LiftEd Transforming Foundational Literacy and Numeracy

EDTECH ACCELERATOR

LiftEd EdTech Accelerator

Impact on Learning Outcomes Study

Think Zone®

Odisha | 2026

Principal Investigator

Prof. Tarun Jain

Professor of Economics, Indian Institute of Management Ahmedabad

This study was conducted by Educational Initiatives under the guidance of
Tarun Jain - Professor of Economics and Reserve Bank of India Chair in Finance & Economics at the Indian Institute of Management Ahmedabad.

Table of Contents

Acknowledgements

We express our profound appreciation to those instrumental in conducting this study and compiling the report. Their support has been indispensable to this endeavour.

We extend our gratitude to the Principal Investigator, Professor Tarun Jain, Associate Professor of Economics and the Reserve Bank of India Chair in Finance and Economics at the Indian Institute of Management Ahmedabad. His guidance in designing the study, overseeing the creation of robust data collection tools and ensuring the quality of insights has been pivotal.

We express our gratitude to the LiftEd EdTech Accelerator consortium partners - Michael & Susan Dell Foundation, Reliance Foundation, UBS Optimus Foundation and British Asian Trust for their flexibility to innovate and generous support throughout the project. We also extend our gratitude to acknowledge the support and counsel from Central Square Foundation.

We extend our deepest gratitude to the team at Educational Initiatives for their expertise in conducting the study, generating insights, and report writing. We are especially grateful to Sridhar Rajagopalan (Co-Founder), Pranav Kothari (CEO), Gayatri Vaidya (Associate Vice President), Arvind Singh (Associate Vice President), and Ei's Product team led by Sangeeta Oak (Associate Vice President) for their leadership and guidance. We would also like to acknowledge Dipak Chalakkal, Vaishali Tiwari, Prashant Kapoor, and Aparna Sanjay, along with the field Operations team and enumerators who conducted the assessments, whose dedication and efforts were instrumental in making this study a reality.

Finally and most importantly, we are indebted to the participants of the study for their time and valuable contributions and the EdTech portfolio partner: ThinkZone for their support in facilitating the planning and coordination of the study.

We hope the study findings will inform stakeholders' decisions on designing and implementing effective EdTech interventions and strengthen the collective evidence on EdTech for foundational learning.

In deep gratitude,
Educational Initiatives

Glossary & Abbreviations

Term	Description
Addition/Subtraction Process	Refers to tasks that assess students' ability to solve multi-digit or column-based addition and subtraction problems, requiring correct sequencing of procedural steps such as carrying over or borrowing. These tasks go beyond fact recall and capture procedural fluency in formal arithmetic methods (e.g., solving 47 + 38 using column addition).
Attrition Buffer	Allowance built into the study's sample design to account for the expected loss of participants over time. As the study was longitudinal and conducted over several months, some students present at baseline were unavailable at endline due to absence or mobility and therefore an attrition buffer was included to preserve statistical power.
Conceptual Skills1	Refers to competencies that require application, reasoning, integration of multiple concepts and problem solving. These skills typically involve multi-step, non-routine, or contextualised tasks and may include spatial, or language-heavy demands, such as solving word problems or applying arithmetic concepts to real-life situations. The tasks from the Numeracy tool classified as conceptual Skills for the purposes of this study are Addition Process (2 digit, column addition), Subtraction Process (2 digit, column addition), Counting in Bundles, Missing Number, Word Problems and Shape Recognition.
Confidence	Refers to the degree of certainty that observed results reflect a true underlying effect rather than chance. It is influenced by sample size, outcome variability, study design and precision of estimates, often reflected through p-values and confidence intervals.
Confidence Intervals (CI)	Represent a statistical range around an estimated effect that conveys the degree of uncertainty in the estimate. They indicate the range within which the true population effect is likely to lie, assuming the underlying model is correct (e.g., an average of estimate of 0.7 year with a 95% confidence interval of 0.6 to 0.9).
Difference-in-Differences (DiD)	Analytical method used to estimate programme impact by comparing changes in outcomes over time between Intervention and Comparison groups. This approach helps account for baseline differences and common time trends affecting both groups.
Intent-to-Treat (ITT)	Impact estimate that captures the effect of being assigned to the intervention, irrespective of actual participation or usage levels. ITT estimates preserve the original group assignment and provide a conservative measure of programme effectiveness under real-world conditions.
Intra-Class Correlation (ICC)	Refers to the degree to which students within the same class or school resemble one another in terms of learning outcomes, reducing the amount of independent information contributed by each student.

1 Ministry of Education, Government of India. (2021). National initiative for proficiency in reading with understanding and numeracy (NIPUN Bharat): Mission guidelines. https://www.education.gov.in/sites/upload_files/mhrd/files/nipun_bharat_guidelines.pdf .

Minimum Detectable Effect Size	The smallest difference between Intervention and Comparison groups that the study is statistically powered to detect, given assumptions about sample size, outcome variability, clustering and significance levels.
Outcome Variability/ Variance	Refers to the amount of natural variation in outcomes across individuals rather than uniform responses. Higher variability in student performance increases uncertainty in estimates and typically requires larger samples to detect programme effects.
Power	Probability that a study will correctly detect a true effect of a specified size if it exists in the population. Studies are commonly designed with power levels of 80 percent or higher to reduce the likelihood of false negatives.
Procedural Skills[2]	Refers to tasks that primarily assess recall, recognition and the execution of well-rehearsed procedures, with minimal reasoning demands. These tasks typically focus on basic number sense and fact recall, involve limited interpretation or transfer and are relatively language-light. For the purposes of this study, the following tasks in the numeracy tool are classified as Procedural Skills: Number Comparison, Addition Facts (untimed), Addition Facts (timed), Subtraction Facts (untimed) and Subtraction Facts (timed).
Quasi-Experimental Design (QED)	Study design used to estimate the effects of an intervention when random assignment of participants to treatment and comparison groups is not feasible. Instead, the design relies on non-randomly formed groups and applies statistical or design-based techniques (such as matching or difference-in-differences) to control for pre-existing differences and approximate causal inference.
Standard Deviation (SD)	Statistical measure of variability that indicates how spread out individual observations are around the mean. When used to express effect sizes, SD units standardise differences across groups or time points, allowing impacts to be compared across outcomes measured on different scales.
Treatment-on-the-Treated (ToT)	Impact estimate that reflects the effect of the intervention on participants who actually received or engaged with it as intended. ToT estimates adjust for differential usage or compliance and are typically derived using assignment to treatment as an instrument for actual participation.

2 Ministry of Education, Government of India. (2021). National initiative for proficiency in reading with understanding and numeracy (NIPUN Bharat): Mission guidelines. https://www.education.gov.in/sites/upload_files/mhrd/files/nipun_bharat_guidelines.pdf .

1. Executive Summary

This chapter outlines the purpose, approach and key findings from the independent evaluation of the ThinkZone EdTech solution in Odisha, examining its effectiveness in improving foundational numeracy outcomes.

1.1 About the Study

Foundational Literacy and Numeracy (FLN) underpin lifelong learning in India, yet significant learning gaps emerge early and persist despite recent strides under NEP 2020, NIPUN Bharat Mission and increased FLN investments. Inclusive EdTech presents a strong opportunity to strengthen FLN at home and in schools, supported by growing smartphone access. However, most existing solutions remain poorly aligned with the languages, needs and contexts of low-income communities.

The LiftEd EdTech Accelerator was created to bridge this gap by supporting contextually relevant, pedagogically sound EdTech for "Low-Income Bharat." Working with governments and ecosystem partners, it enables large-scale adoption through funding, mentorship and technical assistance for eight leading EdTech organisations. A robust evaluation agenda focused on learning outcomes, impact and user experience has generated actionable insights. Collectively, LiftEd-supported solutions have reached over 5.5 million users, directly benefiting more than 3 million children in improving foundational learning.

To build evidence on EdTech's effectiveness in FLN, the Accelerator commissioned an independent evaluation to identify what works, its impact on learning outcomes and key enablers for scale. Conducted by Educational Initiatives (Ei) with Central Square Foundation, the study assessed selected solutions' impact on early mathematics across grades and geographies. As the outcome evaluator, Ei consolidates findings at the partner level, measuring the impact on learning outcomes as a result of the solution usage.

The **EdTech solution evaluated in Odisha is ThinkZone**, which supports foundational learning at home by engaging students in weekly Math and Language activities (Grades 1–5, Odia). Parents receive support through calls, SMS, WhatsApp and teacher groups, along with additional help via a helpline, volunteer calls and regular feedback.

1.2 Study Design

The study assessed the impact of ThinkZone's **at-home FLN intervention** on numeracy outcomes for Grades 2 and 3 students in Puri district in Odisha. It used a **longitudinal quasi-experimental design with a difference-in-differences approach**[3] to compare learning growth between intervention and comparison groups over time, with the resulting impact estimates subsequently converted into Equivalent Years of Schooling (EYOS) to aid interpretation. The Baseline round was conducted in September - October 2024 and Endline round in February - March 2025 with ~2,900 students in each round.

3 The DiD effect size is calculated as: [Avg Delta_Intervention (Δi) - Avg. Delta_Comparison (Δc)] / Pooled Standard Deviation]

Assessments were administered through the Ei NEEV application, capturing student responses via audio and touch-based inputs.

1.3 Key Findings

1.3.1 ThinkZone intervention demonstrated strong gains in foundational numeracy, but shifts in conceptual skills were limited

Students demonstrated meaningful growth in procedural numeracy tasks, most notably in Number Comparison (0.17 SD) and Subtraction Facts (0.16 SD). These gains suggest improved number sense and arithmetic fluency, which serve as essential building blocks for early numeracy development. However, improvements were limited in conceptual tasks such as Word Problems (0.02 SD), highlighting ongoing challenges in fostering deeper conceptual understanding and reasoning within the available intervention window.

The EYOS analysis indicates meaningful learning acceleration of ~1.5 times relative to Business-As-Usual instruction. The intervention generated learning gains equivalent to a substantial fraction of a year of schooling over the study period, demonstrating that students progressed faster than peers in comparison schools. These findings suggest that the model not only improves absolute learning levels but also meaningfully compresses the time required to achieve expected grade-level competencies, highlighting its potential to address foundational learning gaps in low-performing contexts.

1.3.2 Grade-level trends showed large gains within the assessment period in Grade 2 and steadier improvements in Grade 3

The programme produced varied effects across grades. Grade 2 students experienced sharp improvements in basic operations (0.20-0.26 SD), though these were accompanied by null effects in some conceptual tasks (Word Problems: 0.02 SD). In contrast, Grade 3 students recorded steadier and more balanced gains across all assessed skills (0.15–0.18 SD). Overall, the programme appears to be particularly effective at accelerating early-stage numeracy acquisition, while more advanced conceptual or applied skill development (Addition/Subtraction Process, Word Problems) progressed at a slower, but more consistent, pace.

1.3.3 Girls showed steadier numeracy gains, while boys exhibited more variable patterns of progress

Gender-disaggregated results highlight a more mixed response among boys—showing gains in basic operations but negligible or negative shifts in tasks that demand conceptual or visual processing (Counting in Bundles, Word Problems, Shape

Recognition). Girls, on the other hand, displayed more uniform progress across all assessed areas with notable improvements in Missing Number, Addition Facts and Subtraction Process. Girls therefore, demonstrated a more consistent pattern of growth in numeracy throughout the intervention.

1.3.4 Foundational gains outpaced progress in conceptual tasks, with limited mobility among lowest performers

Students showed clear improvement in basic numeracy, with many low-performing learners moving from L1[4] into mid-level proficiency bands. Overall, the Intervention group showed greater upward movement, particularly among mid-level students (L2–L3), who demonstrated more dynamic progression across core arithmetic skills. In contrast, the lowest-performing students (L0) showed limited movement though consistently fewer Intervention students remained at L0 compared to the Comparison group. Students in the highest band (L4) remained largely stable across both groups, reflecting ceiling effects.

The ThinkZone intervention demonstrates clear promise in strengthening foundational numeracy skills, especially among younger learners and girls. At the same time, the limited progress in conceptual competencies and uneven gains across grades and genders underline the need for deeper instructional strengthening. Future iterations may benefit from greater emphasis on strengthening understanding, supporting progression beyond foundational fluency, and examining how gains can be more evenly distributed across learner groups.

4 Categories L0–L4 classify student performance from zero scores to highest proficiency, based on accuracy percentages and fluency relative to the baseline average. These levels help track patterns of improvement or decline across the evaluation timeline. L0 = Zero scorers, L1 = 0 to 25% (accuracy tasks) and 0 to 0.25x Baseline Avg (fluency tasks). Further details on the classification are presented in Section 4.4.

2. Background of the Study

2.1 Inclusive EdTech for Foundational Learning

Foundational Literacy and Numeracy (FLN) - the ability to read, write, comprehend and perform basic mathematical operations by the end of Grade 3 - is widely recognised as the cornerstone of lifelong learning.[5,6,7,8] In India, significant learning deficits emerge early, particularly among children aged 4 to 8, with large proportions of students struggling to acquire basic skills. Repeated rounds of the Annual Status of Education Report (ASER) over the past decade have consistently highlighted low levels of foundational learning. These early gaps make it increasingly difficult for children to grasp more complex concepts as they move through higher grades. The most recent ASER data, however, points to green shoots of progress - improvements in basic reading levels for Grade 3 government students being the highest since 2005 with basic arithmetic levels also showing substantial improvement.[9] Yet, the depth and persistence of learning deficits across the system make clear that far more remains to be done.

To address these challenges, the Government of India has launched multiple initiatives to strengthen FLN, including NEP 2020, which frames FLN as a five-year continuum and emphasises curriculum reform, technology integration and teacher capacity-building. NIPUN Bharat (2021) sets FLN targets for preschool to Grade 3 by 2026–27, supported by initiatives such as the CBSE Reading Mission, the National Curriculum Framework for the Foundational Stage and Jaadui Pitaara, which together provide age-appropriate, multilingual learning resources. Progress is monitored through large-scale assessments like NAS and FLS, while increased funding under Samagra Shiksha and state-level programmes, often implemented with private and civil society partners, aim to further bridge learning gaps[10,11] Further, budget allocations for education under Samagra Shiksha have increased, from ₹31,050 crores (~4200 million USD) in FY 2021-22 to ₹41,250 crore in FY 2025-26.[12] Concurrent to the central government initiatives, state governments, in collaboration with private partners and civil society organisations, have implemented tailored initiatives to bridge learning gaps.

In this context, educational technology (EdTech) has emerged as a potential multiplier, harnessing digital tools to support teaching and learning both in classrooms and at home. A recurring question, however, is whether EdTech is truly accessible to all, especially children from low-income households. Emerging data suggests that access to digital

5 Ministry of Human Resource Development, Government of India. (2020). *National Education Policy 2020*. https://www.education.gov.in/sites/upload_files/mhrd/files/NEP_Final_English_0.pdf
6 Institute for Competitiveness. (n.d.). *State of Foundational Literacy and Numeracy in India*. https://www.competitiveness.in/wp-content/uploads/2021/12/Report_on_state_of_foundational_learning_and_numeracy_web_version.pdf
7 Ministry of Education, Government of India. (n.d.). *About Foundation Literacy and Numeracy*. https://diksha.gov.in/fln.html.
8 Sinha, A (2023). Maximising India's demographic dividend through foundational literacy and numeracy. *Hindustan Times*. (https://www.hindustantimes.com/ht-insight/knowledge/maximising-india-s-demographic-dividend-through-foundational-literacy-and-numeracy-101699332886168.html).
9 Annual Status of Education Report (ASER). (2024). *ASER 2024 National Findings* https://asercentre.org/wp-content/uploads/2022/12/ASER-2024-National-findings.pdf
10 Ministry of Human Resource Development, Government of India. (2020). *National Education Policy 2020*. https://www.education.gov.in/sites/upload_files/mhrd/files/NEP_Final_English_0.pdf
11 Storyweaver (n.d.). *CBSE Reading Mission*.https://storyweaver.org.in/en/about/campaigns/cbse-reading-mission
12 CNBC TV 18. (2025). Budget 2025: *National Education Mission receives outlay of ₹41,250 crore* https://www.cnbctv18.com/budget/budget-2025-national-education-mission-samagra-shiksha-abhiyaan-receives-outlay-of-rs-41250-crore-19548980.htm

infrastructure in these contexts is becoming increasingly feasible, particularly through smartphones. Studies indicate that **90% of households** have access to at least one smartphone and in **75% of these households**, children regularly use the device - typically spending around an hour on it each day.[13] This rise in smartphone penetration, coupled with growing internet availability, presents a significant opportunity to leverage EdTech solutions to improve learning outcomes and narrow educational gaps.

Building on this opportunity, EdTech offers innovative solutions to address critical educational challenges, such as varying teacher quality, diverse learning levels within classrooms and limited access to quality instructional resources. By equipping teachers with tools for effective pedagogy and enabling parents to support their children through interactive content and progress tracking, well-designed, pedagogically sound EdTech solutions have the potential to significantly improve learning outcomes.

Globally, there is emerging evidence on the potential of EdTech to support learning at home. The Global Learning XPRIZE Competition, launched in 2014, incentivised teams from around the world to create open-sourced, scalable software that empowers children to achieve foundational learning skills and saw learning gains for both literacy and numeracy across competing solutions.[14] Similarly, Angrist, Bergman and Matsheng provide experimental evidence on strategies to support learning when schools close.[15] Using a randomised control design, they tested two low-technology interventions in Botswana – SMS messages and phone calls – with parents to support their child's learning and found that combined treatment improves learning by 0.12 standard deviations. This translates to 0.89 standard deviations of learning per USD 100, ranking among the most cost-effective interventions to improve learning.

Despite this promise, most current EdTech solutions in India are designed primarily for middle- and high-income users. Content is often in English, misaligned with the lived realities and languages of children from low-income communities, and offered at price points that place it beyond their reach. As a result, the children who could benefit most from effective EdTech are often the least likely to access it. Unlocking EdTech's transformative potential for "Low Income Bharat" therefore requires inclusive, affordable solutions that are contextually relevant, explicitly address foundational learning and are backed by rigorous evidence of effectiveness. Bridging these gaps will ensure that EdTech becomes a critical lever for equitable and impactful education in India. It was with this objective that the **LiftEd EdTech Accelerator** was set up.

2.2 LiftEd EdTech Accelerator

To bridge the gaps as defined above and to leverage the opportunity that India has, a consortium of non-profit and philanthropic organisations have set up a LiftEd EdTech Accelerator, a two-year initiative from April 2023-25, to support foundational learning of children using EdTech. The Accelerator was set up to support the NIPUN Bharat mission to significantly shape the future of tech-based learning at home for foundational literacy and numeracy in India by reaching 2.5 million children by 2025.

13 Central Square Foundation. (2026). *Bharat Survey for EdTech (BaSE) Report 2026.* Bharat Survey for EdTech (BaSE) Report 2025
14 Global Learning X Prize. (n.d.). *Global Learning X Prize: Executive Summary* https://assets-us-01.kc-usercontent.com/5cb25086-82d2-4c89-94f0-8450813a0fd3/fc467c7f-d8bd-4d05-bba3-2aa9b06833fb/GLEXP_Executive%20Summary.pdf
15 Angrist, N., Bergman, P. & Matsheng, M. (2022). Experimental evidence on learning using low-tech when school is out. *Nature Human Behaviour*, 6, 941-950. https://doi.org/10.1038/s41562-022-01381-z

The LiftEd EdTech Accelerator is anchored by Michael & Susan Dell Foundation, Reliance Foundation and UBS Optimus Foundation as Founding Partners, the British Asian Trust as the Programme Leader and Central Square Foundation as the Design and Technical Partner.

The Accelerator aimed to catalyse the supply of contextually relevant and pedagogically sound learning solutions, generate compelling evidence on their efficacy, work with governments to enhance the efficacy of EdTech adoption and create public goods to address systemic challenges in the ecosystem.

The Accelerator aimed to support eight high-quality EdTech solutions for two years through impact-focused grant funding, dedicated mentorship and capacity-building support to unlock the full potential of the EdTech solutions. The solutions were onboarded into three cohorts, each addressing key challenges in the Indian EdTech ecosystem. The cohorts focused on:

1. Scale – products looking to discover and unlock new pathways to scale - ThinkZone
2. Engagement – products seeking strategies to deepen engagement with the users - Chimple, Ei Mindspark, Pratham, Rocket Learning, Top Parent
3. Product Contextualisation – products developing pedagogically sound and contextually relevant solutions specifically for low-income India - Amira Learning and Sesame Workshop India (SWI)

On the demand side, the Accelerator focused on driving the adoption and institutionalisation of tech based home learning for FLN within State Governments, while also exploring innovative pathways for EdTech integration through partnerships with retail channels, such as gig economy organisations and self-help groups (SHGs).

To tackle the challenge of limited existing evidence on 'what works' in EdTech and to allow for ongoing innovation and progress, the Accelerator's evidence generation agenda included

1. Learning Outcomes Evaluation - to assess the impact on student learning outcomes for ThinkZone, Top Parent and Ei Mindspark.
2. Impact of Acceleration Study - to capture the effectiveness of the strategies implemented within the Accelerator (published).
3. Insights on User Experience Study - a qualitative analysis that gathers feedback from end users on key aspects of the EdTech programme lifecycle, including acquisition, onboarding, engagement and retention.

These evaluations have been conducted under the supervision of the Principal Investigator, Prof. Tarun Jain (Reserve Bank of India Chair Professor of Economics at Indian Institute of Management, Ahmedabad) by experts from Sambodhi Research (qualitative study) and Educational Initiatives (quantitative study) to provide actionable insights to inform future interventions and improvements.

Over 2023-2025, the LiftEd EdTech Accelerator has collaborated with eight leading EdTech organisations to advance the future of tech-enabled at-home learning for foundational literacy and numeracy (FLN) in India. Collectively, these solutions have reached over **5.5 million users**, with more than **3 million** directly benefiting from features and innovations developed through the Accelerator's support.

This report provides insights from the **Learning Outcomes Evaluation Study** for **ThinkZone (TZ)**.

2.3 Overview of EdTech Partner Model

ThinkZone (TZ) seeks to empower parents and educators to improve foundational skills of children in the ages of 3-10 years. The organisation has been running a program to upskill youth volunteers to buttress their teaching capacities to better support improvement in FLN skills of children from low-income communities in Odisha. During the pandemic, the need to explore and enhance at-home learning became increasingly evident. As a result, the home learning program, PRAKASHAK (Parent's Remote Assistance and Knowledge Support for Holistic Advancement of Kids) was developed. The program is rooted in the belief that parental engagement is a key lever for improving learning outcomes of children. There is a corpus of literature that supports this hypothesis – studies underscore the positive association between parental engagement and academic achievement and socio-emotional-cognitive development of the child, particularly in low-income/low-resource contexts.[16]

Under PRAKASHAK, parents are provided with customized learning content and activities covering FLN learning outcomes aligned with the NIPUN Bharat Mission through a variety of delivery channels - messages (SMS), voice calls, Interactive Voice Response System (IVRS) and WhatsApp. The idea is to ensure that parents spend dedicated time with their children every day, going through the content and performing simple learning activities, which can improve their children's FLN skills. Sensitisation messages are also sent to parents, underlining the critical role they play in the continuous learning of their child. Parents and students are onboarded by the youth volunteers, who continuously engage with them through check-in calls.

Under the EdTech Accelerator model, along with building on the existing PRAKASHAK program, a new pathway of running the program in partnership with the government has been introduced. The latter encompasses forging collaborations with the government school ecosystem - teachers and principals in schools and government cadre at the cluster, block and district level (Block education officers (BEOs), officials in Block Resource Centres (BRCs) and Cluster Resource Centres (CRCs). The objective is to catalyse this ecosystem – to ensure buy-in, interest and proactiveness. Specifically, the focus of this relationship is that it translates into a government-promoted program where teachers support in galvanising parental involvement, encouraging them to engage more deeply with their child's education journey. It also involves the government cadre regularly reviewing progress on FLN learning outcomes, deliberating on necessary feed-forward changes and encouraging schools to drive improvements.

16 Lin, Q (2003). Lin, Parent Involvement and Early Literacy / Family Involvement Research Digests / Publications Series / Publications & Resources / HFRP - Harvard Family Research Project. (C) Presidents and Fellows of Harvard College. https://archive.globalfrp.org/publications-resources/publications-series/family-involvement-research-digests/parent-involvement-and-early-literacy. Yan, W., & Lin, Q. (2002, April). Parent involvement and children's achievement: Race and income differences. In annual conference of American Educational Research Association, New Orleans, LA

Outlining the reasons for pivots in ThinkZone's model provides valuable context. First, is the cognisance that to unlock scale, tapping into existing state machinery and ecosystem is critical. Second, at the ThinkZone organisation level, a home learning program centered around youth volunteers is resource intensive (time, talent, money). Lastly, driving parental engagement through a government ecosystem with teachers championing the program is an effective model.

In the current model, ThinkZone collaborates with and catalyses the government school ecosystem (as mentioned above). The ThinkZone team develops guidelines for program implementation. Meetings are held with BEOs / BRCCs / CRCCs in the program blocks to discuss program roll out and support required. The block level officials (and/or ThinkZone staff) conduct meetings with principals and teachers to discuss program details, implementation plan and collation of student database (if available) in their schools of children in Grades 1-3. Subsequently, WhatsApp groups are created for respective schools at the cluster level to facilitate program related communication and to receive a student database (with parents' contact details). Upon receipt of student and parent details, the ThinkZone field team segregates students and parents into WhatsApp and non-WhatsApp users.

Contiguously, the field team coordinates with schools/teachers to map content and activities to school curriculum to personalise learning activities for respective schools School teachers then facilitate the creation of WhatsApp groups with parents and/or add ThinkZone field team to existing groups. Parents are informed about the program by the ThinkZone field team and teachers through parent teacher meetings (PTMs). In these meetings, champion parents and members of school management committees (SMCs) are identified as the point-of-contact. As a next step, follow-up calls and WhatsApp messages are sent to identified champions, to aid dissemination of program information in their communities. Additionally, teachers share program flyers to parents through their children and put up posters in the school. Lastly, parents and teachers receive an automated welcome call, text messages and/or WhatsApp messages (via a chatbot) from ThinkZone.

Parents are provided with a 'primer' on the concept and importance of FLN and the requisite skills, along with visual, image-based content and learning activities aligned with NIPUN Bharat guidelines. Learning activities use simple, easily available household items (for example, using vegetables to help with counting/addition, or using household items to understand the concepts of shapes in geometry). With regards to the parent-facing component, the ThinkZone team, disseminates learning content, worksheets and activities through low-cost, accessible and customised technology – i.e., text messages (for parents without a smartphone), voice calls (pre-recorded descriptive audio versions of the text messages for parents with no/low textual literacy) and WhatsApp (through a WhatsApp chatbot, for parents with a smartphone). Parents with feature phones receive weekly learning activities via text messages and automated calls, along with bi-monthly nudges. Smartphone users access activities through a WhatsApp chatbot, with weekly push notifications and school group updates. All parents are supported through bi-monthly ThinkZone team calls and monthly PTMs.

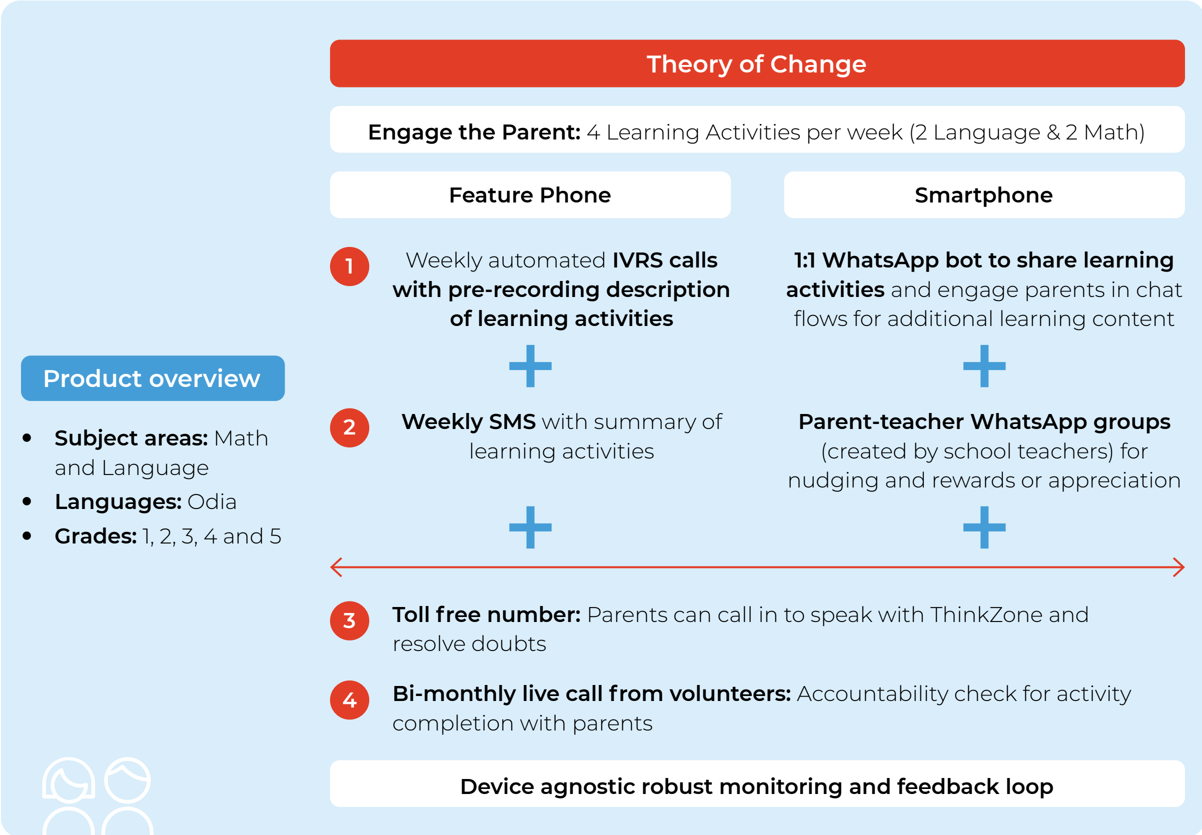


Figure 1: ThinkZone Model

The approach is based on a clear theory of change: engaged parents conducting structured activities at home lead to improved FLN outcomes. By combining government push (for scale through schools) with community pull (through local volunteers), ThinkZone aims to extend reach while ensuring sustained parent participation.

3. Evaluation Design and Approach

This section outlines the study's design, detailing how the quasi-experimental setup, sampling approach and assessment methods were structured to measure ThinkZone's impact on numeracy outcomes. It also summarises the timelines, data collection processes and safeguards used to ensure validity and reliability of findings.

3.1 Study Design

3.1.1 Overview

The study employed a quasi-experimental design (QED), with an Intervention group and a Comparison group identified to benchmark impact. This study seeks to answer the question:

> What impact does the use of the intervention programme have on student learning outcomes?

The Intervention group included students who received structured access to the ThinkZone programme through their teachers. The Comparison group comprised similar students who did not receive the programme during the study period. This setup enabled a clear assessment of the effect of EdTech exposure by comparing progress between the two groups with comparable baseline characteristics. To estimate causal impact, the study used a **difference-in-differences (DiD) approach**, measuring changes in learning outcomes over time while controlling for time-invariant factors.

Students from both Grades 2 and 3 were included in the study, although sampling was not conducted at the grade level. Assessments were carried out in two rounds to measure learning outcomes at the start (baseline) and end (endline) of the evaluation period.

3.1.2 Timelines

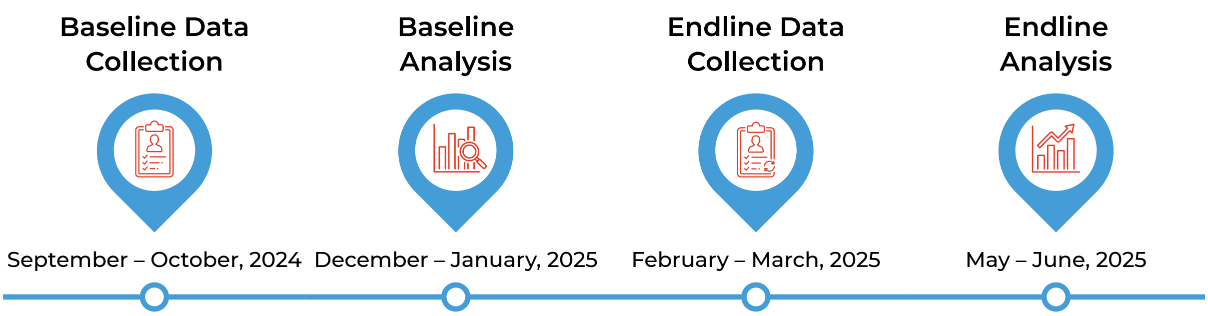


Figure 2: Study Timelines

3.2 Sampling Strategy

3.2.1 Sample Design

The evaluation was conducted in Puri district, Odisha across three different blocks - two for the Intervention group and one for the Comparison group. Since the intervention group was pre-selected, the study's sample size was calculated based on the population of this group.

The following considerations were used for sample size calculation:

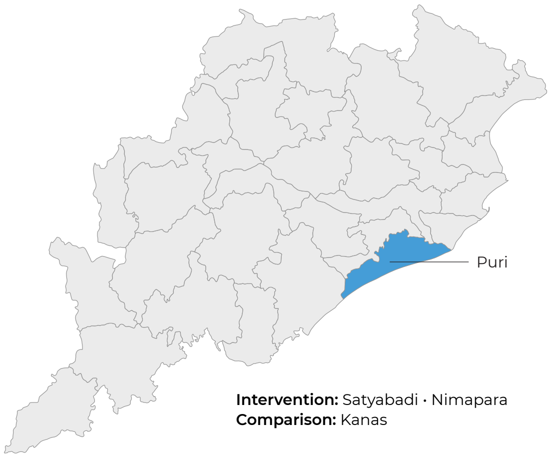
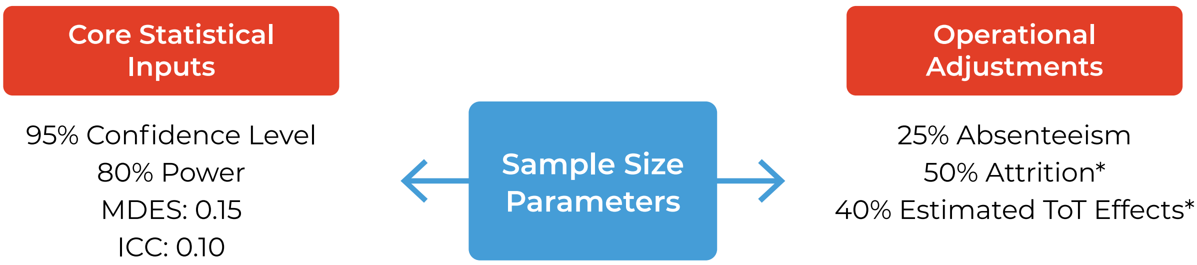


Figure 3: Geography covered

3.2.1 Sample Design

The evaluation was conducted in Puri district, Odisha across three different blocks - two for the Intervention group and one for the Comparison group. Since the intervention group was pre-selected, the study's sample size was calculated based on the population of this group.

The following considerations were used for sample size calculation:



Note: Factors marked with * apply only to the Intervention group.

Figure 4: Sample Size Parameters

Based on the above factors, the final sample size for both the Intervention and Comparison groups was determined as **2,080 students from 139 Intervention schools and 960 students from 64 comparison schools**. Details regarding the sampling process are presented in Annexure 6.1 to 6.3.

3.2.2 Selection of Intervention and Comparison Groups

ThinkZone recruited students for its EdTech Accelerator intervention through a school-based approach, identifying schools with permission to implement the programme and encouraging students to use the FLN solution at home. Intervention group students were sampled from these schools. For the Comparison group, schools from Puri district but outside the intervention blocks were selected using Coarsened Exact Matching (CEM) on socio-economic and demographic factors. Schools from the Kanas block were included as per local permissions.

To address differences in group sizes within schools, many-to-one and one-to-many matching was applied to maximize overlap while keeping an average cluster size of 15 students per school and ensuring logistical feasibility. Initially, 135 schools in Satyabadi with 2,017 students were identified, but 23 schools (340 students) were excluded as they were administered an earlier version of the assessment. These were replaced with 27 schools from Nimapara (403 students).

The matching exercise was repeated twice: once to replace two incorrectly matched secondary schools using UDISE data and again to add comparison schools for 30 Intervention schools (3 from Satyabadi, 27 from Nimapara) identified after the initial round. Further details are presented in Annexure 6.2.

3.2.3 Final Sample Achieved

Since the study was longitudinal, tracking the performance of the same students over the evaluation period, students were excluded from the final dataset if they fell into one of the following categories at the end of data collection:

- Attempting only one subject in a given round due to lack of consent, operational challenges, or other constraints
- Absent or unsynced audio response recordings / assets
- Presence in the Baseline but absence in the Endline

The final sample is as follows:

Table 1: Details of Final Sample

Arm	Sample Size	Minimum Required	Final Sample Achieved
Intervention	2080	343	824
Comparison	960	343	535

Since the minimum required sample size was achieved, no changes were needed to the MDES of the study. A detailed summary of the sampling process is provided in Annexure 6.3.

3.3 Data Collection Methods and Tools

This section outlines the assessment tools, data collection processes and safeguards employed to ensure accuracy and reliability of findings

3.3.1 Assessment Tool

A contextualised version of the **Early Grade Mathematics Assessment (EGMA)** was used for this study. The same assessment tool was administered to students in Grades 2 and 3 and remained unchanged across both the Baseline and Endline rounds. This was feasible given the EGMA-based FLN tool assesses largely common concepts and skills across Grades 2 and 3, with the exception of multiplication and division facts, which are typically introduced in Grade 3. The trade-off of excluding these Grade 3–specific skills is outweighed by the advantages of using a common tool, which enables the calculation of consolidated scores and effect sizes across both grades. This approach allowed the study to examine the impact of the EdTech solution across two grade levels while expanding coverage without increasing sample size or implementation costs. To ensure contextual relevance, the assessment tools were adapted, contextualised and translated into Odia.

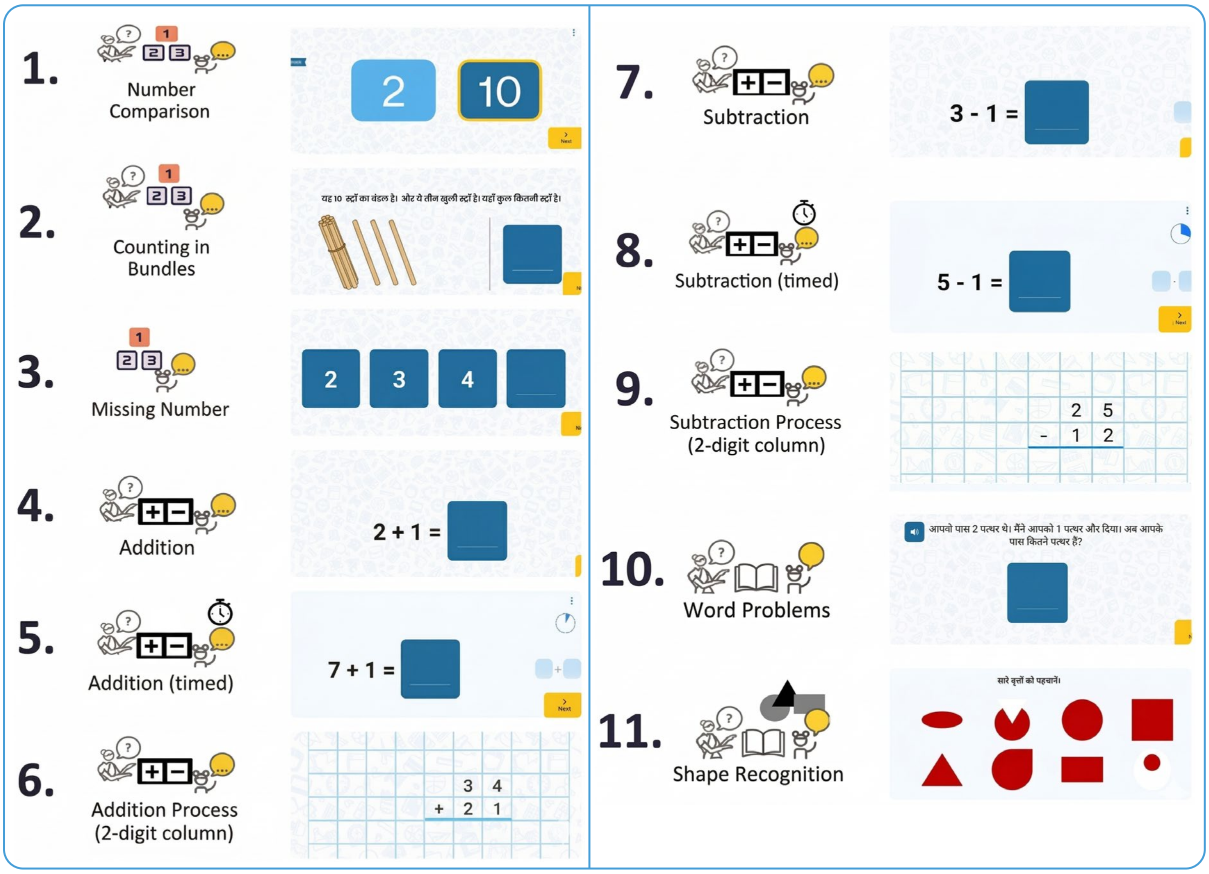


Figure 5: Tasks Assessed

The scaffolded assessments capture foundational skills, ensuring robust measurement of learning outcomes. Further details on the tool are provided in Annexure 6.4.

3.3.2 Data Collection Process

Although ThinkZone's intervention primarily operates as an at-home model, data collection was conducted at the school level. Trained field enumerators visited sampled schools to administer the assessments using tablets. While each assessment was one-on-one in nature - with students responding independently on their own devices - each enumerator simultaneously supervised two children at a time to ensure smooth administration and adherence to protocols. Dedicated enumerator training sessions were conducted prior to each assessment round to standardise procedures and maintain data quality.

Assessments were administered using the Ei NEEV application, installed on tablets provided to students at the start of each session. The application included verbal instructions to guide students through the process. Data was captured as touch-based digital entries, which were auto-scored by the application.

3.3.3 Regional Considerations

The operations plan was designed to account for region-specific factors that could affect student availability and participation. Scheduling of assessments took into consideration public holidays, the local harvest season and the school academic calendar. These adjustments ensured that data collection was minimally disruptive, logistically feasible and allowed for smooth test administration across all sampled schools.

3.3.4 Safeguards for Data Quality

Data collected via the Ei NEEV application was stored in a central database and subsequently transferred to a secure evaluation portal. Tasks were processed using pre-validated scripts that had undergone Quality Analysis testing. For timed tasks, items not reached by students were automatically scored as zero to maintain consistency. Once both Baseline and Endline scoring was completed, further verification checks were conducted, including:

- Analysing scores across competencies[17]
- Checking progression patterns[18]
- Identifying anomalies such as outlier distributions[19]

3.4 Challenges and Implications

The study provides relevant evidence on the intervention's impact, however, the limitations outlined below should be taken into account when considering the findings.

3.4.1 Technical Challenges

- Infrastructure constraints in some schools limited the ability to seat students far apart, occasionally affecting the quietness of the assessment environment. However, verification of the audio files confirmed that the background noise had no impact on validity of the assessments.
- Additionally, during the Baseline, technical issues related to tablet–database synchronization also resulted in the loss of certain audio files. Although recovery efforts were made through system backups, a small portion could not be retrieved. By the Endline, these issues were largely resolved for the purposes of this programme, significantly reducing data loss. Given the large sample size achieved in the study, the minimum required sample size was comfortably exceeded, ensuring no risk to the validity of findings. A detailed breakdown of the final sample determination is included in Annexure 6.3.

3.4.2 Sample Attrition and Participation Gaps

Between the Baseline and Endline rounds, a proportion of students attritioned due to migration, absence from school, or unavailability during the data collection window.

17 Aggregate and task-level scores were reviewed across competency domains to ensure internal consistency and expected variation in performance.
18 Baseline to Endline score changes were examined to verify plausible learning trajectories at the student and group levels.
19 Score distributions were screened to flag extreme values or irregular patterns that could indicate data or scoring issues.

These cases were excluded from the final sample, reducing the overall sample size available for analysis[20].

Lee Bounds analysis was conducted to assess the potential impact of student attrition between baseline and endline. As part of this analysis, students who were present at baseline but missing at endline were assigned a score of zero, thereby generating more conservative estimates of the intervention's impact and ensuring that the reported effects are robust to potential bias arising from differential dropout. Lee Bounds estimates[21] suggest that some of the positive DiD effects reduce in magnitude, lose statistical significance, or change direction once potential attrition bias is accounted for. This indicates that the main results should be interpreted with caution, as differential attrition may influence the estimated impacts.

3.4.3 Operational and Tool Related Challenges

- In a few schools, assessments were completed before it was identified that the tool used Odia numerals, whereas the curriculum uses Hindu–Arabic numerals (1, 2, 3, etc.). The data procured from these schools were excluded from the final analysis. The tool was subsequently revised to align with the curriculum and additional assessments were conducted to meet the required sample size.
- In addition, in certain geographies, delays and constraints in obtaining the necessary permissions from comparison schools limited the ability to conduct assessments as planned. In certain geographies, challenges and delays in securing permissions from comparison schools restricted where assessments could be carried out. Consequently, the comparison group had to be sampled from specific blocks rather than across the full intended geography.

20 Demographic Analysis of the students who dropped out of the study after the Baseline is presented in Annexure 6.5.
21 Lee bounds provide a conservative range of treatment effects by adjusting for potential bias from differential attrition, effectively assuming that all individuals who dropped out would have had the lowest possible outcome (e.g., scored zero).

4. Findings

This section presents the key findings from the evaluation, detailing the overall impact of the ThinkZone programme on numeracy outcomes and how these effects vary by grade and gender. It also examines score movements to provide deeper insight into where the intervention drove progress and where challenges persist.

4.1 Overall Impact

The ThinkZone programme's numeracy evaluation shows meaningful improvements, with higher gains observed in procedural skills. The **strongest effect**s were seen in **Number Comparison** (0.17 SD), indicating strengthened **early numeracy abilities**. A strong grasp of **comparing and judging quantities is widely associated with better performance** in a range of other mathematical tasks[22], underscoring the importance of this gain. **Subtraction Facts** (untimed) also showed notable improvement (0.16 SD), reflecting steady progress in **foundational arithmetic skills**. These results suggest that the programme was effective in reinforcing the basic building blocks on which more complex mathematical competencies rely.

Smaller, **positive gains** appeared in **Addition Facts** (untimed - 0.15 SD, timed - 0.12 SD), **Subtraction Facts** (timed - 0.13 SD), **Subtraction Process** (0.13 SD) **and Missing Number** (0.12 SD), indicating incremental progress in mental arithmetic and problem-solving.

On the other hand, some tasks showed null effects. **Counting in Bundles** (-0.08 SD), **Addition Process** (untimed; 0.06 SD), **Word problems** (0.02 SD) and **Shape Recognition** (-0.07 SD) showed null effect , suggesting no notable improvements in these skills within the evaluation period.

The lower gains in Addition and Subtraction Process (involving column-wise arithmetic operations), as well as Counting in Bundles, are likely because all three require the skills of grouping and regrouping[23].

22 Schneider, M., Beeres, K., Coban, L., Merz, S., Schmidt, S. S., Stricker, J., & De Smedt, B. (2017). Associations of non-symbolic and symbolic numerical magnitude processing with mathematical competence: A meta-analysis. *Developmental Science*, 20(3), e12372. https://doi.org/10.1111/desc.12372
23 Jensen, S., Van Hoof, J., Verschaffel, L., & Van Dooren, W. (2024). Place value and regrouping as helpful constructs to diagnose difficulties in understanding the place value system. *Journal für Mathematik-Didaktik*. https://doi.org/10.1007/s13138-024-00234-8

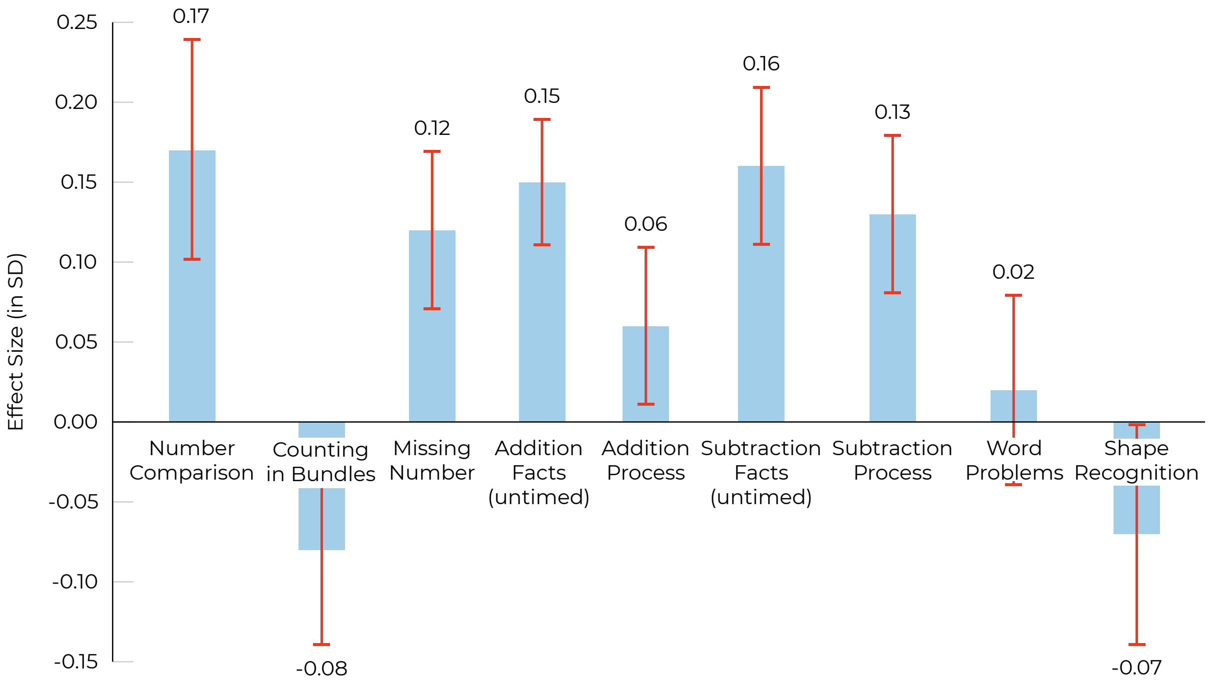


Figure 6: Task wise DiD Effect Sizes (in SD) – Untimed Tasks.

The red lines denote the standard errors.

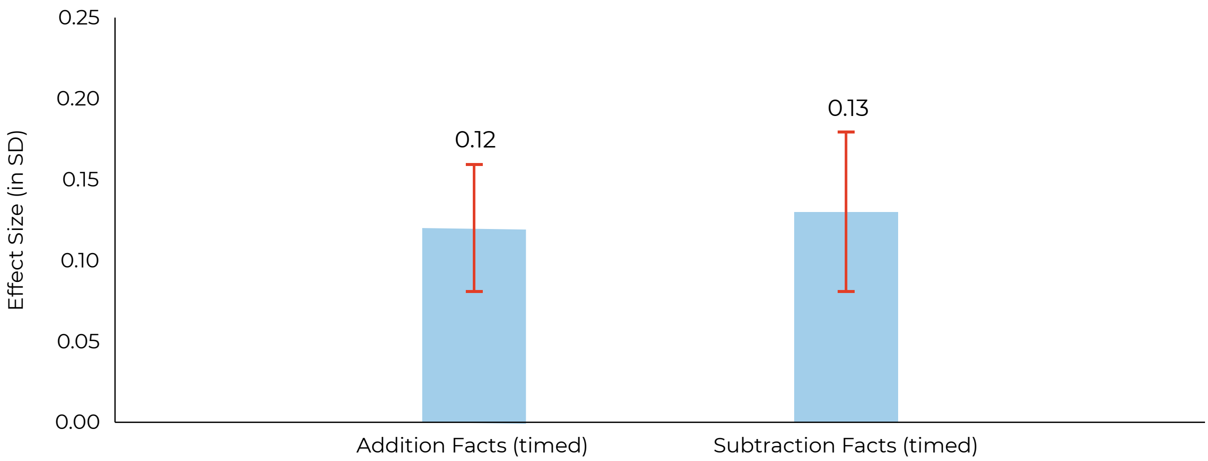


Figure 7: Task wise DiD Effect Sizes (in SD) – Timed Tasks.

The red lines denote the standard errors.

Overall, the findings suggest that the intervention had highest impact on number sense-related skills which form the foundation of early numeracy. However, its effects on other skills such as problem solving, counting strategies, and geometry were weaker, pointing to potential areas for programme strengthening in the future.

4.2 Equivalent Years of Schooling (EYOS)

4.2.1 Methodology

The Equivalent Years of Schooling (EYOS) metric is used to express standardized numeracy learning gains in units of schooling-equivalent years. For Odisha, EYOS is computed using data from the quasi-experimental evaluation, where the control group represents **business-as-usual (BAU)**[24] instruction.

Learning gains are estimated separately for intervention and control groups as the difference between endline and baseline mean numeracy scores. These gains are standardized using a pooled standard deviation to obtain effect sizes at the numeracy-competency level. EYOS is then computed using a **control-as-standard specification**, defined as the ratio of the intervention effect size to the control effect size. This ratio captures learning acceleration relative to the BAU learning trajectory over the same period.

EYOS estimates are aggregated across numeracy competencies using below specifications:

- **Weighted EYOS**, where competencies are aggregated using difference weights based on the average of intervention and control sample sizes across baseline and endline. This specification prioritizes competencies with larger and more reliable samples and is treated as the primary estimate.

Treatment of Extreme Values and Winsorization

Given the heterogeneity in competency-level gains and the presence of auto-scored numeracy tasks, ratio-based EYOS estimates are sensitive to extreme values, particularly when control-group learning gains are small.

To ensure stability and interpretability of the EYOS estimates in Odisha, the following rules are applied:

- **Control effect size floor:** Competency-level observations with very small control effect sizes are excluded from ratio calculations to avoid inflated EYOS ratios driven by near-zero denominators.
- **Upper-tail winsorization:** EYOS ratios are winsorized at an upper cap to limit the influence of extreme values arising from sparse competencies or measurement noise. Values above the cap are set equal to the cap rather than removed.
- **Consistent application:** Floors and caps are applied consistently across weighted and unweighted specifications to ensure comparability.

These thresholds are selected based on inspection of the empirical distribution of effect sizes and documented for transparency. ***For Odisha, EYOS ratios were stabilised by***

24 Angrist, N., Bergman, P., Brewster, C., & Matsheng, M. (2020). Stemming learning loss during the pandemic: A rapid randomized trial of a low-tech intervention in Botswana. Journal of Human Resources, 56(S), S1–S45. https://doi.org/10.3368/jhr.58.S1.0620-12203R1 . Business-As-Usual (BAU) refers to the counterfactual learning trajectory that students would experience in the absence of the intervention, under prevailing instructional practices and system conditions. In impact evaluations, BAU is typically represented by outcomes observed in control or standard-implementation groups and serves as the benchmark against which incremental learning gains from an intervention are measured.

excluding control effect sizes below 0.08 SD and winsorizing ratios above 1.50× to limit instability from sparse or noisy competencies.

Uncertainty around EYOS estimates is quantified using standard errors derived via the delta method, enabling the construction of confidence intervals and hypothesis tests where feasible.

4.2.2 Findings

The EYOS analysis for ThinkZone in Odisha indicates **positive learning gains in numeracy relative to BAU schooling**, though with lower precision than observed in some other contexts.]

Specification	Intervention : Control EYOS Ratio	EYOS Above BAU (Δ years)	Learning Acceleration (%)	Interpretation
Weighted EYOS	1.46×	+0.4–0.5 years	+46%	Preferred estimate; reflects aggregation with sample-size-based difference weights

Under the **weighted specification**, the control-as-standard EYOS ratio is estimated at **1.46**, implying learning progress equivalent to a **46% increase** over the BAU trajectory during the evaluation period. In schooling-equivalent terms, this corresponds to approximately **0.4–0.5 additional years of learning** beyond BAU. While the point estimate indicates meaningful learning acceleration, the associated confidence intervals are wide and the estimate is **not statistically significant at the 95% confidence level**.

The EYOS results for Odisha indicate consistent directional improvements in numeracy outcomes attributable to ThinkZone. These findings should therefore be interpreted as indicative of positive learning progress rather than conclusive evidence of statistically distinguishable gains relative to BAU.

4.3 Impact by Grade

This section examines performance across grades to understand how the intervention affected students at different stages of learning and to highlight grade-specific trends in numeracy skills. Baseline results indicate that Grade 3 students had not yet achieved full proficiency in several Grade 2–level competencies, thereby reducing the risk of ceiling effects and supporting the use of a common assessment framework across grades.

The grade-wise analysis shows a **mixed pattern** - while **both grades benefitted in foundational number skills, Grade 2** recorded greater but uneven gains, whereas **Grade 3 showed steadier yet smaller** improvements. These results indicate that while both grades benefitted, they did so differently, revealing developmental variations that the intervention can strategically respond to.

Untimed tasks reveal both strong progress and some setbacks in Grade 2 across different skills. The largest and **statistically significant gains** were in **Subtraction Facts** (0.26 SD) and **Number Comparison** (0.20 SD) reflecting **strengthened number sense and arithmetic fluency**. At the same time, both Counting in Bundles (-0.21 SD) and Shape

Recognition (-0.20 SD) showed decline in performance, pointing to potential challenges in building conceptual understanding. **Timed tasks** highlight **similar trends**, though with smaller magnitudes. In Grade 2, Subtraction Facts (0.16 SD) was statistically significant, while Addition Facts (0.09 SD) showed smaller, non-significant improvement.

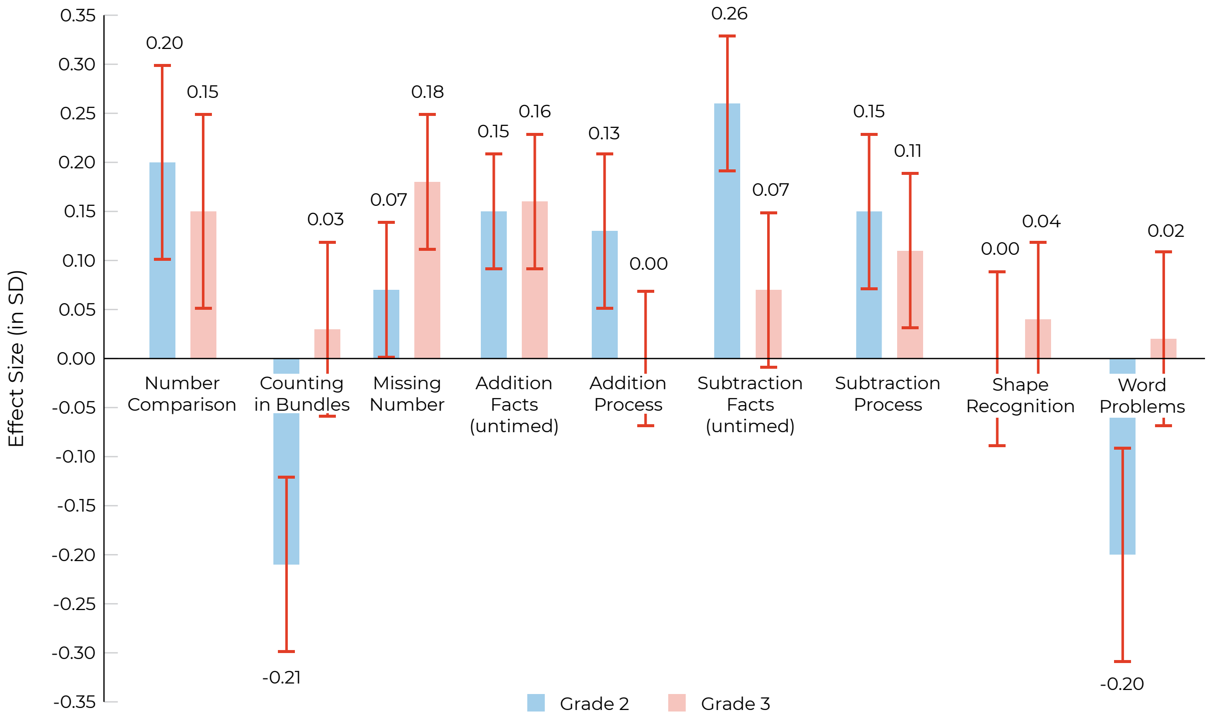


Figure 8: Grade-wise DiD Effect Sizes (in SD) – Untimed tasks.

The red lines denote the standard errors.

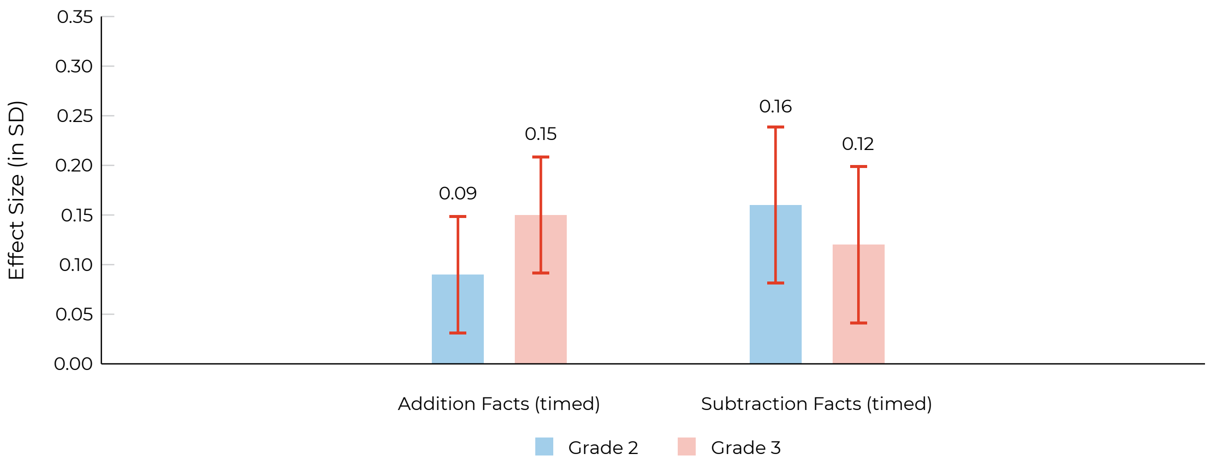


Figure 9: Grade-wise DiD Effect Sizes (in SD) – Timed tasks.

The red lines denote the standard errors.

Grade 3, in contrast, showed more balanced and steady improvement across skills. **Significant effects** were recorded in **Missing Number** (0.18 SD) and **Addition Facts** (untimed - 0.16 SD). Number Comparison (0.15 SD) and Addition Process (0.15 SD) also improved, while Word Problems (0.04 SD) and Shape Recognition (0.02 SD) showed limited progress. Unlike Grade 2, however, **no negative effects were observed**. Timed tasks in Grade 3 also showed steadier gains: Addition Facts (0.15 SD) and Subtraction Facts (0.12 SD) reflected progress but only the former was statistically significant. This suggests that Grade 3 learners may benefit from the programme's structure but require more **intensive scaffolding** to push skills beyond moderate growth—particularly in conceptual tasks that require language comprehension, spatial understanding, or multi-step reasoning[25].

Overall patterns suggest differential effects by grade. Grade 2 students showed larger gains in basic number sense and operations, alongside declines in some conceptual and visual–spatial tasks. Grade 3 students recorded smaller but more consistent improvements, particularly in recognition and arithmetic facts, while performance in reasoning and problem-solving tasks showed limited change. Taken together, the findings indicate that the programme primarily influenced foundational numeracy outcomes, with comparatively less movement observed in conceptual application.

4.4 Impact by Gender

This section examines performance differences between boys and girls to identify potential gender gaps in learning outcomes and assess whether the programme benefits were equitably distributed[26].

The ThinkZone intervention showed **consistent and positive effects across tasks for girls**. Girls registered gains in **Subtraction Proces**s (0.19 SD). They also showed some improvement in conceptual areas such as **Missing Number** (0.18 SD) and **Word problems** (0.10 SD). This stability suggests that the intervention supported girls reliably across domains. The consistency of progress also indicates that girls may respond well to structured practice formats that build gradually across skill levels[27].

25 Bakker, A., Smit, J., & Wegerif, R. (2015). Scaffolding and dialogic teaching in mathematics education: Introduction and review. ZDM – Mathematics Education, 47(7), 1047–1065. https://doi.org/10.1007/s11858-015-0738-8

26 Boys and girls were largely comparable in their numeracy skills at study onset, with 11 out of 13 measures demonstrating gender equivalence ($p>0.05$). Only 2 numeracy subtasks showed significant baseline differences: number comparison untimed ($p = 0.03$) and subtraction facts timed ($p = 0.01$).

27 Di Tommaso, M. L., Contini, D., De Rosa, D., Ferrara, F., Piazzalunga, D., & Robutti, O. (2024). Tackling the gender gap in mathematics with active and cooperative learning: Evidence from a randomized intervention. *Economics of Education Review, 100*, Article 102538. https://doi.org/10.1016/j.econedurev.2024.102538. The study finds that structured, guided, and cooperative learning formats significantly improve girls' mathematics outcomes, suggesting that girls may respond particularly well to instructional approaches that provide incremental practice and clear task sequencing in early skill development.

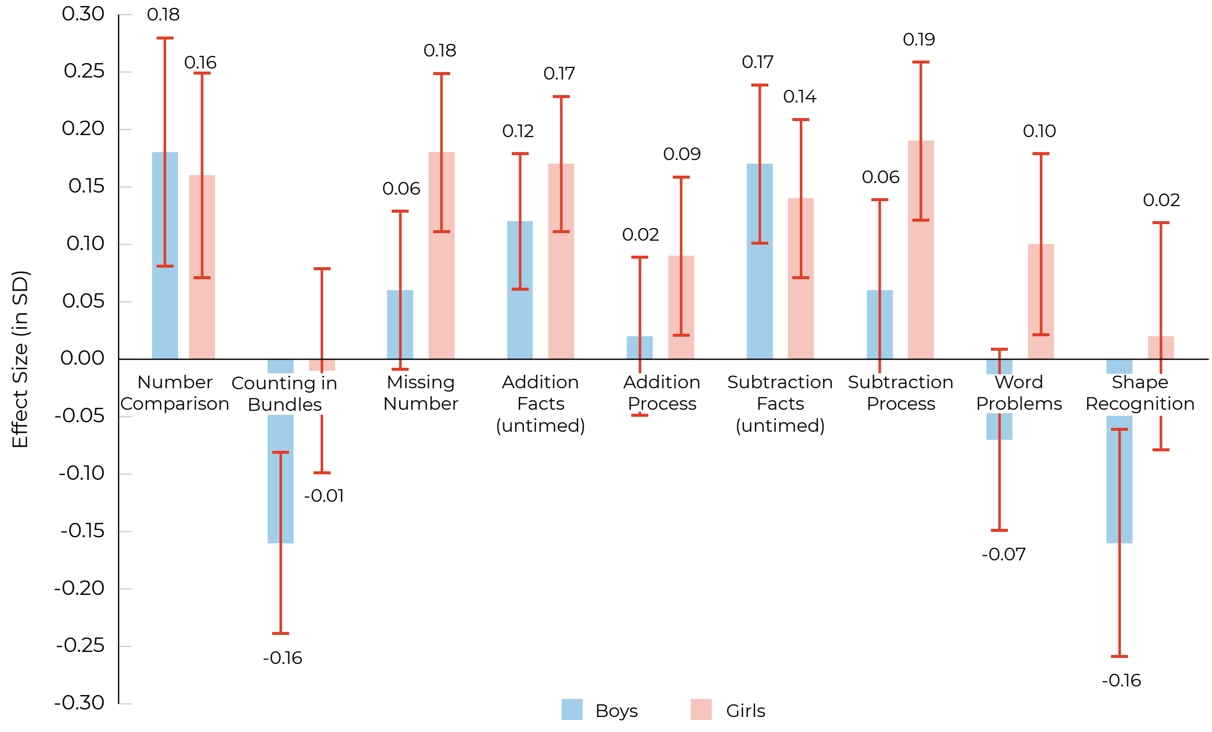


Figure 10: Gender-wise DiD Effect Sizes (in SD) – Untimed tasks.

The red lines denote the standard errors.

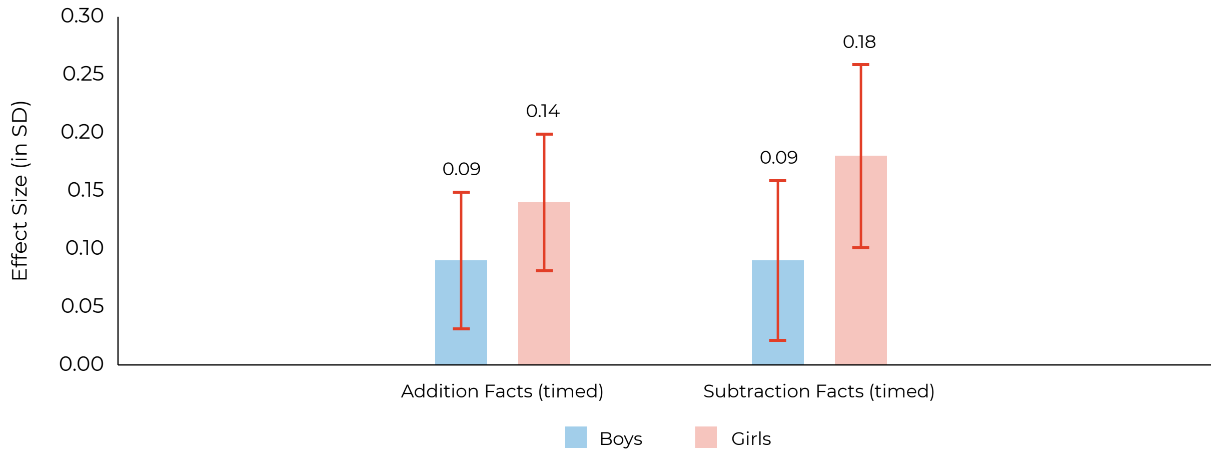


Figure 11: Gender-wise DiD Effect Sizes (in SD) – Timed tasks.

The red lines denote the standard errors.

In contrast, the programme recorded **mixed effects on boys' numeracy outcomes**. Boys demonstrated significant gains in certain skills such as **Number Comparison** (0.18 SD) and **Subtraction Facts** (0.17 SD, timed and untimed). However, improvements were **less pronounced in other areas**, with null effects (not statistically significant) in **Addition Process** (0.02 SD), **Word Problems** (-0.07 SD) and **Shape Recognition** (-0.16 SD) with negative effects in Counting in Bundles (-0.16 SD). This uneven pattern suggests that while boys responded well to fact-based and fluency-oriented tasks, they **struggled with spatial, or language-heavy**[28] demands.

Overall, the findings indicate clear gender-based differences in how students responded to the intervention. Girls showed a more consistent pattern of improvement across tasks, while boys' outcomes were more uneven, with gains concentrated in some skills and declines observed in others. These patterns point to differential learning trajectories by gender within the same programme context, highlighting variation not in access to benefits but in how those benefits manifested across skill types[29].

4.6 Score Distribution and Movement Analysis

This section presents the distribution of students across performance bands over assessment rounds, along with observed progressions and declines in student scores. Student performance at the sub-task level has been classified into five levels (L0–L4), defined separately for accuracy-based and fluency-based tasks, as outlined below.

Table 2: Categorisation for Accuracy and Fluency Tasks

Category	Accuracy Tasks (%)	Fluency Tasks (CPM - Correct Per Minute)
L0	0%	0
L1	0 to 25%	0 to 0.25 times Baseline Average (BL Avg.)
L2	25 to 50%	0.25 to 0.5 times BL Avg.
L3	50 to 75%	0.5 to BL Avg.
L4	75 to 100%	> BL Avg.
Note: All category lower limits are inclusive, except for L1; for example, 25–50% includes 25% but excludes 50%.		

These categories enable tracking of the magnitude and direction of change in student performance across the evaluation timeline.

The tables that follow present the percentage distribution of students across performance bands at Baseline and Endline for each assessed sub-task. Percentages reflect the proportion of students in each category at the respective assessment round, such that L0 + L1 + ... + L4 = 100% for Intervention and Comparison groups[30]:

28 Classification of skills, for the purposes of this study, are defined in the Glossary.
29 Evans, D. K. (2022). What we learn about girls' education from interventions. World Bank; Research on Improving Systems of Education (RISE). Synthesising evidence from education impact evaluations, Evans (2022) notes that gender-neutral education interventions often generate learning gains for girls that are equal to or larger than those for boys. These patterns are attributed to differences in baseline skill distributions and to girls disproportionately benefiting from improvements in instructional quality, structured pedagogy, and learning supports.
30 Detailed counts of students in each category, disaggregated by Intervention and Comparison groups, are provided in Annexure 6.9.

Table 3: Baseline Performance Category - Student Distribution

Sub-task	Intervention					Comparison				
	L0	L1	L2	L3	L4	L0	L1	L2	L3	L4
Number Comparison (untimed)	16%	1%	7%	12%	65%	16%	1%	2%	7%	74%
Counting in Bundles (untimed)	36%	0%	14%	10%	40%	33%	0%	13%	10%	44%
Missing Number (untimed)	18%	7%	19%	21%	35%	10%	6%	18%	20%	45%
Addition facts (untimed)	20%	0%	6%	10%	65%	10%	0%	5%	10%	76%
Addition facts (timed)	16%	5%	7%	21%	51%	8%	4%	11%	18%	59%
Addition Process (untimed)	25%	14%	10%	28%	24%	16%	11%	12%	32%	30%
Subtraction facts (untimed)	24%	0%	13%	13%	49%	16%	0%	11%	8%	66%
Subtraction facts (timed)	28%	5%	6%	18%	43%	19%	3%	4%	19%	55%
Subtraction Process (untimed)	41%	10%	15%	22%	12%	30%	9%	12%	27%	22%
Word problem (untimed)	35%	19%	17%	14%	15%	28%	21%	15%	14%	21%
Shape Recognition (untimed)	70%	0%	14%	11%	5%	74%	0%	17%	6%	3%

Across most sub-tasks, Intervention students show a clear upward shift from lower bands (L0–L2) to higher bands (L3–L4) at endline, especially in untimed number sense and operations (number comparison, addition/subtraction facts). Comparison schools also improve, but gains are smaller and uneven, with persistent concentration in L2–L3 for process-based tasks.

Table 4: Endline Performance Category - Student Distribution

Sub-task	Intervention					Comparison				
	L0	L1	L2	L3	L4	L0	L1	L2	L3	L4
Number Comparison (untimed)	15%	0%	3%	8%	74%	18%	0%	1%	7%	73%
Counting in Bundles (untimed)	33%	0%	12%	13%	42%	26%	0%	11%	14%	49%
Missing Number (untimed)	12%	6%	19%	21%	42%	10%	4%	17%	23%	45%
Addition facts (untimed)	14%	0%	6%	7%	73%	9%	0%	4%	7%	80%
Addition facts (timed)	11%	4%	5%	19%	60%	8%	2%	10%	19%	61%
Addition Process (untimed)	21%	10%	14%	32%	22%	16%	9%	15%	32%	28%
Subtraction facts (untimed)	19%	0%	14%	12%	55%	13%	0%	15%	10%	63%
Subtraction facts (timed)	21%	4%	5%	18%	51%	19%	3%	4%	16%	57%
Subtraction Process (untimed)	33%	14%	14%	26%	14%	30%	10%	13%	25%	21%
Word problem (untimed)	31%	22%	20%	13%	15%	28%	18%	20%	15%	20%
Shape Recognition (untimed)	81%	0%	11%	4%	4%	85%	0%	10%	3%	2%

In addition to examining the distribution of students across performance bands at Baseline and Endline, this study also tracked changes in individual student performance over time. As a longitudinal study involving the same students, this allows for an analysis of broad trends in movement across performance bands between the two rounds.

The score-transition patterns show that the **Intervention group demonstrated comparatively greater upward movement** across several tasks. Among the lowest-performing students (L0), progress was limited overall—especially on reasoning-heavy tasks such as Word Problems, where **53% of students in both groups remained at L0**[31]. On conceptual skills like Subtraction Process as well, many L0 students did not progress, though **fewer students in the Intervention group stayed at L0 (59%) compared to the Comparison group (65%)**. This trend repeats for other skills such as Missing Number, where **41% of Intervention students vs. 55% of Comparison students** remained at L0.

Mid-level performers (L2 and L3) in the Intervention group showed **more dynamic movement**, with students advancing, stagnating, or occasionally dropping levels—particularly on Addition and Subtraction Facts. For instance, in untimed Addition Facts, **60% moved upward (L2 to L3 or L4), 13% stayed at L2, and 27% moved down to L0 or L1**.

Table 5: Score Movement Analysis - Numeracy[32]

Upward movewwment >10% | Downward movement >10%

Addition Facts (untimed)					
Baseline Performance Category	Endline Performance Category				
	L0	L1	L2	L3	L4
Intervention					
L0	48%	0%	15%	13%	25%
L1	-	-	-	-	-
L2	37%	0%	4%	10%	49%
L3	10%	0%	12%	6%	72%
L4	2%	0%	2%	5%	91%
Comparison					
L0	61%	0%	18%	12%	10%
L1	-	-	-	-	-
L2	24%	0%	8%	12%	56%
L3	8%	0%	10%	17%	65%
L4	2%	0%	1%	4%	92%

Word Problems (Untimed)					
Baseline Performance Category	Endline Performance Category				
	L0	L1	L2	L3	L4
Intervention					
L0	53%	26%	12%	5%	4%
L1	30%	29%	26%	10%	5%
L2	18%	23%	28%	17%	14%
L3	13%	13%	25%	24%	25%
L4	12%	11%	14%	19%	45%
Comparison					
L0	53%	21%	15%	7%	4%
L1	28%	21%	25%	16%	10%
L2	17%	28%	25%	17%	12%
L3	11%	14%	27%	24%	24%
L4	12%	8%	11%	18%	51%

31 Verschaffel, L., Greer, B., & Van Dooren, W. (2020). Mathematical word problems: A critical analysis. Routledge. Word-problem solving requires the coordinated development of linguistic comprehension, contextual interpretation, and mathematical reasoning—capacities that tend to emerge gradually and are difficult to strengthen through short-term interventions.

32 Detailed analysis has been shared in Annexure 6.9.3 and 6.9.4.

Students in the highest band (L4) exhibited the most stability in both groups. For Number Comparison, **85% (Intervention) and 91% (Comparison)** stayed at L4 and for Missing Number, **64% and 67%**, respectively, remained at L4.

This indicates that high performers maintained their proficiency regardless of group assignment, which is expected given **ceiling effects** and points to the need for **enrichment-level tasks** for the programme to extend gains for advanced learners[33].

Table 6: Heatmap representing task-wise gains for each group

Gains of <5% | Gains of 5/>5% | Gains of 10/>10% | Decline

Task	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6	
	BL Avg.	EL Avg.	BL Avg.	EL Avg.	BL Avg.	EL Avg.	BL Avg.	EL Avg.	BL Avg.	EL Avg.	BL Avg.	EL Avg.
Number Comparison (untimed)	67%	80%	74%	74%	72%	78%	66%	73%	72%	80%	66%	83%
Counting in Bundles (untimed)	33%	33%	43%	47%	54%	44%	38%	47%	48%	51%	47%	49%
Missing Number (untimed)	46%	54%	48%	53%	52%	58%	46%	53%	52%	58%	57%	57%
Addition facts (untimed)	58%	70%	66%	71%	69%	77%	59%	72%	67%	76%	75%	73%
Addition facts (timed)	5.4	5.2	5.4	6.1	5.4	6.5	4.9	6	5.5	6.8	5.9	6.9
Addition Process (untimed)	39%	37%	42%	42%	46%	45%	42%	44%	45%	47%	53%	50%
Subtraction facts (untimed)	43%	55%	55%	55%	55%	60%	53%	59%	55%	62%	55%	62%
Subtraction facts (timed)	3.3	4	3.8	4.2	4.2	4.5	3.8	4.0	3.8	4.7	4.3	4.1
Subtraction Process (untimed)	24%	24%	29%	31%	35%	36%	29%	32%	31%	39%	36%	36%
Word Problems (untimed)	23%	27%	32%	30%	34%	34%	30%	31%	33%	36%	29%	35%
Shape Recognition (untimed)	14%	9%	13%	8%	21%	10%	16%	13%	18%	13%	25%	14%

Overall, the Intervention generated modest gains in procedural numeracy skills and reduced stagnation among the lowest performers, though progress in reasoning-oriented competencies remained limited. Compared to the largely static patterns in the Comparison group, even these small shifts suggest that the Intervention enabled marginal upward mobility that may not have occurred otherwise.

33 Ziernwald, L., Hillmayr, D., & Holzberger, D. (2022). Promoting high-achieving students through differentiated instruction in mixed-ability classrooms: A systematic review. Journal of Advanced Academics, 33(4), 540–573. https://doi.org/10.1177/1932202X221112931. Ziernwald, Hillmayr, and Holzberger (2022) argue that ceiling effects due to high pretest performance can mask true effects for advanced learners, highlighting the importance of differentiated, enrichment-oriented tasks to sustain growth for high performers in intervention settings.

5. Discussion and Implications

This evaluation of the ThinkZone intervention in Odisha offers substantial insights into its effectiveness, reach and the challenges present in advancing FLN for young learners in low-resource settings. Using a structured research design and standard assessment tools, the study traces the impacts of ThinkZone's approach across multiple layers - student performance, grade levels and gender differences.

- **The EYOS estimate of approximately 1.5 years indicates that the intervention enabled students to achieve learning gains equivalent to one and a half years of Business-As-Usual instruction over the study period.** This magnitude reflects substantial learning acceleration rather than incremental improvement, suggesting that the programme meaningfully compresses the time required for students to reach expected competency levels. Such gains are particularly consequential in low-learning contexts, where delayed foundational skills compound over time. The EYOS result underscores the potential of well-designed, high-intensity EdTech models to not only raise learning levels but also close accumulated learning gaps within a relatively short timeframe, strengthening the case for scale-up where system capacity and implementation fidelity can be sustained.
- **Clear Gains in specific Foundational Skills**
  The intervention delivered statistically significant improvements in Number Comparison and Subtraction Facts. This suggests that the programme effectively strengthened students' foundational numeracy skills, particularly in comparing quantities and recalling basic subtraction facts—an encouraging indication of its core instructional strength. However, the gains were limited to these specific areas, implying that other targeted competencies may not have improved as intended; future iterations should refine or intensify those components to ensure more balanced impact across all learning outcomes.
- **Greater Impact in Lower Grade**
  Grade 2 students benefited more from the programme than Grade 3 students. This suggests that introducing such interventions earlier in primary school may maximise impact as children are in the most formative stages of skill acquisition[34].

34 Kim, J., Gilbert, J., Yu, Q., & Gale, C. (2021). Measures matter: A meta-analysis of the effects of educational apps on preschool to Grade 3 children's literacy and math skills. AERA Open, 7(1), 1–19. https://doi.org/10.1177/23328584211004183. In their meta-regression moderator analysis, estimated impacts were 0.18 SD higher (on average) in younger students than higher grade samples, suggesting educational-app effects may be larger when implemented earlier in children's learning trajectories.

- **Gendered Outcomes**
  Girls showed consistent and statistically significant improvements across foundational numeracy tasks. In contrast, boys exhibited uneven gains—performing better on fact-based, fluency skills (e.g., number comparison and subtraction) but showing weak or negative effects on tasks such as word problems—highlighting the need for gender-sensitive instructional adaptations.
- **Progression Among Lower Performers**
  The intervention was most effective in lifting students from the lowest performance bands in basic tasks, whereas movement in conceptual skills such as Word Problems, was more limited. Most foundational gains were seen among students starting at the lowest proficiency levels.

The ThinkZone model demonstrates the potential for a scalable, community-driven EdTech approach that supports foundational learning for children in low-resource settings. Future efforts could focus on extending skill development beyond the basics, addressing equity gaps, and maintaining engagement through adaptive and responsive strategies. The study's findings provide insights for ThinkZone and other programs with similar priorities, supporting efforts to advance foundational numeracy in India and comparable contexts.

6. Technical Annexures

6.1 Details of Sampling Design

- For the evaluation, using 95% confidence[35], 80% power[36] and 0.1225 variance[37] the required sample size is 325 students each in the Intervention and Comparison groups.
- Minimal Detectable Effect Size (MDES)[38] The study aims to be able to detect a 0.15 SD difference between the baseline and endline rounds. The formula used to calculate MDES is:

$$MDE = (t_K + t_\alpha) \cdot \sqrt{\frac{1}{P(1-P)} \cdot \frac{\sigma^2}{N}} \cdot \sqrt{1 + (m-1) \cdot ICC}$$

MDE = Minimal detectable effect
t_K and t_α = Critical values from Student's t for power K and significance level α
σ^2 = Variance; N = Sample size; P = Proportion in treatment; m = Cluster size;
ICC = Intraclass correlation

- Since sampling by school introduces intra-class correlation (ICC[39]), a design effect (DEFF) adjustment was applied. Assuming ICC = 0.1 and an average cluster size of 15 students, DEFF was calculated as:

$$DEFF = 1 + 0.1 \times (15 - 1) = 2.4$$

- This increased the intervention and comparison sample size to 343 each.
- To account for absenteeism[40] (25% buffer), the sample increased to 458 per group. Factoring in 50% retention for a longitudinal study[41], the required sample became 915 per group.
- Additionally, to estimate the Treatment-on-Treated (ToT) effect, it was assumed 40% of intervention students would meet the ideal usage threshold[42]. This required sampling 2,286 intervention students. Since the total intervention population is 2,080, a full census was conducted for the intervention group.

6.2 Selection of Comparison Group

The comparison group for this quasi-experimental study would consist of schools that have not been given the intervention but are as similar as possible to the schools in the

35 The degree of certainty that the results reflect the true effect in the population rather than chance.
36 The probability of correctly detecting an effect if it truly exists.
37 The amount of natural variation in the outcome across individuals rather than uniform responses.
38 The smallest difference between groups that the study is designed to detect.
39 A measure of how similar students are within the same class or school, which affects the extent of independent information each student contributes. For this study, the intra-class correlation (ICC) was assumed to be 0.1, resulting in a Design Effect of 2.4.
40 The expected loss of sample due to students being absent during assessment, estimated at 25% for this study.
41 As the study was longitudinal and conducted over several months, some students present at baseline were unavailable at endline, hence an attrition buffer was included.
42 The estimated impact only on those who actually received the intervention as expected, estimated at 40% for this study.

intervention group in terms of baseline (pre-intervention) characteristics. The objective of the comparison group is to capture what would have been the student learning outcomes in schools in the demonstration districts if the intervention was not given to these schools (i.e., the counterfactual). Hence, it is critical to identify the factors that are most likely to affect the outcome variable, and then match the intervention and comparison observation units on these factors to create a valid counterfactual. While there are different techniques for creating a valid comparison group such as Propensity Score Matching, Regression Discontinuity Design, Mahalanobis Distance Matching, **Coarsened Exact Matching (CEM)** was used in this study for the following reasons:

1. A Monotonic Imbalance Bounding (MIB) matching method, CEM aims to balance the intervention and comparison groups ex ante
2. CEM also bounds through ex ante user choice both the average treatment effect estimation error and the degree of model dependence
3. CEM has been shown to produce good covariate balance between exposure groups and, thus, to reduce the impact of confounding in observational causal inference

6.2.1 Coarsened Exact Matching (CEM)[43]

The basic idea of CEM is to coarsen each variable by recoding it, so that substantially indistinguishable values are grouped and assigned the same numerical value. An exact matching algorithm is then applied to the coarsened data to determine the matches and to prune unmatched units.

In simple terms, the CEM process does not try to match each student in the Intervention group with a comparable student in the comparison group. Instead, it groups the characteristics of all the students in a particular school in the intervention group and then tries to find a school in the comparison group whose grouped characteristics exactly match the grouped characteristics of the original school. For example, let us assume that there is a school in the intervention group with 5 students in Grade 2, whose age (in years) and gender (M / F) is as follows:

- 6, M
- 6, M
- 6, M
- 7, F
- 7, F

The CEM method will coarsen this data into groups or bins and represent the school as one with less than 10 students in Grade 2, between 50% to 75% of whom are male and whose ages are either 6 or 7 years, with 50% to 75% of them being 6 years old. As a result, this school will be an exact match for the following school in the same district where the intervention is not being implemented and which has 9 students in Grade 2, whose age (in years) and gender (M / F) is as follows:

43 Iacus, S. M., King, G., & Porro, G. (2011). Causal inference without balance checking: Coarsened exact matching. *Political Analysis, 20*(1), 1–24. https://doi.org/10.1093/pan/mpr013; Iacus, S. M., King, G., & Porro, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association, 106*(493), 345–361. https://doi.org/10.1198/jasa.2011.tm09599.

- 6, F
- 6, F
- 6, F
- 6, F
- 6, M
- 6, M
- 7, M
- 7, M
- 7, M

Thus, the result of the CEM process in this case will be that each school in the sampling frame for the intervention group will be paired with a comparable school in the sampling frame for the comparison group and the sampling of the intervention and comparison groups will be done from these two sets of schools. To account for unforeseen challenges and situations on the ground during the data collection, a list of replacement schools will also be identified for the schools that make up the sampling frame for the comparison group.

6.2.2 Variables for the CEM Process:

Given below is the list of school-level variables that were used for the CEM process, to conduct the pairwise matching of schools, based on the data available on the UDISE website. These variables were chosen due to their potential relevance to student learning outcomes, informed by prior project experience.

Table 7: School-level variables

#	School-level variables
1	Total number of students in Grade 2
2	Total number of students in Grade 3
3	Pupil teacher ratio for the entire school
4	Location (urban vs. rural)
5	School (national) management type (e.g., dept. of education, local body, govt. aided, etc.)
6	Composite Index of Technological Infrastructure related variables (comprising of: count of functional laptops, desktops, tablets, scanners, printers, webcams, digi-boards and access to the internet)[44]
7	School category (e.g., primary, primary with upper primary, etc.)

In scenarios where no exact match was present in the district for an intervention group school using all 7 variables mentioned above, some of these school-level variables were either coarsened further or dropped entirely in the subsequent rounds of the matching process, to ensure that a match was identified, while excluding the already matched intervention group schools. For example, in the absence of an exact match, the number of students in Grade 2 was coarsened into class intervals with a width of 20 students instead of 10 students for the next round of matching. However, the first 4 variables mentioned above were always retained in the CEM process.

44 All these variables are dichotomous in the UDISE dataset (meaning they have values of 1 for yes and 0 for no). The composite index was created by multiplying by taking the maximum value of the component variables, i.e., if any one of the variables had a value of 1, the composite variable would also have a value of 1.

6.2.3 Summary Table of Matching Exercise

Based on the approach detailed above, a summary of the results of the matching exercise is given below:

Table 8: Summary of Matching Exercise

Matching Round	No. of unique intervention group schools	No. of unique comparison group schools	Details of matching variables
1	110	61	The count of students in Grades 2 & 3 and the pupil-teacher ratio, were coarsened into class intervals of width 10 from 1 to 100, with a separate class for 0 and then into intervals of width 50 from 101 to 300, with a separate class for any values above 300. All 110 intervention groups schools matched in this round could have been covered by 27 comparison group schools, but the number was increased to 61 using a many-to-many matching approach, to sample the minimum number of comparison group schools required to maintain a cluster size of at least 15.
2	8	0	The school category variable was dropped.
3	11	0	The aggregated tech infra variable was also dropped.
4	2	0	The width of the class intervals for the coarsened count of students in Grades 2 & 3 and the pupil-teacher ratio, was increased from 10 to 20 for values between 1 to 100. The school category and aggregated tech infra variables were also re-introduced.
5	3	2	The width of the class intervals for the coarsened count of students in Grades 2 & 3 and the pupil-teacher ratio, was increased from 20 to 50 for values between 1 to 100.
6	2	1	The pupil-teacher ratio variable was dropped.
7	2	0	The pupil-teacher ratio variable was re-introduced and the count of students in Grade 3 variable was dropped. The width of the class intervals for the coarsened count of students in Grade 2 and the pupil-teacher ratio was reverted to 10 for values between 1 to 100.
8	1	0	The pupil-teacher ratio variable was dropped again, along with the count of students in Grade 3 variable that was dropped in the previous step.
Total	139	64	

The final matching process paired all 139 intervention group schools with 64 comparison group schools. One comparison group school was matched with up to 40 intervention schools and one intervention school with up to 35 comparison schools. While technically

32 comparison schools would have covered all intervention schools, 64 comparison schools were selected to maintain an average cluster size of 15 students per school, leading to a comparison group sample of 960 students.

6.3 Final Sample Achieved

The table below presents a detailed summary of the number of students assessed as part of the study, along with the number excluded from the analysis due to the following reasons:

- Attempting only one subject in a given round due to lack of consent, operational challenges, or other constraints
- Absent or unsynced audio response recordings / assets
- Presence in the Baseline but absence in the Endline

Table 9: Final Sample Achieved in Intervention and Comparison groups

Description	Intervention				Comparison			
	Baseline		Endline		Baseline		Endline	
	Literacy	Numeracy	Literacy	Numeracy	Literacy	Numeracy	Literacy	Numeracy
Total number of assessments administered	1414	1357	1320	1232	739	727	708	677
Students excluded due to operational constraints[45]	67	10	97	9	27	15	45	14
Students excluded due to technical issues[46]	164		134		42		25	
Students excluded as per study design criteria[47]	359		265		135		103	
Final Student Sample	824				535			

45 Students who attempted only one of the two subjects within a given assessment round.
46 The student's assessment having absent or unsynced recordings/assets.
47 Students not present in both assessment rounds i.e. Baseline and Endline.

6.4 Modified EGMA Tool

The below table summarises the tool utilised in this evaluation.

Table 10: Details of modified Numeracy Tool and Skills Covered

Concept / Skill Assessed	Total Items/ Questions	Nature of Task	Metric	Objective	Mode of Evaluation
Number discrimination (Identify and read out the greater number among the given pairs of numbers)	10	Untimed	Number of questions answered correctly	The objective of this task is to check if the child can compare and order numbers up to 1,000 accurately.	Automated
Counting in bundles (Give the total count of straws when given a few bundles of 10 straws each, along with some single straws. Also choose the correct count of bundles of 10 straws and single straws that add up to a given number.)	4	Untimed	Number of questions answered correctly	The objective of this task is to check if the child understands the concept of place value, i.e., if they understand how many tens and ones make a 2-digit number.	Automated
Number patterns (Identify a pattern in numbers and call out the missing number in each pattern)	8	Untimed	Number of missing numbers identified correctly	The objective of this task is to check if the child can recognize and complete patterns involving skip counting in 1s, 10s, 5s and 2s, for numbers up to 1,000.	Automated
Single-digit addition facts (Solve single-digit addition problems)	3	Untimed	Number of questions answered correctly	The objective of this task is to check if the child can add single-digit numbers fluently.	Automated
	20	Timed: 1 minute			Automated
Multi-digit addition in vertical format (Solve multi-digit addition problems)	5	Untimed	Number of questions answered correctly	The objective of this task is to check if the child can add multi-digit numbers accurately.	Automated
Subtraction facts within 18 (Solve subtraction problems involving numbers within 18 and with single-digit answers)	3	Untimed	Number of questions answered correctly	The objective of this task is to check if the child can subtract numbers up to 18 and with single-digit answers, fluently.	Automated
	20	Timed: 1 minute			Automated

Concept / Skill Assessed	Total Items/ Questions	Nature of Task	Metric	Objective	Mode of Evaluation
Multi-digit subtraction in vertical format (Solve multi-digit subtraction problems)	5	Untimed	Number of questions answered correctly	The objective of this task is to check if the child can subtract multi-digit numbers accurately.	Automated
Operations involving 0 (Solve addition and subtraction problems involving 0)	4	Untimed	Number of questions answered correctly	The objective of this task is to check if the child understands the concept of zero accurately. 1 question each in addition facts (untimed), addition process, subtraction facts (untimed) and subtraction process[48].	Automated
Word problems on single-digit number operations (Give an answer to narrated word problems)	6	Untimed	Number of questions answered correctly	The objective of this task is to check if the child can apply basic addition and subtraction in real-world scenarios.	Automated
Shape recognition (Identify a specific shape among a collection of shapes)	3	Untimed	Number of shapes recognized correctly	The objective of this task is to check if the child can recognize basic shapes like circles, rectangles and triangles. One question each on circles, rectangles and triangles. Each question involves showing the student multiple shapes (8, 10 and 11 respectively), of which only a few (2, 4 and 4 respectively) are correct examples of the shape the student has been asked to identify.	Automated

48 These items are already included in the count of the respective sub-tasks.

6.5 Demographic Analysis of Students Who Dropped Out after Baseline

Table 11: Demographic Analysis of Baseline Dropouts

Indicator		Intervention		Comparison	
		Students absent after Baseline	Overall Student Pool	Students absent after Baseline	Overall Student Pool
No. of Students		359	1341	135	729
Student Age (avg.)		7.5 (7.4 - 7.6)*	7.6 (7.5 - 7.6)	8.1 (7.9 - 8.4)	7.4 (7.4 - 7.5)
Grade Split	Class 2	46% (38% - 53%)	46% (42% -50%)	63% (53% - 73%)	51% (41% - 62%)
	Class 3	54% (47% - 61%)	54% (50% - 58%)	37% (24% - 50%)	49% (35% - 63%)
No. of Sisters (avg.)		0.7 (0.6 - 0.8)	0.9 (0.8 - 0.9)	0.8 (0.7 - 1.0)	0.9 (0.8 - 0.9)
No. of Brothers(avg.)		0.7 (0.6 - 0.8)	0.8 (0.7 - 0.8)	0.7 (0.6 - 0.9)	0.8 (0.7 - 0.8)
Socio-economic Context*	Low	6% (-4% - 16%)	8% (3% - 13%)	14% (0% - 29%)	14% (-1% - 28%)
	High	94% (91% - 96%)	92% (90% - 93%)	86% (82% - 90%)	86% (83% - 90%)

* 95% Confidence Interval mentioned in brackets
** Binary flag created on the basis of appliances/transport available at home (i.e. feature phone v. smart phone; cycle v. motor cycle)

Across both Intervention and Comparison groups, students who dropped out after baseline were broadly similar in age to the overall student pool, as indicated by overlapping 95% confidence intervals. In the Intervention group, the average age of dropouts (7.5 years; CI: 7.4–7.6) closely aligned with the overall pool (7.6 years; CI: 7.5–7.6). In the Comparison group, dropouts were slightly older on average (8.1 years; CI: 7.9–8.4) than the overall pool (7.4 years; CI: 7.4–7.5), with limited overlap in confidence intervals, suggesting a small age difference. Grade composition patterns were largely comparable within the Intervention group, with overlapping intervals for Classes 2 and 3. In the Comparison group, however, dropouts were more concentrated in Class 2 (63%; CI: 53%–73%) relative to the overall pool (51%; CI: 41%–62%), indicating a possible skew towards lower grades among those absent after baseline.

Household composition, measured through the average number of sisters and brothers, showed minimal differences between students who dropped out after baseline and the overall student pool. In both Intervention and Comparison groups, the confidence intervals for the number of sisters and brothers overlapped substantially across dropouts and the full sample. For instance, in the Intervention group, dropouts reported an average of 0.7 sisters (CI: 0.6–0.8) compared to 0.9 (CI: 0.8–0.9) in the overall pool, while brother counts were similarly aligned. These overlaps suggest that sibling composition is unlikely to be a distinguishing factor associated with baseline dropout in either group.

Socio-economic status distributions among students who dropped out after baseline were largely consistent with those of the overall student pool, as reflected by overlapping confidence intervals for both low and high socio-economic categories. In the Intervention group, the proportion of dropouts from low socio-economic households (6%; CI: –4% to 16%) was comparable to the overall pool (8%; CI: 3%–13%), while the high socio-economic category showed similarly close alignment. In the Comparison group, point estimates for low socio-economic status were identical for dropouts and the overall pool (14%), with wide and overlapping confidence intervals.

Overall, the confidence interval evidence suggests that baseline dropout was not strongly patterned by socio-economic context and students who dropped out were demographically similar to the broader study population across most measured characteristics[49].

6.6 Overall Performance

Table 12: Detailed Results

ThinkZone											
Task	Task Type	Unit	Endline Avg.		Baseline Avg.		Delta (EL - BL)		Pooled SD	Effect Size (SD)	p-value
			I	C	I	C	I	C			
Number Comparison	Untimed	%	77%	76%	72%	77%	5%	-1%	36%	0.17	0.01
Counting in Bundles	Untimed	%	47%	53%	45%	47%	2%	6%	40%	-0.08	0.19
Missing Number	Untimed	%	56%	60%	50%	58%	6%	2%	33%	0.12	0.01
Addition Facts	Untimed	%	74%	80%	66%	78%	7%	2%	36%	0.15	0.00
Addition Facts	Timed	CPM	6.4	7.2	5.4	6.7	0.9	0.5	3.89	0.12	0.00
Addition Process	Untimed	%	44%	50%	44%	52%	1%	-1%	34%	0.06	0.25
Subtraction Facts	Untimed	%	59%	66%	54%	68%	4%	-2%	38%	0.16	0.00
Subtraction Facts	Timed	CPM	4.4	4.9	3.8	4.8	0.6	0.1	3.43	0.13	0.01
Subtraction Process	Untimed	%	34%	40%	31%	40%	4%	-1%	33%	0.13	0.01
Word Problems	Untimed	%	33%	38%	32%	38%	1%	0%	32%	0.02	0.65
Shape Recognition	Untimed	%	10%	8%	17%	12%	-6%	-5%	25%	-0.07	0.36

6.7 Grade-wise Results

49 It is important to note that demographic information was not mandatory during data collection and therefore, while representative, it may be incomplete.

Table 13: Detailed Results - Grade 2

ThinkZone - Grade 2											
Task	Task Type	Unit	Endline Avg.		Baseline Avg.		Delta (EL - BL)		Pooled SD	Effect Size (SD)	p-value
			I	C	I	C	I	C			
Number Comparison	Untimed	%	72%	70%	65%	71%	7%	-1%	38%	0.20	0.03
Counting in Bundles	Untimed	%	36%	47%	36%	38%	0%	8%	39%	-0.21	0.02
Missing Number	Untimed	%	45%	51%	39%	48%	6%	3%	33%	0.07	0.32
Addition Facts	Untimed	%	64%	75%	53%	70%	11%	5%	40%	0.15	0.02
Addition Facts	Timed	CPM	5.1	6.0	4.1	5.3	1.0	0.7	3.7	0.09	0.10
Addition Process	Untimed	%	35%	41%	33%	44%	1%	-3%	33%	0.13	0.10
Subtraction Facts	Untimed	%	51%	57%	43%	60%	8%	-3%	40%	0.26	0.00
Subtraction Facts	Timed	CPM	3.3	3.8	2.8	3.8	0.5	0.0	3.1	0.16	0.05
Subtraction Process	Untimed	%	24%	31%	21%	32%	3%	-2%	30%	0.15	0.04
Word Problems	Untimed	%	24%	31%	23%	30%	1%	1%	29%	0.00	0.85
Shape Recognition	Untimed	%	7%	5%	15%	8%	-8%	-3%	23%	-0.20	0.09

Table 14: Detailed Results - Grade 3

ThinkZone - Grade 3											
Task	Task Type	Unit	Endline Avg.		Baseline Avg.		Delta (EL - BL)		Pooled SD	Effect Size (SD)	p-value
			I	C	I	C	I	C			
Number Comparison	Untimed	%	81%	82%	77%	82%	4%	-1%	34%	0.15	0.13
Counting in Bundles	Untimed	%	56%	59%	52%	56%	4%	3%	40%	0.03	0.77
Missing Number	Untimed	%	64%	67%	59%	67%	6%	0%	31%	0.18	0.01
Addition Facts	Untimed	%	81%	85%	77%	85%	4%	0%	30%	0.16	0.02
Addition Facts	Timed	CPM	7.4	8.3	6.6	8.0	0.8	0.2	3.7	0.15	0.01
Addition Process	Untimed	%	52%	59%	52%	59%	0%	0%	33%	0.00	0.96

ThinkZone - Grade 3											
Task	Task Type	Unit	Endline Avg.		Baseline Avg.		Delta (EL - BL)		Pooled SD	Effect Size (SD)	p-value
			I	C	I	C	I	C			
Subtraction Facts	Untimed	%	65%	74%	64%	75%	1%	-1%	35%	0.07	0.34
Subtraction Facts	Timed	CPM	5.2	5.9	4.6	5.7	0.6	0.2	3.5	0.12	0.10
Subtraction Process	Untimed	%	42%	48%	38%	48%	4%	0%	34%	0.11	0.10
Word Problems	Untimed	%	40%	44%	39%	45%	1%	-1%	32%	0.04	0.65
Shape Recognition	Untimed	%	12%	10%	17%	15%	-5%	-6%	27%	0.02	0.78

6.8 Gender-wise Results

Table 15: Detailed Results - Boys

ThinkZone - Boys											
Task	Task Type	Unit	Endine - Boys		Baseline - Boys		Delta - Boys		Pooled SD	Effect Size	p-value
			I	C	I	C	I	C			
Number Comparison	Untimed	%	75%	75%	73%	80%	2%	-5%	36%	0.18	0.07
Counting in Bundles	Untimed	%	47%	56%	45%	48%	2%	8%	40%	-0.16	0.06
Missing Number	Untimed	%	52%	60%	48%	58%	4%	2%	34%	0.06	0.38
Addition Facts	Untimed	%	70%	82%	63%	80%	7%	2%	37%	0.12	0.08
Addition Facts	Timed	CPM	5.9	7.1	5.1	6.7	0.8	0.4	3.9	0.09	0.09
Addition Process	Untimed	%	42%	52%	42%	53%	0%	-1%	35%	0.02	0.74
Subtraction Facts	Untimed	%	55%	64%	51%	67%	4%	-3%	39%	0.17	0.01
Subtraction Facts	Timed	CPM	4.0	4.7	3.5	4.6	0.4	0.1	3.5	0.09	0.21
Subtraction Process	Untimed	%	30%	39%	28%	39%	2%	0%	33%	0.06	0.39
Word Problems	Untimed	%	32%	39%	31%	37%	0%	2%	32%	-0.07	0.50
Shape Recognition	Untimed	%	11%	9%	17%	11%	-6%	-2%	26%	-0.16	0.15

Table 16: Detailed Results - Girls

ThinkZone - Girls											
Task	Task Type	Unit	Endline - Girls		Baseline - Girls		Delta - Girls		Pooled SD	Effect Size	p-value
			I	C	I	C	I	C			
Number Comparison	Untimed	%	79%	76%	70%	73%	9%	3%	37%	0.16	0.08
Counting in Bundles	Untimed	%	48%	51%	45%	47%	3%	3%	40%	-0.01	0.95
Missing Number	Untimed	%	59%	60%	52%	58%	7%	1%	32%	0.18	0.01
Addition Facts	Untimed	%	77%	78%	69%	76%	8%	2%	35%	0.17	0.01
Addition Facts	Timed	CPM	6.8	7.2	5.7	6.7	1.0	0.5	3.9	0.14	0.01
Addition Process	Untimed	%	46%	48%	45%	51%	1%	-2%	34%	0.09	0.19
Subtraction Facts	Untimed	%	62%	67%	57%	68%	4%	-1%	37%	0.14	0.03
Subtraction Facts	Timed	CPM	4.8	5.1	4.1	5.0	0.7	0.1	3.3	0.18	0.02
Subtraction Process	Untimed	%	38%	40%	33%	42%	5%	-1%	33%	0.19	0.00
Word Problems	Untimed	%	34%	36%	32%	38%	1%	-2%	31%	0.10	0.19
Shape Recognition	Untimed	%	9%	6%	15%	13%	-6%	-7%	24%	0.02	0.83

6.9 Score Movement Analysis

6.9.1 Baseline Performance Distribution - Count of Students

Table 17: Baseline Performance Category - Count of Students

Sub-task	Intervention					Comparison				
	L0	L1	L2	L3	L4	L0	L1	L2	L3	L4
Number Comparison (untimed)	128	8	54	99	535	83	6	11	39	396
Counting in Bundles (untimed)	295	0	115	81	333	176	0	67	56	236
Missing Number (untimed)	148	56	157	173	290	56	31	98	108	242
Addition facts (untimed)	162	0	49	81	532	51	0	25	52	407
Addition facts (timed)	130	38	56	176	424	44	23	57	94	317

Sub-task	Intervention					Comparison				
	L0	L1	L2	L3	L4	L0	L1	L2	L3	L4
Addition Process (untimed)	202	112	82	227	201	83	56	63	169	162
Subtraction facts (untimed)	198	0	110	109	407	83	0	57	43	350
Subtraction facts (timed)	230	45	46	145	358	101	16	21	101	294
Subtraction Process (untimed)	334	83	120	185	102	158	48	66	143	118
Word problem (untimed)	288	156	144	112	124	150	113	81	75	114
Shape Recognition (untimed)	580	0	117	87	40	396	0	90	31	15

6.9.2 Endline Performance Distribution - Count of Students

Table 18: Endline Performance Category - Count of Students

Sub-task	Intervention					Comparison				
	L0	L1	L2	L3	L4	L0	L1	L2	L3	L4
Number Comparison (untimed)	123	4	24	65	607	95	2	7	38	392
Counting in Bundles (untimed)	270	0	98	108	347	139	0	60	75	260
Missing Number (untimed)	100	49	157	169	344	55	21	91	125	240
Addition facts (untimed)	112	0	46	59	601	48	0	22	36	426
Addition facts (timed)	88	32	44	159	493	44	9	53	103	321
Addition Process (untimed)	171	85	116	263	179	85	47	78	170	149
Subtraction facts (untimed)	153	0	113	100	448	69	0	77	51	330
Subtraction facts (timed)	175	35	43	147	414	101	18	22	84	302
Subtraction Process (untimed)	266	111	113	208	116	159	55	70	132	111
Word problem (untimed)	253	177	159	103	122	145	96	103	80	103
Shape Recognition (untimed)	661	0	88	31	33	446	0	52	17	12

6.9.3 Student Score Movement from Baseline to Endline - Intervention

Table 19: Score Movements - Intervention

>10% Downward Movement >10% Upward Movement

Task	Baseline Performance Category	Endline Performance Category				
		L0	L1	L2	L3	L4
Number Comparison (untimed)	L0	27%	0%	9%	7%	58%
	L1	50%	0%	13%	0%	38%
	L2	13%	4%	15%	24%	44%
	L3	27%	1%	2%	19%	51%
	L4	10%	0%	0%	4%	85%
Counting in Bundles (untimed)	L0	52%	0%	13%	12%	23%
	L1	-	-	-	-	-
	L2	42%	0%	18%	15%	25%
	L3	25%	0%	10%	19%	47%
	L4	15%	0%	9%	12%	64%
Missing Number (untimed)	L0	41%	13%	21%	13%	12%
	L1	14%	23%	38%	23%	2%
	L2	12%	8%	36%	27%	18%
	L3	5%	2%	20%	30%	44%
	L4	2%	1%	5%	15%	77%
Addition facts (untimed)	L0	48%	0%	15%	13%	25%
	L1	-	-	-	-	-
	L2	37%	0%	4%	10%	49%
	L3	10%	0%	12%	6%	72%
	L4	2%	0%	2%	5%	91%
Addition facts (timed)	L0	41%	10%	12%	24%	13%
	L1	31%	3%	14%	37%	14%
	L2	16%	11%	13%	35%	25%
	L3	5%	5%	8%	35%	47%
	L4	1%	1%	0%	8%	89%

Task	Baseline Performance Category	Endline Performance Category				
		L0	L1	L2	L3	L4
Addition Process (untimed)	L0	52%	14%	14%	18%	3%
	L1	21%	24%	21%	29%	5%
	L2	17%	9%	22%	33%	19%
	L3	9%	6%	17%	43%	25%
	L4	5%	6%	5%	36%	48%
Subtraction facts (untimed)	L0	53%	0%	21%	11%	15%
	L1	-	-	-	-	-
	L2	18%	0%	20%	21%	40%
	L3	14%	0%	13%	17%	56%
	L4	4%	0%	9%	9%	78%
Subtraction facts (timed)	L0	50%	7%	8%	17%	18%
	L1	45%	14%	11%	23%	7%
	L2	13%	7%	15%	28%	37%
	L3	10%	3%	7%	24%	55%
	L4	6%	2%	1%	14%	77%
Subtraction Process (untimed)	L0	59%	14%	12%	13%	3%
	L1	29%	20%	20%	23%	9%
	L2	16%	17%	21%	37%	9%
	L3	12%	13%	13%	41%	22%
	L4	7%	7%	11%	27%	47%
Word Problems (untimed)	L0	53%	26%	12%	5%	4%
	L1	30%	29%	26%	10%	5%
	L2	18%	23%	28%	17%	14%
	L3	13%	13%	25%	24%	25%
	L4	12%	11%	14%	19%	45%
Shape Recognition (untimed)	L0	86%	0%	7%	3%	3%
	L1	-	-	-	-	-
	L2	79%	0%	10%	5%	6%
	L3	72%	0%	21%	1%	5%
	L4	63%	0%	24%	8%	5%

6.9.4 Student Score Movement from Baseline to Endline - Comparison

Table 20: Score Movements – Comparison

>10% Downward Movement   >10% Upward Movement

Task	Baseline Performance Category	Endline Performance Category				
		L0	L1	L2	L3	L4
Number Comparison (untimed)	L0	27%	1%	2%	12%	58%
	L1	33%	0%	0%	33%	33%
	L2	45%	0%	9%	9%	36%
	L3	33%	0%	5%	13%	49%
	L4	13%	0%	1%	5%	81%
Counting in Bundles (untimed)	L0	44%	0%	16%	8%	32%
	L1	-	-	-	-	-
	L2	30%	0%	20%	20%	30%
	L3	20%	0%	11%	21%	48%
	L4	13%	0%	6%	15%	67%
Missing Number (untimed)	L0	55%	5%	16%	5%	18%
	L1	19%	19%	29%	23%	10%
	L2	9%	6%	42%	30%	12%
	L3	3%	1%	21%	34%	41%
	L4	2%	2%	4%	21%	71%
Addition facts (untimed)	L0	61%	0%	18%	12%	10%
	L1	-	-	-	-	-
	L2	24%	0%	8%	12%	56%
	L3	8%	0%	10%	17%	65%
	L4	2%	0%	1%	4%	92%
Addition facts (timed)	L0	51%	7%	19%	16%	7%
	L1	24%	10%	38%	24%	5%
	L2	5%	7%	37%	39%	12%
	L3	5%	0%	7%	37%	50%
	L4	3%	0%	3%	11%	83%
Addition Process (untimed)	L0	57%	12%	19%	12%	0%
	L1	25%	16%	21%	29%	9%
	L2	11%	16%	18%	34%	20%
	L3	5%	7%	17%	43%	28%
	L4	5%	4%	8%	31%	53%

Task	Baseline Performance Category	Endline Performance Category				
		L0	L1	L2	L3	L4
Subtraction facts (untimed)	L0	52%	0%	25%	14%	10%
	L1	-	-	-	-	-
	L2	21%	0%	32%	11%	36%
	L3	9%	0%	14%	16%	60%
	L4	3%	0%	9%	8%	79%
Subtraction facts (timed)	L0	48%	8%	9%	10%	24%
	L1	67%	13%	7%	13%	0%
	L2	40%	0%	10%	30%	20%
	L3	16%	3%	5%	23%	53%
	L4	6%	2%	2%	14%	76%
Subtraction Process (untimed)	L0	65%	7%	13%	10%	5%
	L1	29%	25%	15%	19%	13%
	L2	22%	14%	20%	38%	6%
	L3	11%	13%	17%	35%	25%
	L4	12%	4%	5%	29%	50%
Word Problems (untimed)	L0	53%	21%	15%	7%	4%
	L1	28%	21%	25%	16%	10%
	L2	17%	28%	25%	17%	12%
	L3	11%	14%	27%	24%	24%
	L4	12%	8%	11%	18%	51%
Shape Recognition (untimed)	L0	89%	0%	7%	2%	1%
	L1	-	-	-	-	-
	L2	77%	0%	15%	4%	4%
	L3	58%	0%	23%	8%	12%
	L4	73%	0%	20%	7%	0%

6.10 Analysis of Literacy Investigation

6.10.1 Overview

The literacy assessment was administered alongside the numeracy assessment and covered an identical sample of students. The final analytical sample includes only those students who participated in both literacy and numeracy assessments across the Baseline and Endline rounds. Literacy outcomes were measured using a contextualised version of the Early Grade Reading Assessment (EGRA), which was administered uniformly to students in Grades 2 and 3 and remained consistent across both rounds. To ensure alignment with the programme context, the assessment tool was adapted and contextualised and translated into Odia.

This literacy investigation is indicative and should be interpreted with appropriate caution. All the literacy sub-tasks relied on audio-based assessments. Variations in audio clarity and inconsistencies in recording conditions have influenced the measurement precision.

6.10.2 Indicative Investigation

The ThinkZone literacy assessment indicates significant positive gains in literacy, with the largest improvements observed in Letter Recognition (untimed: 0.36 SD; timed: 0.21 SD), reflecting gains in both accuracy and fluency. Familiar Word Reading also showed notable gains (untimed: 0.28 SD; timed: 0.18 SD), indicating improved ability to recognise common words, an essential bridge between letter knowledge and fluent reading. Listening Comprehension also improved by 0.20 SD, suggesting gains in students' ability to understand spoken text.

Table 21: ThinkZone Literacy Results

ThinkZone - Literacy											
Task	Task Type	Unit	Endline Avg.		Baseline Avg.		Delta (EL - BL)		Pooled SD	Effect Size (SD)	p-value50
			I	C	I	C	I	C			
Listening Comprehension	Untimed	%	49%	49%	38%	46%	10%	3%	35%	0.20	0.00
Oral Vocabulary	Untimed	%	74%	74%	70%	73%	4%	1%	24%	0.12	0.11
Letter Recognition	Untimed	%	83%	82%	73%	82%	10%	1%	26%	0.36	0.00
Letter Recognition	Timed	CPM	26.3	25.9	23.2	25	3.1	1	9.98	0.21	0.00
Familiar Word Reading	Untimed	%	86%	86%	79%	87%	7%	-1%	28%	0.28	0.00
Familiar Word Reading	Timed	CPM	21.2	22	18.2	20.9	3	1.1	10.07	0.18	0.00
Nonword Reading	Timed	CPM	18.9	19.7	16.5	18.4	2.4	1.3	9.83	0.11	0.03
Oral Reading Fluency	Timed	CPM	35.2	39.8	28	36.4	7.1	3.3	28.22	0.13	0.00
Reading Comprehension	Untimed	%	45%	50%	34%	41%	11%	9%	39%	0.06	0.17

Oral Reading Fluency (0.13 SD) and Nonword Reading (0.11 SD) demonstrated statistically significant gains of smaller magnitude. In contrast, Oral Vocabulary (0.12 SD) and Reading Comprehension (0.06 SD)) showed non-significant gains. Overall, this indicates that the programme may be effectively strengthening foundational decoding and recognition skills.

50 A p-value of less than 0.05 is considered statistically significant.

LiftEd Transforming Foundational Literacy and Numeracy

EDTECH ACCELERATOR

To learn more about the

LiftEd EdTech Accelerator,

visit our website:

www.edtechaccelerator.org

Contact us at:
edtechaccelerator@centralsuarefoundation.org