

By literacy subskill, Figure 4 compares the percentage of correctly answered test questions for those students speaking their school's language at home (white bars) and other students who speak another language at home (yellow bars). Students who speak another language at home perform worse on every literacy subskill. Yet, this gap is comparatively small for listening comprehension (a gap of 5.0 percentage points). This suggests most students can comprehend the school's official language of instruction, independent of whether they speak it at home. In turn, we observe the largest performance gap in students' ability to produce written text (a gap of 13.3 percentage points). We also observe a gap in students' vocabulary skills (7.7 percentage points). This may suggest students who do not speak the school's language of instruction at home struggle with language production in that language.

5.2 Insights from teacher surveys

5.2.1 Teacher background characteristics

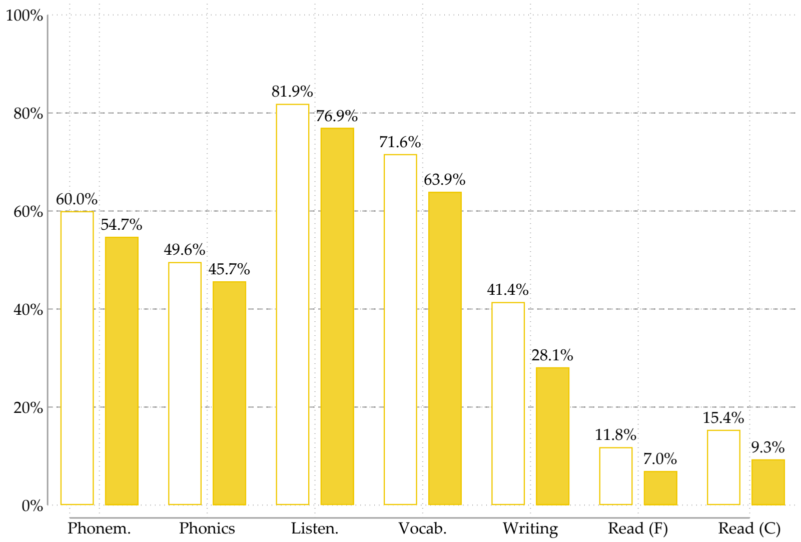
In Table 1, Panel A summarizes information on basic teacher demographics and teachers' access to phones. About two-thirds of the teachers are female (64 percent). On average, teachers are 37 years old and have about nine years of teaching experience. Almost all (94 percent) have a teaching certificate, and the great majority (81 percent) are in a permanent and pensionable position. Nearly all (98 percent) of the teachers own a phone, about four out of five (79 percent) own a smartphone,

[19] Speaking a different language at home is a prevalent phenomenon: 33.0 percent of students do not speak their school's official language as the primary language at home.

[20] In our sample, 31.2 percent of students said they did not have a "best friend" in their grade and school.

[21] Rodriguez-Segura (2022) points to the "Simple View of Reading" and suggests oral literacy skills complement decoding skills as students learn to read.

Figure 4: Literacy subskills by students' home language



Note: By literacy subdomains, this figure compares the percentage of correctly answered test questions for those students who speak their school's language at home (white bars) and other students who speak another language at home (yellow bars). Abbreviations for subdomains in literacy are as follows: "Phonem.": Phonemic awareness; "Listen.": Listening comprehension; "Vocab.": Vocabulary; "Read (F)": Reading with fluency; "Read (C)": Reading with comprehension.

and more than half (55 percent) are members of a school-internal WhatsApp group together with their colleagues.

Panel B of this table reports on teachers' performance on some of the same assessment questions we administered to students. Four out of five teachers (79 percent) could solve one (out of two) division questions correctly, but less than two-thirds (62 percent) solved both questions correctly. Virtually all (92 percent) could read a short story fluently (as opposed to haltingly or not being able to read), and almost all could answer two comprehension questions about this text (95 percent). This is even though more than a quarter of teachers (27 percent) mainly speak a language at home that differs from the school's official language of instruction.

5.2.2 Teacher beliefs and awareness of student learning levels

In Table 1, Panel C provides information on teachers' awareness of their students' learning levels. 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 and reading a paragraph). We then compare the teacher's estimate with the observed proportion in her school, as per the student assessments. For subtraction, we compare teachers' guesses with students' performance on a single subtraction question (with carrying); we also compare teachers' guesses with students' classification as per the ASER test, which requires that students solve two subtraction questions (with carrying). We find teachers drastically overestimate their students' ability (by 47 percentage points for the subtraction question and by 27 percentage points for the share of students who can read a paragraph).

Panel D of this table reports on whether teachers believe they, as opposed to external forces, impact

Table 1: Teacher characteristics, beliefs, and awareness of student learning levels

	N (1)	Mean (2)	SD (3)	Min (4)	Max (5)
Panel A: Background characteristics					
Female	309	0.64	0.48	0.00	1.00
Age (in years)	289	37.13	8.09	19.02	67.00
Teaching experience (in years)	292	8.66	7.70	0.00	35.00
Has a teaching certificate	309	0.94	0.24	0.00	1.00
Employment is permanent and pensionable	309	0.81	0.39	0.00	1.00
Owns a phone	309	0.98	0.13	0.00	1.00
Owns a smartphone	309	0.79	0.41	0.00	1.00
Member of a WhatsApp group with colleagues	308	0.55	0.50	0.00	1.00
Panel B: Content knowledge and home language					
One of two division questions	309	0.79	0.41	0.00	1.00
Both division questions	309	0.62	0.49	0.00	1.00
Reads fluently	302	0.92	0.28	0.00	1.00
One of two reading comprehension questions	302	0.99	0.08	0.00	1.00
Both reading comprehension questions	302	0.95	0.22	0.00	1.00
Home language different from schools' language	309	0.27	0.45	0.00	1.00
Panel C: Awareness of student learning levels					
Guessed proportion for subtraction question	309	0.59	0.21	0.04	1.00
Overestimate (vs. subtraction question)	309	0.47	0.21	–0.13	1.00
Overestimate (vs. ASER subtraction level)	309	0.49	0.21	–0.10	1.00
Guessed proportion for paragraph	309	0.41	0.24	0.00	1.00
Overestimate (vs. ASER paragraph level)	309	0.27	0.22	–0.23	0.91
Panel D: Locus of control					
There is little I can do to help a student's learning.	309	0.32	0.47	0.00	1.00
Pupils come unprepared from previous grades.	309	0.61	0.49	0.00	1.00
Parents do not seek feedback from the teacher on pupil performance.	309	0.65	0.48	0.00	1.00
Parents do not have the necessary education to help their child.	309	0.75	0.43	0.00	1.00
If parents would do more for their children, I could do more.	309	0.83	0.38	0.00	1.00

Notes. This table provides descriptive statistics about teachers' background characteristics, their beliefs with respect to whether they, as opposed to external forces, can impact student learning ("locus of control"), and awareness of student learning levels. Some teachers refused to answer individual questions (see Column 1 for the number of observations). Except for age and years of teaching experience, all rows reflect proportions. Teachers' age as of 31 December 2022. Years of experience as per the calendar year of the teacher's first posting (counting 2022 as zero years of experience). "Guessing" refers to a teacher's estimate of the proportion of her students who can solve a specific question from the study's student assessments (solving a subtraction problem, and reading a paragraph). "Overestimate" compares the teacher's estimate with the observed proportion in her school, as per the student assessments. For subtraction, we compare teachers' guesses with students' performance on a single subtraction question (with carrying); we also compare teachers' guesses with students' classification as per the ASER test, which requires that students solve two subtraction questions (with carrying). Under "locus of control," the table reports on the proportion of teachers who said they agree (vs. disagree) with each of the five statements.

student learning (their perceived "locus of control"). About a third of teachers (32 percent) agreed with the statement that they could do little to help a student's learning. Six out of ten teachers (61 percent) agreed that their students came unprepared from earlier classes. Finally, about two-thirds (65 percent) said parents did not request feedback on students' performance, three-quarters (75 percent) believed parents lacked the necessary education to help their child learn, and even more (83 percent) stated that they could do more if parents did more for their children.

5.2.3 Teachers' participation in continuous professional development and collaboration

Figure 5 reports on teachers' (self-reported) frequency of participation in professional development activities (top panel). More than half of the teachers (52 percent) said that, in the current term, they had not participated in someone else's practical demonstration of how to teach something. Similarly, in the current term, more than half (51 percent) had not received one-on-one feedback on their teaching. Also, more than half of the teachers (53 percent) did not comply with the expected frequency of participating in teacher group meetings every two weeks.

The figure also describes teachers' (self-reported) frequency of collaborating with other teachers (bottom panel). Teachers stated they frequently collaborated with their colleagues outside the classroom. About three-quarters of the teachers said that, at least twice a month, they exchanged learning materials (72 percent), and discussed the learning development of specific students (77 percent). About two-thirds (64 percent) said they worked at least twice a month with other teachers to use similar standards for assessing student progress. At the same time, about a third had never observed another teacher's class (31 percent), and more than half (54 percent) of the teachers had never taught a lesson together with a colleague. Also, despite the mandate to hold teacher group meetings every other week, more than a third of the teachers (36 percent) had never attended a team conference.[22]

5.3 International comparison of student performance

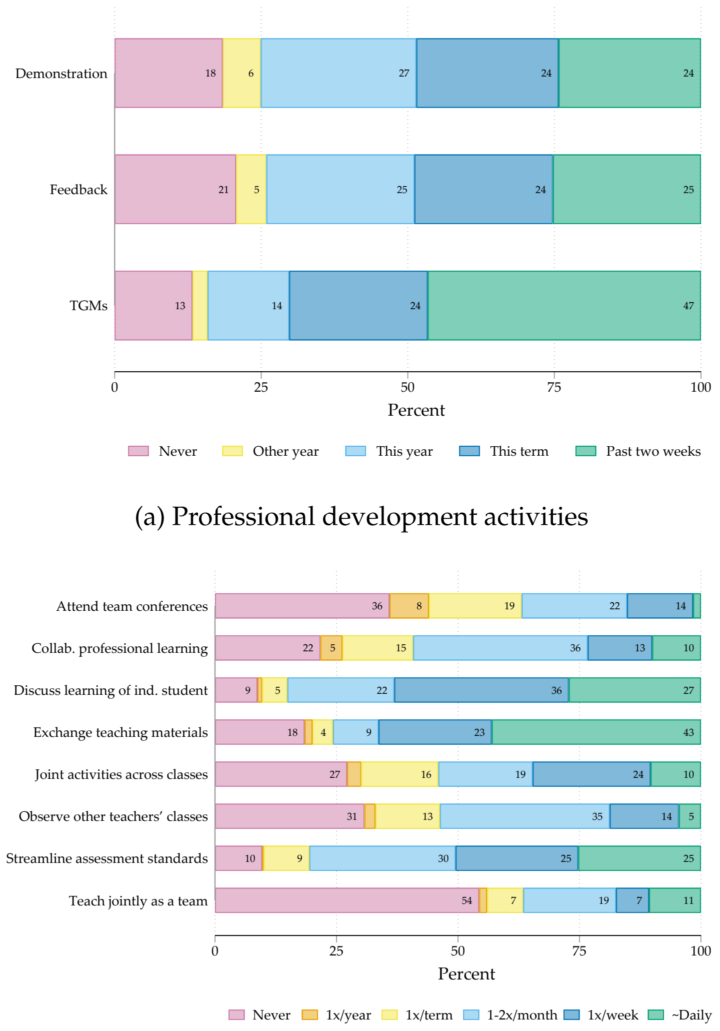
This section compares the mathematics results from Zambia with other assessments from government schools in Karnataka (India) and Kano State (Nigeria). The Indian data stems from the 2018 baseline of a randomized controlled trial one of us conducted across 292 public primary schools (de Barros et al., 2022); the Nigerian data stems from the 2021 endline of a quasi-experimental evaluation of TaRL Africa's work across 180 public primary schools. As in the Zambian baseline, both cases estimate results for in-person, one-on-one assessments. Using common test questions as "anchors", item response theory allows us to map the ability estimates from these two countries onto the same scale we use for Zambia (Kolen and Brennan, 2004).

As shown in Figure 6, we document very large performance differences across the three contexts. Nigeria's third-graders performed approximately 1.2 standard deviations below their peers from Zambia, and even the fourth-graders performed two-thirds of a standard deviation below Zambia's third-graders. In contrast, Karnataka's third-graders performed 0.6 standard deviations better than the Zambian students. These findings are also corroborated by each country's ASER results (not shown in the figure). For example, in the Nigerian sample, only 3.6 percent of grade-3 students knew how to subtract (or divide); this number is 9.3 percent in our baseline data from Zambia (see above) and 32.4 percent in the sample from Karnataka.[23]

[22] Note that only 13 percent of teachers said that they had never attended a teacher group meeting. This discrepancy (36 vs. 13 percent) may suggest teachers do not perceive these teacher group meetings as "team conferences."

[23] The ASER reports for government schools in rural Karnataka put this number at 23.5 percent in 2018 and 19.6

Figure 5: Teachers' participation in professional development activities and collaboration

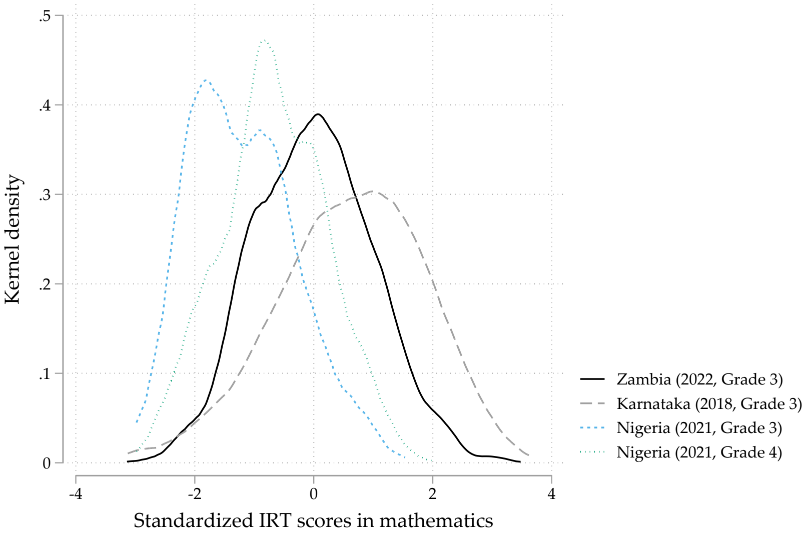


(a) Professional development activities

(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).

Figure 6: International comparison of mathematics performance



Note: This figure provides kernel density plots comparing the mathematics results from Zambia with other assessments from government schools in Karnataka (India) and Kano State (Nigeria). All results are linked onto the Zambian scale, using anchor items and item response theory; item parameters from India and Nigeria are constrained to match those of the Zambian results. The Indian data stems from the 2018 baseline of a randomized controlled trial (across 292 public primary schools) (de Barros et al., 2022); the Nigerian data stems from the 2021 endline of a quasi-experimental evaluation of TaRL Africa's work (across 180 public primary schools). All three studies estimate results for in-person, one-on-one assessments.

6 Technical considerations

6.1 Baseline balance

In this section, we present evidence that randomization led to three comparable groups of students. In Table 2, Panel A compares the study's main outcomes in math and literacy, including one math score we expect to be more closely related to the program and the ASER (covering number sense and arithmetic) and another math score that is not (or less closely) related to the ASER (covering the remaining subskills, such as geometry and data display). Panel B compares the three subgroup's performance on more fine-grained math subdomains; Panel C provides a similarly fine-grained comparison for literacy subdomains. Panels D and E present comparisons for all ASER levels. Finally, Panels F and G show group comparisons for student and school characteristics.

Across the three experimental groups, after controlling for randomization stratum fixed effects, we do not detect group differences beyond what would have been expected by chance. Importantly, the main baseline test scores are balanced, with no statistically significant differences ($p > 0.1$). In fact, not just the means but the distribution of baseline achievement was quite similar across experimental groups (see Appendix Figure A.2). Table 2 shows minor differences for a small number of subdomains and background characteristics. They do not include the main outcomes, and there is no clear pattern. Yet, to the extent that these variables predict students' endline performance, we will include them as covariates in our analyses of program impacts (selecting controls via Lasso).

6.2 Analytical strategy and power calculations

Now that the study's sample size is known, we can update our estimations about the probability that our statistical tests will detect an impact on student learning should such an impact exist (i.e., statistical power). To this end, we first present our analytical strategy; thereafter, we report on the minimal detectable effect sizes for a commonly accepted level of statistical power (0.8).

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

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

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 overall test scores. In our secondary analyses, Y_{isr}^t are either measures of subskills or potentially mediating variables. 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 $Y_{isr}^{t=0}$ and $\boldsymbol{X}_{isr}^{t=0}$ as covariates. Measured at baseline ($t = 0$), $Y_{isr}^{t=0}$ is a student's initial outcome of interest, and $\boldsymbol{X}_{isr}^{t=0}$ is a vector of baseline controls selected by a Lasso procedure on student, teacher, and school characteristics. The coefficients of interest, β_1^t and β_2^t, reflect the interventions' intent-to-treat (ITT) effects, for

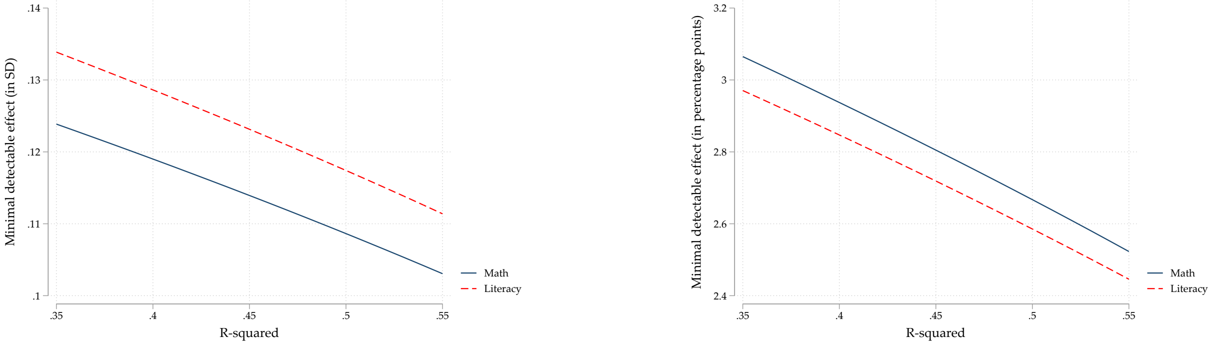
percent in 2022 (ASER Centre, 2023). This discrepancy may be due to different sampling strategies. First, the ASER reports on the entire state, not just the two districts covered by de Barros et al. (2022) (Bijapur and Tumkur). Second, the ASER reports focus on rural locations only. Third, the randomized trial did not track students to their homes.

Table 2: 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 7: Power calculations for a worst-case scenario of attrition



(a) IRT scores

(b) ASER

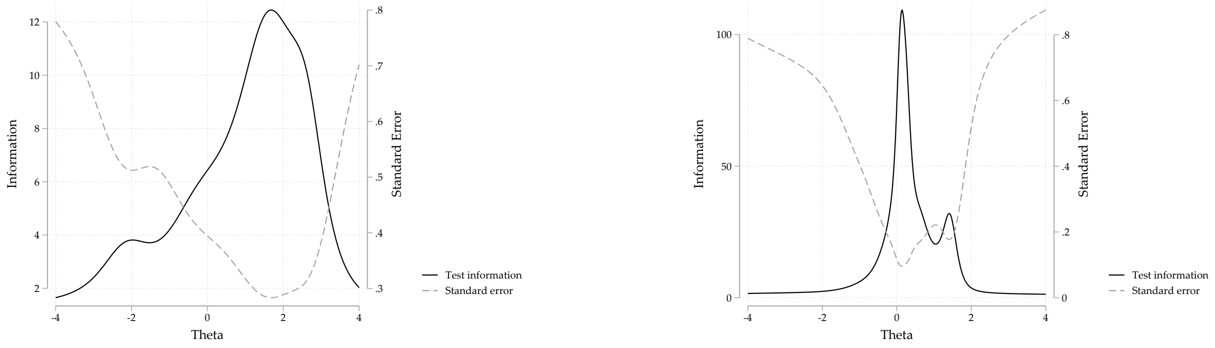
Note: This figure reports on the study's minimal detectable effect sizes for impacts on the continuous test scores, as per a two-parameter logistic (2PL) item response theory (IRT) model (to the left, measured in standard deviations) and for the ASER (to the right, measured in percentage point increases). 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).

follow-up round t (where t may reflect process monitoring rounds or the endline data collection). To assess whether the program effects differ, we will also test their equality. We cluster standard errors at the zone level (cf. Abadie et al., 2022).[24]

Figure 7 reports on the study's minimal detectable effects (given the above analytical strategy, power of 0.8, and a statistical significance level of 0.05). Three assumptions go into these calculations. Firstly, while attrition levels should be at 10 percent or lower, we do not yet know how many students cannot be tracked from baseline to endline. We (conservatively) show calculations for a worst-case scenario of 20 percent. Secondly, we do not know 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 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 (see Evans and Yuan, 2022), and they reflect TaRL Africa's minimal expectations for the program's success.

[24] We will explore an alternative specification to estimate β_2^t, which replaces α_r with zone fixed effects. If this approach improves precision, we will prefer this specification instead. Even in this case, we will cluster standard errors at the zone level (cf. de Chaisemartin and Ramirez-Cuellar, 2020).

Figure 8: Test reliability and information across varying ability levels



(a) Mathematics

(b) Literacy

Note: This figure reports on the reliability of ability estimates, across the range of student ability in both subjects. Solid lines show the test information function, as per the study's two-parameter logistic (2PL) item response theory (IRT) models. Dashed lines show the corresponding standard error of measurement.

6.3 Test and item characteristics

This section reports on the psychometric characteristics of the study's student assessments and their items. In summary, we find these measurement properties to be favorable.

To begin, we investigate the tests' overall reliability or the level of noise captured by the instruments. Judging by Cronbach's α, we find acceptable levels of reliability (i.e., low levels of noise) for both subjects (α is above 0.72 for mathematics and above 0.80 for literacy).[25]

Item response theory further allows us to assess the instruments' level of precision at varying levels of student ability. Ideally, we would like the instruments to cover a wide range of student abilities with similarly high levels of precision. However, by design, very high and very low ability levels are measured with more noise. Figure 8 reports the results from our analysis of reliability. As expected, the tests' information is lower at the tails of the distribution and, respectively, the standard error of ability estimates becomes higher in this range. Yet, even two standard deviations below and above the mean, the standard error of measurement remains acceptable: it ranges from approximately 0.3 to 0.6 (which corresponds to reliability levels of 0.91 and 0.64, respectively, at these extreme ability levels). For literacy, we note that the test's precision sharply spikes (to extremely high levels) around students' ability to read words. That is, while this and more advanced levels of ability were very well covered by the instruments, the assessments could have been improved even further had we included additional, easy questions around phonics, phonemic awareness, listening comprehension, and vocabulary.

Finally, for completeness, we also report on the performance of individual test questions. Appendix Tables A.1 and A.2 provide item characteristics, for the two subjects, as per the study's item response theory model. We find acceptable to high levels of discrimination for most items, with a median discrimination parameter of 1.3 for mathematics and 2.1 for literacy. We also observe how the difficulty level of items relates to common theories of child development and lan-

[25] For simplicity, to calculate these overall reliability levels with classical test theory, we impute a zero if the adaptive tests did not administer a too-difficult item to a student. Recall that our overall continuous test scores rely on item response theory, which improves over this approach.

guage acquisition (with those questions related to oral language acquisition being easier than writing or reading, for example). Lastly, the results suggest adding additional test questions not only improves the assessments' content validity (capturing those subdomains of foundational skills not covered by the ASER)—this strategy also improves the assessments' reliability and covers a wider range of student abilities with more fine-grained test items.

6.4 Pre-registration and pre-analysis plan

Following best practices for experimental research, we will register this study in a public trial registry (the AEA Trial Registry). This trial registration will include a list of our hypotheses and statistical tests we will conduct (a "pre-analysis plan"). Appendix Table A.3 provides this list.[26] In lieu of a separate plan, we will upload the present document. We may update the plan after finalizing the study's process monitoring strategy. We also consider submitting the study to the Journal of Development Economics before we know its results (on the journal's "pre-results review track").

[26] The study will control the false discovery rate (FDR). Multiple hypothesis testing and advancements over "basic" FDR methods (such as Storey's q) are an active area of research; we will choose a "modern" method such as Boca and Leek's FDR regression (see Korthauer et al., 2019).

Appendix A Additional figures and tables

Table A.1: Item characteristics: Mathematics

	Test	Mapping			Item parameters		Proportion correct
		Content domain	Cognitive domain	Grade level	Discr.	Diff.	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Q7004	non-ASER	Geometric shapes and measures	Applied	1	0.56	-6.2	0.97
Q9	non-ASER	Geometric shapes and measures	Procedural	1	0.82	-3.67	0.94
Q41186	non-ASER	Data display	Procedural	2	0.8	-2.16	0.82
1-digit numbers	ASER	Number sense	Procedural	1	2.52	-2.14	
Q1110	non-ASER	Arithmetic	Applied	1	0.86	-2.12	0.83
Q7005	non-ASER	Data display	Procedural	1	0.51	-1.35	0.66
Q1126	non-ASER	Geometric shapes and measures	Procedural	1	0.42	-1.25	0.62
2-digit numbers	ASER	Number sense	Procedural	1	2.21	-0.42	
Q22	non-ASER	Arithmetic	Applied	2	1.23	-0.08	0.52
Q7026	non-ASER	Number sense	Applied	1	0.87	-0.07	0.51
Q1106	non-ASER	Arithmetic	Procedural	2	2.03	0.07	0.48
1 addition item	ASER	Arithmetic	Procedural	2	1.52	0.68	
Q40	non-ASER	Number sense	Applied	1	0.74	1.02	0.34
Q7027	non-ASER	Number sense	Procedural	1	1.7	1.06	0.23
1 subtraction item	ASER	Arithmetic	Procedural	2	2.3	1.28	
Q7021	non-ASER	Arithmetic	Applied	2	1.49	1.43	0.17
3-digit numbers	ASER	Number sense	Procedural	2	1.57	1.49	0.16
2 addition items	ASER	Arithmetic	Procedural	2	2.53	1.54	
Q8	non-ASER	Number sense	Procedural	2	1.32	1.6	0.16
2 subtraction items	ASER	Arithmetic	Procedural	2	2.87	1.63	
Q41	non-ASER	Number sense	Applied	2	1.09	1.82	0.16
Q3024	non-ASER	Arithmetic	Applied	2	0.79	1.83	0.22
Q7006	non-ASER	Data display	Applied	2	0.86	1.84	0.20
1 multiplication item	ASER	Arithmetic	Procedural	3	1.61	1.9	
1 division item	ASER	Arithmetic	Procedural	3	2.24	2.21	
2 multiplication items	ASER	Arithmetic	Procedural	3	2.41	2.23	
Q3036	non-ASER	Geometric shapes and measures	Procedural	3	0.39	2.57	0.28
2 division items	ASER	Arithmetic	Procedural	3	3.66	2.64	
Q3032	non-ASER	Geometric shapes and measures	Applied	2	0.29	2.84	0.31
Q7030	non-ASER	Data display	Applied	2	0.33	3.54	0.24

Notes. This table provides item characteristics as per a two-parameter logistic (2PL) item response theory (IRT) model (Columns 5 and 6). Items are sorted in ascending difficulty. Item names refer to study-internal question IDs. For reference, the table also provides whether the item comes from the ASER test or not (Column 1). It also shows each item's mapping to content domains (Column 2), cognitive domains (Column 3), and to curricular grade-level expectations (Column 4). The average proportion of correct answers during the baseline (Column 7) is missing if an item was not administered to all students (due to the adaptive nature of the tests).

Table A.2: Item characteristics: Literacy

	Test	Mapping		Item parameters		Proportion correct
		Subdomain	Grade level	Discr.	Diff.	
	(1)	(2)	(3)	(4)	(5)	(6)
Q10	non-ASER	Vocabulary	1	0.41	-3.3	0.78
Q12	non-ASER	Comprehension (listening)	2	0.72	-3.11	0.88
Q11	non-ASER	Vocabulary	1	0.58	-2.15	0.76
Q13	non-ASER	Comprehension (listening)	2	0.57	-1.81	0.72
Q1	non-ASER	Phonemic Awareness	1	0.59	-1.29	0.67
Q2	non-ASER	Phonemic Awareness	1	0.69	-1.12	0.66
Read 1 letter, more	ASER	Reading (letters, words)	1	3.2	-0.79	
Q6	non-ASER	Phonics	1	0.43	-0.48	0.55
Q4	non-ASER	Phonemic Awareness	1	0.37	-0.42	0.54
Read 2 letters, more	ASER	Reading (letters, words)	1	5.2	-0.31	
Q9	non-ASER	Vocabulary	1	0.61	-0.24	0.53
Read 3 letters, more	ASER	Reading (letters, words)	1	7.3	-0.06	
Read 1 word, more	ASER	Reading (letters, words)	1	5.09	0	
Q5	non-ASER	Phonics	1	0.43	0.04	0.49
Read 2 words, more	ASER	Reading (letters, words)	1	11.16	0.12	
Read 4 letters, more	ASER	Reading (letters, words)	1	5.82	0.14	
Q7	non-ASER	Phonics	1	0.77	0.19	0.46
Read 3 words, more	ASER	Reading (letters, words)	1	11.3	0.25	
Q15	non-ASER	Writing	1	1.52	0.3	0.41
Q3	non-ASER	Phonemic Awareness	1	0.41	0.35	0.46
Q8	non-ASER	Phonics	1	0.81	0.37	0.43
Q14	non-ASER	Writing	1	2	0.39	0.37
Read para (haltingly)	ASER	Reading (fluency)	2	5.11	0.45	0.32
Read all 5 letters	ASER	Reading (letters, words)	1	4.04	0.5	
Q16	non-ASER	Writing	1	2.08	0.54	0.33
Q17 (haltingly)	non-ASER	Reading (fluency)	2	11.34	0.61	
Q18	non-ASER	Comprehension (reading)	3	4.75	0.86	
Read story (fluently)	ASER	Reading (fluency)	2	3.3	0.92	
Read para (fluently)	ASER	Reading (fluency)	2	4.23	1.26	0.13
Q17 (fluently)	non-ASER	Reading (fluency)	2	9.22	1.4	
Q19	non-ASER	Comprehension (reading)	3	5.34	1.4	

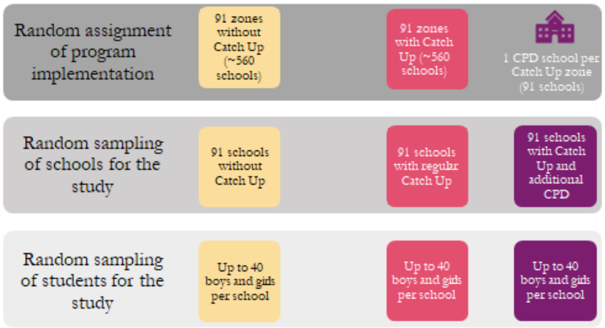
Notes. This table provides item characteristics as per a two-parameter logistic (2PL) item response theory (IRT) model (Columns 4 and 5). Items are sorted in ascending difficulty. Item names refer to study-internal question IDs. For reference, the table also provides whether the item comes from the ASER test or not (Column 1). It also shows each item's mapping to subdomains (Column 2), and to curricular grade-level expectations (Column 3). The average proportion of correct answers during the baseline (Column 6) is missing if an item was not administered to all students (due to the adaptive nature of the tests).

Table A.3: List of tests to be conducted

Test #	Sample	Treatment	Outcome
		Panel A: Main hypotheses	
Test #	**Sample**	**Treatment**	**Outcome**
1	Overall	Regular Catch Up	Overall score, math
2	Overall	Regular Catch Up	Overall score, literacy
3	Overall	Catch Up with CPD	Overall score, math
4	Overall	Catch Up with CPD	Overall score, literacy
5	Overall	Regular Catch Up	Score strongly reflecting the program's focus, math
6	Overall	Regular Catch Up	Score weakly reflecting the program's focus, math
7	Overall	Catch Up with CPD	Score strongly reflecting the program's focus, math
8	Overall	Catch Up with CPD	Score weakly reflecting the program's focus, math
		Panel B: Subgroup analyses	
Test #	**Sample**	**Treatment**	**Outcome**
9	Weakest quartile	Catch Up with CPD	Overall score, math
10	Weakest quartile	Catch Up with CPD	Overall score, literacy
11	Subgroup: Girls	Catch Up with CPD	Overall score, math
12	Subgroup: Girls	Catch Up with CPD	Overall score, literacy
13	Strongest quartile	Catch Up with CPD	Overall score, math
14	Strongest quartile	Catch Up with CPD	Overall score, literacy
15	Subgroup: Boys	Catch Up with CPD	Overall score, math
16	Subgroup: Boys	Catch Up with CPD	Overall score, literacy
		Panel C: Exploratory analyses	
Test #	**Sample**	**Treatment**	**Outcome**
17-18	Overall	Catch Up with CPD	By grade-level, math
19-20	Overall	Catch Up with CPD	By grade-level, literacy
21-22	Overall	Catch Up with CPD	By cognitive domain, math
		Panel D: Supplemental analyses	
Test #	**Sample**	**Treatment**	**Outcome**
23	Overall	Difference in treatment effects	Overall score, math
24	Overall	Difference in treatment effects	Overall score, literacy
25	Overall	Difference in treatment effects	Score strongly reflecting the program's focus, math
26	Overall	Difference in treatment effects	Score weakly reflecting the program's focus, math

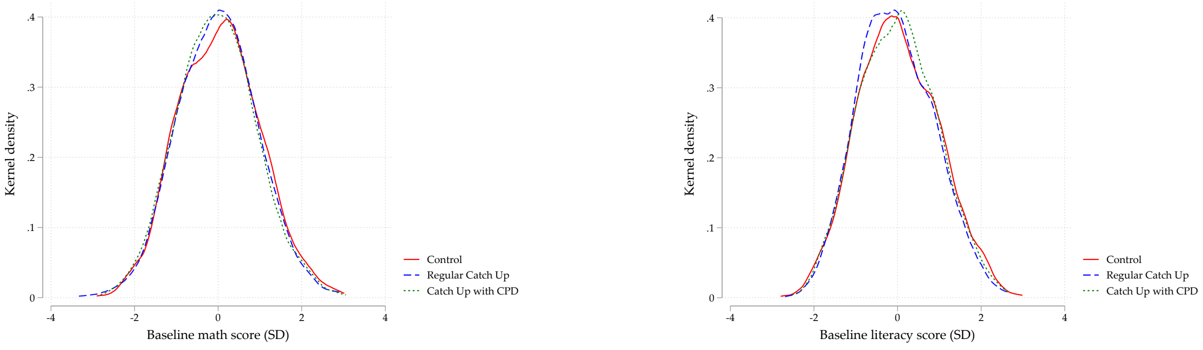
Notes. This table pre-specifies the study's tests and their order. The study will control the false discovery rate (FDR). Multiple hypothesis testing and advancements over "basic" FDR methods (such as Storey's q) are an active area of research; we will choose a "modern" method such as Boca and Leek's FDR regression (see Korthauer et al., 2019).

Figure A.1: Sampling and randomization procedure


Random assignment of program implementation
91 zones without Catch Up (~560 schools)
91 zones with Catch Up (~560 schools)
1 CPD school per Catch Up zone (91 schools)
Random sampling of schools for the study
91 schools without Catch Up
91 schools with regular Catch Up
91 schools with Catch Up and additional CPD
Random sampling of students for the study
Up to 40 boys and girls per school
Up to 40 boys and girls per school
Up to 40 boys and girls per school

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 A.2: 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.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2022, October). When Should You Adjust Standard Errors for Clustering? *The Quarterly Journal of Economics*, qjac038.

ASER Centre (2023). Annual Status of Education Report (Rural) 2022. Provisional Report, ASER Centre, New Delhi.

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg (2020, April). A Theory of Experimenters: Robustness, Randomization, and Balance. *American Economic Review 110*(4), 1206–1230.

Bruhn, M. and D. McKenzie (2009, October). In Pursuit of Balance: Randomization in Practice in Development Field Experiments. *American Economic Journal: Applied Economics 1*(4), 200–232.

de Barros, A., J. Fajardo-González, P. Glewwe, and A. Sankar (2022, December). Large-scale efforts to improve teaching and child learning: Experimental evidence from India. *Journal of Development Economics (conditionally accepted via pre-results review)*.

de Chaisemartin, C. and J. Ramirez-Cuellar (2020, July). At What Level Should One Cluster Standard Errors in Paired and Small-Strata Experiments? Working Paper 27609, National Bureau of Economic Research, Cambridge, MA.

Evans, D. K. and F. Yuan (2022, April). How Big Are Effect Sizes in International Education Studies? *Educational Evaluation and Policy Analysis*, 01623737221079646.

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

Korthauer, K., P. K. Kimes, C. Duvallet, A. Reyes, A. Subramanian, M. Teng, C. Shukla, E. J. Alm, and S. C. Hicks (2019, June). A practical guide to methods controlling false discoveries in computational biology. *Genome Biology 20*(1), 118.

Penney, J. (2023). Cautions when normalizing the dependent variable in a regression as a z-score. *Economic Inquiry 61*(2), 402–412.

Rodriguez-Segura, D. (2022, September). A closer look at reading comprehension: Experimental evidence from Guatemala. *International Journal of Educational Development 93*, 102630.