What Works Hub for Global Education

Cheaper (and more effective) by the dozen: Evidence from 12 randomised A/B tests optimising tutoring for scale

Working Paper

Noam Angrist, Claire Cullen, Janica Magat

October 2025

Blavatnik School of Government
University of Oxford
UK International Development
Partnership | Progress | Prosperity
Youth Impact

Cheaper (and more effective) by the dozen: Evidence from 12 randomised A/B tests optimising tutoring for scale

Working Paper

Noam Angrist

University of Oxford, Youth Impact

noam.angrist@bsg.ox.ac.uk

Claire Cullen

Youth Impact

ccullen@youth-impact.org

Janica Magat

Youth Impact

jmagat@youth-impact.org

The What Works Hub for Global Education is an international partnership, funded by the UK government's Foreign, Commonwealth & Development Office and the Bill & Melinda Gates Foundation, working out how to effectively implement education reforms at scale.

Please cite this as:

Angrist, N., Cullen, C. & Magat, J. 2025. Cheaper (and more effective) by the dozen: Evidence from 12 randomised A/B tests optimising tutoring for scale. What Works Hub for Global Education Working Paper Series. 2025/001. https://doi.org/10.35489/BSG-WhatWorksHubforGlobalEducation-WP_2025/001

Cheaper (and more effective) by the dozen: Evidence from 12 randomized A/B tests optimizing tutoring for scale

Noam Angrist, Claire Cullen, Janica Magat*

Abstract

Over the course of 12 rapid randomized experiments, we optimize an educational tutoring program. Tutoring is one of the most effective educational approaches yet has remained difficult to scale due to high costs. We adaptively test and improve a technology-enabled tutoring program to enhance cost-effectiveness and scalability. Results show that seven of twelve tests led to efficiency improvements, a "rate of discovery" of 58%. This compares favorably to the tech sector where 10-40 percent of tests generate improvements, demonstrating the potential for A/B testing to yield large efficiency gains in the education sector. The largest efficiency gains were driven by cost-reducing modifications that streamlined labor-intensive implementation processes and effectiveness-enhancing innovations that actively involved caregivers in their child's education, more than doubling impact at minimal additional cost. We explicitly measure practitioner prior and posterior beliefs, and find that rigorous testing facilitates more accurate identification of 'what works.' Our findings both reveal the returns to iterative testing in social programs and contribute new evidence on simple, cost-effective strategies to improve learning outcomes.

*We are grateful for funding and thought partnership from UBS Optimus Foundation, the Mulago Foundation, the Douglas B. Marshall Foundation, the Jacobs Foundation, and the Agency Fund. We are also grateful for strong technical and research collaborations and comments from colleagues at the What Works Hub for Global Education and the Center for the Study of African Economies at the University of Oxford as well as from J-PAL and IPA. We thank colleagues for comments at the American Economic Association Annual Meetings, the What Works Hub for Global Education conference, and the CSAE conference, and for useful conversations with Amrita Ahuja, Valmik Ahuja, Jenny Aker, Norma Altshuler, Tahir Andrabi, Susan Athey, Avery Bang, Michela Carlana, Stefan Dercon, Sharnic Djaker, Sasha Gallant, Paul Glewwe, James Habyarimana, Michelle Kaffenberger, Dean Karlan, Eliana La Ferrara, Axel Larrinaga, Clare Leaver, Susanna Loeb, Temina Madon, Rob On, Richard Sedlmayr, Kevin Starr, Caitlin Tulloch, and Paul Youn. We especially thank the current and former Youth Impact team, in particular the program management team (Colin Crossley, Moitshepi Matsheng, Sunshine Ntshambiwa, Tumisang Pifelo, and Amogeleng Klaas), the research team members leading ConnectEd monitoring and evaluation (Ayodotun Ayorinde, Sophie Ochmann, and Phodiso Mogotsi, and previous team members Bonno Balopi, Amy Jung, Lovemore Mawere, and Gaone Moetse), and research team members leading A/B testing efforts more broadly (Amanda Beatty and Konstantin Buchel). AEA registry: AEARCTR-0016323. Oxford University Ethics Review Ref.: 2212164.

I Introduction

Despite growing school enrollments worldwide, millions of children still do not acquire essential foundational skills in literacy and numeracy (Angrist et al., 2021; Ganimian and Djaker, 2022). Given the scale of this learning crisis, there is high demand from governments, international organizations, and non-government organizations for cost-effective education programs that can be scaled up.

A growing evidence base identifies effective educational programs to improve learning (Murnane and Ganimian, 2014). One of the most effective educational approaches is tutoring (Carlana and La Ferrara, 2025; Cortes et al., 2024; Fryer Jr, 2017; Kraft, Schueler and Falken, 2024; Nickow, Oreopoulos and Quan, 2020). Yet high costs have remained a barrier to scale (Kraft et al., 2022). Moreover, when effective programs such as tutoring are expanded at scale, effectiveness often decreases (Davis et al., 2017; Mobarak, 2022). This phenomenon is increasingly referred to as the "voltage drop" (List, 2022).

Iterative A/B testing can help address these scaling challenges, addressing both sides of the equation – reducing costs and enhancing effectiveness. A/B testing is often defined by randomized, rapid, and regular program optimization in the technology sector and can enable substantial efficiency improvements over time.[1] Technology companies like Google and Amazon run thousands of A/B tests monthly to optimize products at scale (Kohavi, Tang and Xu, 2020; Koning, Hasan and Chatterji, 2022). Yet such iterative experimentation remains rare in education and other social sectors, despite recent developments in adaptive experimentation and evidence-based innovation suggesting that regular randomized evaluations can result in substantial social returns (Athey et al., 2023; Kasy and Sautmann, 2021; Kremer, 2020).

In this paper, we conduct a dozen iterative randomized A/B tests of a numeracy tutoring program to optimize efficiency and scalability. We focus on a mobile phone tutoring program in line with efforts to experiment with technology-enabled approaches to lower the costs of tutoring at scale (Bhatt et al., 2024; Carlana and La Ferrara, 2025; Ganimian, Vegas and Hess, 2020; Gortazar, Hupkau and Roldán-Monés, 2024; Robinson et al., 2024; Zoido et al., 2024). Mobile phones are a particularly widespread and low-cost technology (Bergman and Chan, 2021), especially in lower income countries (Aker and Mbiti, 2010). Teachers deliver weekly numeracy tutorials to primary school children through phone calls. In addition to instruction via phone, short weekly assessments ensure instruction is targeted to each student's learning level: children who hadn't mastered addition were taught addition; those who knew addition progressed to subtraction, and so forth.[2] Earlier versions of the program have improved learning across RCTs in six countries —Botswana, India, Kenya, Nepal, Philippines, and Uganda— demonstrating both high internal and external validity (Angrist et al., 2023; Angrist, Bergman and Matsheng, 2022). With this strong foundation of evidence, we address frontier questions of efficiency and scalability in Botswana through a dozen rapid randomized A/B tests optimizing cost-effectiveness.

[1] Using the language in List (2024), each successive A/B test cycle can consider a more scalable 'Option C' model.

[2] This approach is aligned with growing evidence on the importance of targeted instruction (Angrist and Meager, 2023; Banerjee et al., 2017; Duflo, Kiessel and Lucas, 2024; Muralidharan, Singh and Ganimian, 2019).

We measure learning outcomes on identical foundational numeracy tests collected repeatedly every school term. This consistency in outcomes enables high comparability across tests. In addition, consequential, rapid outcomes – such as high-frequency learning indicators – enable iterative testing which can inform decision-making. A/B testing is often conducted in the technology sector using proxy indicators that are realized quickly (e.g. clicks on a website). When applying A/B testing in the education sector, we aimed to collect data on outcomes that change quickly enough to enable real-time optimization while also going beyond engagement and clicks to collect indicators connected to social impacts of interest. By collecting data on foundational numeracy skills – an outcome that is both consequential and can change quickly – across successive testing rounds, we leverage a unique opportunity to assess the relevance of A/B testing for social outcomes.

We conduct two types of A/B tests: cost-reducing tests and effectiveness-enhancing tests. Cost-reducing tests are rarely conducted in the social sciences, yet reducing costs is essential for programs to scale successfully (Al-Ubaydli, List and Suskind, 2017; List, 2024). In each test, we ask whether a lower-cost version can be as effective as the status quo program. These tests resemble non-inferiority tests in medicine which compare whether new treatments are 'just as good' as the status quo (Laster and Johnson, 2003). We systematically reduce costs across successive rounds of A/B tests. For example, we optimize scheduling efficiency, since substantial time and money is typically lost scheduling tutoring sessions. Holding total dosage constant, we compared longer sessions (40 minutes) every other week with shorter sessions (20 minutes) each week. This shift increased the share of tutor time devoted to instruction and reduced costly, labor-intensive weekly scheduling. In a related follow-up test, we vary whether rotating tutors is as effective as having the same tutor each week. If equally effective, this allows more flexible matching of tutors to households, lowering coordination and scheduling costs. Taken together, these cost-reducing tests aim to reduce labor-intensive time spent scheduling, one of the biggest costs, while preserving program impact.

In terms of effectiveness-enhancing tests, we examine various margins to enhance impact at low marginal cost. We explore three types of enhancements: tech add-ons, motivational nudges, and parent engagement. Technology add-ons and motivational nudges, such as additional whatsapp and SMS messages, are cheap additions with potentially high cost-effectiveness. Encouraging greater parent engagement also has promising potential to improve effectiveness at low cost. Correlations in the literature suggest more engaged parents can enable students to learn more. We test the effect of additional caregiver engagement through randomized encouragement in a set of A/B tests to establish causal effects.

Results show that seven of twelve modifications yielded significant efficiency gains, a rate of discovery of 58%; several tests yielded up to 30% efficiency gains each. In all cases, cost-reducing innovations streamlined implementation and reduced costs while maintaining learning outcomes. In terms of effectiveness-enhancing tests, the largest improvements came from caregiver engagement strategies that more than doubled program impact at low cost. This result demonstrates that parent engagement is a promising but underutilized lever for promoting education outcomes in low-resource settings, similar to results from high-income settings (Bergman, 2021).

We conduct a comprehensive cost-effectiveness analysis. Given a central contribution of A/B testing is returns to cost-effectiveness, we examine cost carefully and consider multiple cost categories. We incorporate both direct programmatic financial expenditures, as well as estimating time spent by tutors and administrators; we also consider opportunity costs for beneficiaries. Even when considering all of these costs, cost-effectiveness estimates reveal striking returns for successful modifications: caregiver engagement, for example, generates learning gains up to 65 standard deviations per $100—among the highest returns documented in the education literature.

Do A/B tests generate intuitive results that frontline experts could have predicted without the test, or do A/B tests reveal novel insights? We directly measure practitioner prior beliefs and find that program experts often underestimate the returns to modifications while overestimating the value of maintaining the status quo. For example, we find that frontline workers predict that a cost-reduction will result in lower impact. Yet our A/B tests find that cost-reducing tests can successfully reduce cost without a loss in impact. After observing rigorous A/B test results, implementers update their beliefs toward the true effects, demonstrating the value of experimental methods to complement expert intuition.

Our overall 'rate of discovery', which finds efficiency enhancements in 58% of the tests conducted (seven of twelve tests), compares favorably to technology sector benchmarks of 10-40%, demonstrating that systematic A/B testing can be highly effective in optimizing social programs. Moroever, the efficiency gains from cost-reductions and effectiveness enhancements not only averted the typical voltage drop when scaling social programs, but increased total effectiveness over time.

These results contribute to the literature on effective educational strategies, such as tutoring, which are highly effective but remain difficult to scale, in part due to high costs. We contribute evidence on efficiency improvements to lower costs and enhance the effectiveness of tutoring at scale (Carlana and La Ferrara, 2025; Kraft et al., 2022; Robinson et al., 2024). In addition, we contribute new evidence on another effective educational strategy – engaging parents in their child's educational instruction – a margin that has gained attention in high-income countries (Avvisati et al., 2014; Doepke, Sorrenti and Zilibotti, 2019; List, Pernaudet and Suskind, 2021; Ziege and Kalil, 2025) but remains underutilized in low-income countries (Angrist, Kabay, Karlan, Lau and Wong, 2025; Dizon-Ross, 2019). We show that parents can be effective conduits for education even in low-resource, low-literacy settings. By identifying more efficient approaches to improve learning, we also contribute to the broader evidence on addressing the learning crisis through cost-effective solutions (Angrist, Evans, Filmer, Glennerster, Rogers and Sabarwal, 2025; Global Education Evidence Advisory Panel, 2023).

Our findings further contribute to a growing 'science of scale' literature by demonstrating that A/B testing in the social sector can help mitigate and even reverse potential voltage drops as programs are expanded (Al-Ubaydli, List and Suskind, 2017; Davis et al., 2017; List, 2022; Mobarak, 2022). This study presents one of the first systematic applications of ongoing A/B testing program optimization — widely used in the technology sector — over multiple rounds of successive iterations to improve educational interventions and consequential social outcomes.

II Intervention Context and Modifications Tested

II.A The base intervention

We examine a foundational numeracy program that targets instruction to individual student learning levels, and leverages phone calls to reach students cheaply at home to provide low-cost tutoring. The program, shown to improve learning across contexts in multiple randomized controlled trials (Angrist et al., 2023; Angrist, Bergman and Matsheng, 2022), is now being scaled-up through government systems in Botswana, the Philippines, and India, among other settings.

The program was designed with two key features in mind: platform and pedagogy. In terms of platform, the program uses the widely accessible platform of mobile phones accessible to most households even in low-income countries (UN, 2023). In terms of pedagogy, the program targets instruction, adapting to each student's learning level through short high-frequency assessments conducted at the end of every tutoring call. Based on these assessments, instructors guide students through numeracy practice problems, helping them master key skills in addition, subtraction, multiplication, and division over the course of the term. For example, students who struggled with addition were taught addition, while those who had mastered addition but not subtraction were taught subtraction. Calls were made to caregivers who then invited the child to join, either handing them the phone or using speakerphone for the tutoring session.

A typical phone call followed the following format: the tutor called the household at a mutually agreed time, directing the conversation toward the student while encouraging caregivers to put the phone on speaker and be available for support. Tutors worked with the student at their identified learning level from the previous week, guiding them through simple steps to solve the operation being taught. After the student completed a series of problems with the tutor, the call ended with a 'checkpoint question' – a single math problem at the level taught that week. This checkpoint allowed the tutor to evaluate and update the student's learning level, ensuring the next week's instruction was targeted at the right level.

While earlier evidence suggests high internal and external validity for this approach (Angrist et al., 2023), initial proof-of-concepts often fail to scale when taken to new or larger settings (List, 2022; Mobarak, 2022). Thus, questions remain about scalability and efficiency, necessitating further program modification and optimization.

II.B Program modifications

We conduct 12 successive rounds of rapid randomized A/B testing, testing various program modifications. We developed and evaluated innovations specifically designed to address both sides of the scaling equation – reducing implementation costs as well as enhancing program effectiveness. Figure A.1 visualizes the cost-reducing program modifications for each respective A/B test; Figure A.2 visualizes the effectiveness-enhancing program modifications. Table 1 provides a brief summary of all A/B tests.

Panel A in Table 1 shows cost reduction adjustments, which aim to reduce costs without compromising learning outcomes. These modifications systematically targeted scheduling efficiency, one of the program's most time-consuming components and a large cost driver. We tested three innovations to improve scheduling efficiency. First, we adjusted the dosage distribution. In the original model, each tutoring call lasted 20 minutes, while scheduling these calls required roughly 30 minutes. To improve efficiency, in the modification, we shifted to a bi-weekly 40 minute call, maximizing instructional time once a household was reached, and minimizing time spent scheduling. A second set of tests to improve scheduling efficiency involved rotating different tutors each call for a given household rather than having consistent tutors each call (the business-as-usual model). If different tutors were equally effective, households could be matched with any available tutor, reducing scheduling frictions and costs. A final scheduling efficiency evaluated whether centralized call center allocations of available tutors to households, relative to more decentralized tutor-coordinated scheduling, could minimize scheduling inefficiency and maximize instructional time.

The program modifications in Panel B in Table 1 focus on enhancing program effectiveness at low marginal cost. We consider three effectiveness-enhancing categories. First, technology enhancements which added cheap complementary components that encouraged independent practice outside of tutoring sessions. These low-cost 'add-ons' included SMS messages with weekly math problems for independent practice, short WhatsApp video lessons reinforcing basic operations, and homework assignments discussed in subsequent calls. These additions costed only \$0.18-\$0.87 per child. Second, we test motivational nudges, designed to increase educational involvement through light-touch encouragement. These messages included inspirational testimonials from past program participants describing positive learning impacts and caregiver involvement strategies. Third, and finally, we included a series of program modifications designed to actively involve caregivers in tutoring sessions—for instance, encouraging caregivers to lead the second half of tutoring calls. By 'taking over' part-way through the session, caregivers could enable students to follow instructions more seamlessly than listening to instructions through the phone and to maximize time spent on practice problems. Moreover, by only taking over halfway through the call, this provided a scaffolding of support to parents in low-resource households where this type of educational engagement remains rare. Caregivers could first listen to the call as the tutor taught virtually; halfway through the call, caregivers started to teach directly, helping their child with practice problems based on concepts covered in the first half of the call by the tutor. This approach was designed to engage caregivers in their child's education in a low-resource context and overcome barriers to entry.

II.C Iterative A/B testing

While the social sector often struggles to scale effective programs, the technology industry has developed sophisticated methods for maintaining – and even improving – effectiveness at scale. Central to this success is a culture of continuous experimentation and optimization. Large technology companies routinely conduct thousands of A/B tests each month, enabling them to iteratively refine products and services based on rigorous evidence (Kohavi and Thomke, 2017). This experimental

mindset is deeply embedded in their operational practices, allowing for performance optimization at unprecedented scale (Azevedo et al., 2020; Siroker and Koomen, 2015).

Despite clear benefits, education and social sectors rarely employ systematic experimentation, missing opportunities to create more effective, sustainable programs at scale. Adaptive experimentation offers particular promise in education by accelerating learning and improving interventions over time through real-time, data-driven adjustments (Athey et al., 2023).

We characterize our application of iterative A/B testing using three 'R's: Randomized, Rapid, and Regular (Angrist et al., 2024). In terms of randomization, similar to the typical randomized trial, treatment groups are randomly allocated between group A and B, enabling identification of causal effects due to any modification in the program. In contrast to the typical randomized controlled trial, we don't include a pure control group, but rather compare program modifications to the status quo program without the modification. Thus, the question is not 'does the program work?' but rather 'is the modified program cheaper or more effective'? This primarily answers an efficiency question rather than an effectiveness question. A typical use case of A/B testing is when a program has already been verified to work relative to a pure control group in a full-scale randomized controlled trial, as we have done for this program in prior work (Angrist et al., 2023). With proof-of-concept established, the frontier questions become those about efficiency margins to facilitate implementation at scale. Of note, in some cases, effectiveness can still be inferred from an A/B test. For example, if B is better than A, as long as A does not generate negative effects, B is also better than zero. Thus, A/B testing can still be an option to infer effectiveness in some circumstances.

In terms of rapid cycles of learning, the typical A/B test aims to generate insights within weeks or months, rather than years. This speed facilitates high-frequency learning and ensures rigorous learning informs real-time decision-making. This is facilitated by having indicators that are both collected quickly and can change quickly. For example, in the technology industry, clicks on a website generate rapid results. A challenge, however, when collecting rapid indicators is ensuring that they are consequential enough to influence decision-makers. This is particularly challenging in the social sector where changing social outcomes can take time. We refer to these as 'golden indicators' – indicators that can change quickly enough to facilitate rapid learning while also being consequential enough to inform decision-making. In our setting, we collect data on numeracy skills using rapid learning outcome assessments, indicators that have been pressure-tested over years in order to facilitate a unique combination of speed and substance.

In terms of regular learning, iterative A/B testing is designed to be an ongoing optimization process rather than a 'single slice'. Each test should inform the next, leading to cumulative and compounding gains. To this end, we conduct multiple experiments sequentially rather than running only a single experiment or comparing all treatments at once in parallel. This enables program decision-makers to have time and space in between tests to respond to the results. This also enables each successive round of experiments to update A/B test questions in response to results from prior experiments, facilitating cumulative learning. This regular iteration produces continuous

optimization throughout the scaling process, which can help mitigate and even reverse voltage drops in effectiveness as programs scale.

III Sample and Data

The total study sample consists of 14,818 students in Botswana, primarily from the South East, Kgatleng, North East, and Kweneng regions (Figure A.3). We collaborated with government schools to distribute consent forms to parents and guardians of students in grades 3 to 6. Within each consenting household, we identified a primary school student and main caregiver responsible for educational support. Students received a brief baseline learning assessment to determine appropriate tutoring levels, ensuring instruction was tailored to individual needs from the program outset.

A/B tests were conducted over 12 successive school terms with preparation starting in mid 2020 and the trials running between January 2021 and December 2024 (see timeline in Figure A.4). These A/B tests were conducted by Youth Impact, which has run randomized trials since 2014 and has been regularly conducting A/B tests since 2018. Households were randomly assigned to status quo or modification arms. The median grade level is 4 and 52% of students are female. At baseline, student's math levels were well below grade-level expectations: 17% knew no basic operations, 21% could do addition only, 20% could do only up to subtraction, 27% could do up to multiplication, and only 15% could answer two-digit division problems. Among caregivers, 73% were mothers, 11% fathers, 8% grandparents, and 2% siblings. A set of descriptive statistics are reported in Table B.1.

Data collection involved baseline and endline surveys for each test, administered via 30-minute phone calls with children and caregivers, covering topics such as student learning outcomes, parental engagement, and treatment compliance. To ensure data integrity, tutors did not conduct baseline or endline assessments for their own students. Each test round lasted approximately 12 weeks with 3-week intervals between rounds. This rapid, iterative approach allowed for results to inform programming in real-time and enabled cumulative learning about the program's cost-effectiveness over multiple cycles of testing.

The learning assessment tested basic numeracy competencies measured in various high-profile studies in the education literature (Banerjee et al., 2017, 2007) and identified as core proficiencies in numerical operations by the UNESCO Global Proficiency Framework. The test included multiple numeracy items, such as problems involving two-digit addition, subtraction, multiplication, and division. In addition, we asked students to solve a place value word problem and a fraction problem to capture a broader range of mathematical understanding. To ensure the reliability of the assessment, we implemented a series of quality-assurance measures (Angrist et al., 2023).[3] These numeracy

[3]For example, to minimize the likelihood of family members assisting the child with the math problems, we implemented a two-minute time limit per question and asked each child to explain their work. A problem was marked correct only if the child accurately explained their solution and if instructors were confident that no parental assistance was provided. Angrist et al. (2023) show that this approach to measuring learning is reliable through a series of validity checks. For instance, they compared in-person with phone-based assessments for the exact same set of students and found no significant differences. They also tested a random subset of students twice on the same competencies to ensure consistency and randomized various problems of the same proficiency to test robustness.

measures enabled A/B testing with outcomes beyond day-to-day engagement data, enabling us to assess impacts on high frequency social outcomes.

IV Empirical Strategy

In each A/B test, households were randomly assigned either to the "status quo arm", which received the current program model, or to the "modification arm", where participants received a new program version designed to improve cost-effectiveness.

We exploit the random assignment of households to identify causal effects and quantify the impact of the different program innovations on learning outcomes. We estimate the intent-to-treat effect of being assigned to a treatment arm as follows:

$$Y_{ij} = \alpha + \beta_1 Modification_j + \gamma X_j + \delta_k + \epsilon_{ij} \tag{1}$$

where Y_{ij} is a learning outcome for individual i in household j. $Modification_j$ is a dummy variable which takes on the value 1 if a household has been assigned to the modification arm or zero if in the status quo arm. X_j denotes a vector of baseline control variables to enhance statistical power and precision.[4] δ_k refers to relevant strata such as geographic region. We assess impact on learning outcomes, both in raw units and standardized (standardized relative to status quo group standard deviations and centered at mean zero). We estimate round-by-round and pooled regressions using ordinary least squares (OLS) with robust standard errors. All standard errors are clustered at the household level, as the randomization was conducted at this level. In practice, this is equivalent to clustering at the student level, since each household had only one participating student.

We also assess heterogeneous treatment effects using the following interaction specification:

$$Y_{ij} = \alpha + \beta_1 Modification_j + \beta_2 M_j + \beta_3 Modification_j \times M_j + \gamma X_j + \delta_k + \epsilon_{ij} \tag{2}$$

where M_j is the moderator of interest. In addition to learning outcomes, our analysis incorporates alternative specifications that examine other key outcomes beyond learning, such as educational engagement and parental beliefs, to provide a broader understanding of the program's impact.

Due to randomization, we expect the status quo and modification arms to be balanced at baseline in expectation across studies. Appendix Table B.2 presents balance tests using baseline data for key demographics and learning. We find no statistically significant differences across multiple dimensions, including gender, age, and baseline learning. Additionally, Appendix Table B.3 shows that the endline response rate is high, at 71 percent, and attrition between baseline and endline is balanced across treatment groups. All tests had high compliance rates. In all arms, weekly monitoring data show high take-up (about 70 to 80 percent) and fidelity (80 to 98 percent of lessons were accurately targeted to student learning levels).

[4] These variables included: baseline learning levels, student age, and gender.

Additionally, for a subset of A/B tests, we estimate effects of caregiver engagement on the calls using a randomized encouragement design, since only around two-thirds of group B caregivers chose to take-up the encouragement to lead part of the tutoring call. We both calculate intention-to-treat effects as well as use an instrumental variable (IV) approach to estimate the impact of actual caregiver engagement on learning. We leverage random assignment to the encouragement treatment as an instrument for the degree of caregiver engagement in practice, following standard approaches (Angrist, Imbens and Rubin, 1996). The random assignment provides exogenous variation in caregiver engagement that is uncorrelated with other household characteristics, satisfying a key identifying assumption for instrumental variables estimation. We also plausibly satisfy the exclusion restriction, such that random encouragement of caregiver engagement in the tutoring call is unlikely to affect learning through any other channel but caregiver engagement. This is plausible since the treatment provides no additional resources or training which could otherwise affect learning. This approach estimates the local average treatment effect (LATE) among compliers – caregivers who would engage when encouraged but not otherwise.

We employ the following instrumental variables two-stage least squares (2SLS) estimation approach. We first estimate the first-stage effect of the random encouragement instrument on caregivers leading part of the tutoring call:

$$Engagement_j = \tau + \beta_1 Modification_j + \epsilon_{ij} \tag{3}$$

We then estimate the second-stage to identify the effect of caregivers leading the call on learning:

$$Learning_{ij} = \phi + \beta_1 \widehat{Engagement}_j + \epsilon_{ij} \tag{4}$$

where $Learning_{ij}$ is a learning outcome for individual i in household j, and $\widehat{Engagement}_j$ is the estimated share of caregivers leading the call from Equation (3).

V Results

V.A Cost-reducing tests

A central question for scaling is whether programs can be delivered more cheaply without sacrificing impact. Cost-reducing A/B tests examine whether removing or streamlining program elements lowers costs while preserving impact relative to the status quo. For these tests, the absence of significant learning differences alongside reduced implementation costs implies substantial efficiency gains.

In these tests, we focus on a key cost-reduction mechanism: improving the scheduling efficiency of tutoring sessions, which accounts for a large share of program labor costs. First, we tested the impact of shifting the dosage distribution of the tutoring calls from 20-minutes once a week to 40-minutes every other week – the same dosage, but distributed differently. The 40-minute biweekly model maximizes instructional time once a tutor reaches a household and reduces the need for costly

scheduling each week, which can otherwise occupy substantial tutor time. Second, we examined whether having a different tutor each week was as effective as the same tutor. While the same tutor might be more familiar to households, if different tutors can be as effective, this opens the door to more efficient scheduling, matching households to the first available tutor, and reducing time lost on scheduling. Third, we examined a centralized call center assignment of households to available tutors, again optimizing scheduling efficiency. Each of these tests aims to streamline labor costs in terms of time spent on scheduling, in turn enhancing efficiency and cost-effectiveness.

Table 2 and Figure 1 show no statistically significant learning differences between status quo and modification groups across all cost-reducing tests (columns 1-4). Results in Table 2 show effects on our main learning outcome of foundational numeracy skills measured using average student level. These null effects demonstrate that we can reduce high-cost operational elements without compromising program quality and impact. Results are robust to different estimation approaches, for instance, including baseline controls (columns 2 and 4).

We repeated the dosage distribution and implementer type tests in additional A/B testing rounds and pooled results to increase statistical power (column 5). Since cost-reducing tests are designed to detect small treatment differences, this can require repeat testing in order to have sufficient statistical power. After pooling, these tests are powered at the 80% level to detect differences of 0.1 standard deviations or greater, indicating that we are adequately powered to detect typical effect sizes (Rainey, 2024). Our repeat test results reinforce that null effects are not due to insufficient power; rather, we obtain relatively precise null differences between groups. These null results highlight the potential to substantially streamline program costs without a reduction in program effectiveness, identifying critical margins for efficiency gains. This typology of cost-reducing tests further reveals another margin where null results can be highly informative (Abadie, 2020).

V.B Effectiveness-enhancing tests

For effectiveness-enhancing tests, we examine if adding program elements can increase impact at low marginal cost relative to the status quo. For these tests, finding statistically significant learning differences between groups A and B at minimal marginal cost yields substantial efficiency gains.

We target various low-cost margins for increasing program effectiveness: providing additional content via low-cost tech such as SMS and WhatsApp; motivational nudges; and encouraging active caregiver engagement.

When providing additional content, one test involved sending supplementary video lessons through WhatsApp, while another involved sending math problems that students could complete on their own time. We also tested whether explicitly assigning problems as homework could encourage more practice outside of tutoring calls. Given the low cost of sending SMS and WhatsApp videos – in some cases just a few cents per implementation period per child – this approach represents a potentially cost-effective enhancement to the program.

When providing motivational nudges, the additional content consisted of short encouragement messages and testimonials from previous beneficiaries emphasizing the benefits of caregiver

participation.

Finally, in terms of active caregiver involvement, tutors in the treatment group requested caregivers to take over the tutoring call halfway through and lead a portion of the tutoring session by walking through and explaining a math problem to the child themselves. The marginal cost of this intervention is extremely low, as the average length of the tutoring call remains similar regardless of whether the instruction is delivered by the tutor or the caregiver, and most calls took place in the evenings or on weekends when caregivers were not working. The existing literature suggests increasing parental engagement in education could lead to improved learning outcomes; we test low-cost ways to do this in the context of an existing program and in a low-resource setting.

Table 3 and Figure 2 show the results of A/B tests aimed at enhancing program effectiveness. In panel A, column 1 shows that sending additional math problems via SMS above and beyond a phone call leads to a 0.12 standard deviation improvement in learning ($p = 0.11$), robust to adding baseline controls (column 2). Additionally, providing homework SMS messages and discussing solutions during the next call improved learning by 0.08 standard deviations ($p = 0.13$). Although only marginally statistically significant at conventional levels, given very low costs, these approaches could represent cost-effective enhancements to the base phone-call model, reinforcing the value of complementing calls with SMS components. In contrast, multiple other modifications show no effects on learning, including WhatsApp videos and motivational nudges to caregivers.

In panel C, results on active caregiver involvement innovations show the largest and most statistically significant learning differences between the status quo and modification groups. Columns 1 and 3 show that encouraging caregivers to actively engage and lead part of the tutoring calls produces substantial learning gains of 0.20 (p = 0.008) to 0.25 (p = 0.006) standard deviations. These effects remain robust to the inclusion of baseline controls (columns 2 and 4). To contextualize these results, the median effective education intervention produces only a 0.1 standard deviation gain in learning (Evans and Yuan, 2022), and over half of education interventions have no impact at all (Angrist, Evans, Filmer, Glennerster, Rogers and Sabarwal, 2025). These caregiver engagement effects are thus double those of a typical full education intervention. Moreover, these additional learning gains build on an already effective base intervention received by both treatment groups. Given that the marginal cost of this modification is extremely low—requiring no additional materials or technology — this represents a highly cost-effective educational innovation, demonstrating the substantial untapped potential of leveraging existing household support. We further explore the cost-effectiveness of this modification in later sections of the paper.

V.C Aggregate results

We pool results in Table 4 by high-level categories to assess the overall effectiveness of different types of A/B tests. On average, we find that cost-reducing innovations achieve substantial cost savings with no statistically detectable change in learning. This pooled result with a precisely estimated null effect confirms that cost reductions represent genuine efficiency gains rather than underpowered statistical tests.

We also find that effectiveness-enhancing tests generate significant learning improvements, with pooled results indicating an average gain of 0.09 standard deviations. This represents a 71% increase over the baseline effect of 0.12 standard deviations found in the original Botswana study (Angrist, Bergman and Matsheng, 2022), demonstrating that iterative optimization can substantially amplify program impact.

Taken together, these results demonstrate the power of systematic A/B testing in optimizing education interventions for cost-effectiveness and scalability. Continuous refinement enables the dual objectives of reducing implementation costs while enhancing educational effectiveness — a combination that challenges traditional assumptions about trade-offs between program quality and affordability.

V.D Heterogeneous treatment effects

Analysis of heterogeneous treatment effects reveals that program modifications benefit students similarly across key demographic and academic dimensions (Table 5). Neither cost-reducing nor effectiveness-enhancing modifications show differential impacts by gender (columns 1 and 3) or baseline learning level (columns 2 and 4). The absence of heterogeneous effects likely reflects the program's individualized design: one-on-one tutoring with adaptive instruction targeted to each student's specific learning level ensures that all participants can benefit regardless of their starting point or demographic characteristics. This is consistent with findings from earlier randomized controlled trials of this program (Angrist et al., 2023; Angrist, Bergman and Matsheng, 2022) and related targeted instruction studies (Muralidharan and Singh, 2025), where effects did not differ by student characteristics.

These findings demonstrate that iterative optimization can achieve equity goals as well as efficiency goals, delivering broad-based learning gains.

V.E A particularly high-return innovation: caregiver engagement

The results in panel C of Table 3 demonstrate the significant impact of caregiver engagement on children's learning outcomes. Encouraging caregivers to co-lead tutoring calls resulted in substantial learning improvements, with intention-to-treat (ITT) effects of up to 0.25 standard deviations ($p = 0.006$). These results demonstrate that simple modifications with extremely low marginal costs can more than double the impact of the base program tested in Botswana (Angrist, Bergman and Matsheng, 2022).

Given the striking effectiveness of this simple modification, we aim to understand the full extent of this modifications' impact. We estimate effects for caregivers who indeed *co-instructed*. Intention-to-treat effects capture average effects for all caregivers *encouraged* to co-instruct, but not for those caregivers who took up the encouragement to take over tutoring when prompted.

We estimate effects for those who co-lead instruction using instrumental variables analysis, leveraging random encouragement as an exogenous shock to co-leading the tutorial. Approximately

two-thirds of caregivers agreed to co-lead calls when encouraged. Results in Table 6 show that when indeed caregivers actively engage, this yields learning gains of 0.377 standard deviations (p = 0.006).

We explore potential mechanisms through which co-tutoring encouragement generates learning improvements. A primary mechanism appears to be effective direct instruction by caregivers. While caregivers in our setting are relatively low-literacy, often a reason cited for caregiver's inability to support their child's education in low-resource contexts, we find caregivers effectively engage in supporting instruction. Both the large first stage in Table 6 – based on caregivers reporting that they led tutorials – as well the share of tutors perceiving that caregivers were actively engaged in calls – as shown in columns (1) and (2) in Table 7 – confirm that the treatment led to large and substantial direct parental participation in their child's education. In addition to these large gains in direct instruction, we find that encouraging caregivers to co-lead tutoring calls can shift caregiver beliefs about both their child's capabilities and the value of mathematics education. Specifically, as shown in Table 7, caregivers in the treatment group reported an increase in their perception of their child's mathematical proficiency by 0.16 levels (similar to actual learning gains of their child) and a 6.8 percentage point increase in the share who believe mathematics is very important for children to learn, although these effects are only marginally statistically significant.[5]

These results suggest that when caregivers co-tutor their children, their beliefs about their child's performance and the value of education shift alongside their direct investments in education. This aligns with evidence that educational interventions can be particularly cost-effective when they influence caregiver beliefs through direct engagement (Dizon-Ross, 2019). The combination of hands-on tutoring experience and observable improvements in their child's problem-solving may create a cycle where increased parental engagement leads to better learning outcomes, which reinforces positive beliefs about education.

The large magnitude of these caregiver engagement effects underscores their practical importance. While parent engagement remains an underutilized lever in education programming, these findings demonstrate that actively involving caregivers in the educational process can generate large learning gains. The magnitude of the effects of caregiver active engagement — nearly four times larger than the typical successful intervention (Evans and Yuan, 2022) — suggests significant untapped potential for enhancing educational outcomes through simple, low-cost family involvement strategies.

These results also reveal the value of systematic innovation. Our findings align with Kremer et al. (2021), who demonstrate that just a handful of highly successful and cost-effective investments can pay for an entire portfolio of investments. While several of the innovations we tested showed modest or null effects, caregiver engagement generated returns at the highest end of the education literature, illustrating the value of systematic experimentation to identify breakthrough modifications.

[5] Caregivers most likely to co-tutor do not systematically differ from other caregivers by employment status or education level.

V.F Cost-effectiveness and efficiency gains to iterative innovation

To assess the cost-effectiveness of the various tutoring program modifications tested, we conducted a comprehensive cost analysis. We calculate efficiency gains using two complementary approaches designed to capture different types of improvements. First, for effectiveness-enhancing tests, we follow the literature (Dhaliwal et al., 2012; Kremer, Brannen and Glennerster, 2013) and conduct traditional cost-effectiveness analysis (CEA) in terms of standard deviation gains in learning per $100 as follows:

$$CEA = \Delta_{impact}/\Delta_{cost} * 100 \tag{5}$$

However, this equation does not capture efficiency gains for cost-reducing tests well. For example, for a successful cost reduction which achieves lower costs without compromising impact, the numerator in Equation (5) converges to zero (no change in impact). This generates a flooring effect, failing to capture any increased efficiency gains even as costs increasingly go down.

To address this limitation and consistently evaluate both types of innovations – cost-reducing and effectiveness-enhancing – we use an efficiency (e) gains metric that applies to both cost-reducing and effectiveness-enhancing tests as follows:

$$e = \Delta/\alpha \tag{6}$$

where Δ refers to a change either in impact or costs (depending on the type of test), and α refers to the baseline cost or impact in the status quo group.

In calculating cost, we drew from the literature in economics on cost-effectiveness (Dhaliwal et al., 2012), incorporating both financial and economic costs per child by treatment arm. Financial costs include labor, phone calls, data, SMS, baseline and endline assessments, organizational staffing (management, oversight, innovation), and modification-specific expenses. Economic costs capture household opportunity cost of a caregiver's time when being actively involved in the tutoring scheduling, calls, and homework.

We calculate efficiency gains using the metric from Equation (6) for all tests in Table 8, as well as conduct standard cost-effectiveness analysis as per Equation (5) for effectiveness-enhancing tests. Results demonstrate substantial returns to systematic experimentation. Seven of twelve tests generated significant efficiency gains, with cost-reducing innovations showing particularly high success rates. All cost-reducing tests achieved substantial efficiency improvements. The biweekly implementation model reduced costs by 11% while maintaining learning outcomes. Assigning the next available tutor, through leveraging different tutors and centrally assigned tutors, reduced scheduling costs by 5-10% without reducing educational impact. These efficiency gains stemmed from reducing the time tutors spent scheduling sessions — a time and labor-intensive component of the program. Because labor accounts for a large share of program costs, even small reductions in scheduling time translated into meaningful efficiency gains.

For effectiveness-enhancing tests, we include efficiency calculations for interventions where

p-values were marginally significant ($p < 0.15$). The SMS and homework add-ons, despite being marginal statistically significant, are extremely low cost and have the potential to be highly cost-effective. These types of innovations are relevant for practical decision-making, where low-cost interventions with potential benefits might warrant adoption even with marginal statistical significance at conventional thresholds.

Among effectiveness-enhancing tests, caregiver co-tutoring innovations achieved large efficiency gains at 22-30% and striking cost-effectiveness, yielding gains in learning of up to 65 standard deviation gains per $100 spent. This modification ranks among the most cost-effective educational interventions in the literature (Angrist, Evans, Filmer, Glennerster, Rogers and Sabarwal, 2025; Kremer, Brannen and Glennerster, 2013). This is due both to its large effectiveness and very low marginal cost. This degree of efficiency and cost-effectiveness gain is a standout feature of iterative A/B testing, which enables optimization exactly on these margins.

Overall, we find 58% (seven of twelve tests) of modifications tested yield efficiency gains, comparing favorably to technology sector benchmarks of 10-40% (Kohavi et al., 2020), demonstrating that iterative A/B testing can be highly effective in optimizing social programs. In addition, we find a higher 'hit-rate' for cost-reducing tests, with all cost reductions translating into some efficiency gains, although we find higher gains for effectiveness-enhancing modifications when they work.

V.G A/B tests improve on decision-makers' intuition and update beliefs

We assess the potential for A/B testing to inform and influence frontline decision-makers' beliefs. Do practitioners under-estimate the effects of program modifications? If yes, A/B testing can offer novel insight into the true effect of a program modification. Moreover, when results emerge, do practitioners respond to the results from rigorous learning? If yes, the decision-making return to A/B testing is particularly high.

We collected detailed data on tutor beliefs to understand the relevance of A/B tests to frontline decision-makers. We collect data on priors and posteriors. We then compare prior and posterior beliefs among practitioners to actual effects found in A/B tests.

Figure 3 shows practitioner predictions of modification impacts relative to the status quo. We focus on cost-reducing tests for which we collected the most detailed data on tutor beliefs. In terms of prior beliefs, we find that practitioners assume that cost-reduction modifications would reduce program effectiveness. This is natural and intuitive – removing program components is plausibly likely to reduce impacts. However, as indicated earlier in the paper, results show that cost-reductions did not result in lower impacts. The red dotted line reveals the true effect: cost-reduction modifications were no less effective than the status quo. Thus, practitioners seem to have systematically low prediction accuracy pre-A/B test and underestimate the returns to cost-reductions. These results on inaccurate predictions — where implementers underestimated modification benefits — are consistent with other literature showing that both experts and non-experts alike are poor forecasters of intervention effects (DellaVigna, Pope and Vivalt, 2019; Vivalt and Coville, 2023). This reveals the value of conducting A/B testing to identify efficiency gains that otherwise might

have been left on the table.

We also explore posterior beliefs after the A/B test. We find that implementers update their beliefs accurately post-intervention, moving closer to the true effects after observing results. This reveals that rigorous learning has high returns to frontline decision-makers and can successfully influence practitioner posterior beliefs.

Of note, systematic collection of practitioner beliefs can also help elicit which program modifications and A/B tests are worth trying in the first place. Bayesian theory predicts that when priors are weakly held (e.g., they are distributed more widely), these types of beliefs are more likely to be malleable and responsive to new information and evidence. Thus, in terms of influencing decision-making and increasing the likelihood of new evidence being used in practice, the best program modifications to test via iterative A/B testing are likely those where there is some uncertainty to begin with, and thus room for new information to influence decisions.[6]

Altogether, these results demonstrate two key values of systematic experimentation: generating new insights that challenge conventional wisdom, and equipping implementers with evidence-based knowledge on what works. These findings reinforce that rigorous, rapid, and regular experimentation can play an instrumental role in improving decision-making on the frontline of implementation.

VI Conclusion

This paper identifies concrete margins to enable tutoring programs – one of the most effective yet hard-to-scale approaches in education – to become substantially more efficient and scalable. These margins primarily include reducing labor-intensive costs and enhancing parental engagement in education. We further demonstrate the value of prioritizing cost-reducing tests alongside the more typical effectiveness-enhancing tests conducted in the social sciences. We also showcase the importance of explicitly quantifying practitioner prior and posterior beliefs to inform evidence production and decision-making.

This paper further demonstrates that iterative A/B testing can reverse the typical "voltage drop" associated with scaling social interventions, simultaneously reducing costs and enhancing effectiveness. Through 12 successive experiments of a tech-enabled tutoring program, we achieved up to an 11% reduction in program costs in a given round and improved learning outcomes by up to 30% in some tests. Most notably, caregiver engagement modifications generated learning gains of 0.20-0.25 standard deviations at very low marginal cost, effects multiple times larger than the median successful education intervention, with cost-effectiveness ratios reaching 65 standard deviation gains in learning per $100 invested. Our 'rate of discovery' – with 58% of innovations tested improving efficiency – exceeds technology sector benchmarks (10-40%), demonstrating that iterative experimentation has the potential to be as effective, if not more effective, in social sectors as in technology and commercial sectors.

[6] Initial uncertainty represents just one factor among several—including potential costs and benefits—that inform which questions are worth testing in an A/B test.

As governments and organizations seek cost-effective solutions to address global challenges, iterative A/B testing offers a powerful tool for maximizing impact while minimizing costs. More broadly, iterative A/B testing addresses persistent tensions between traditional impact evaluation and implementation realities by embedding rapid optimization experiments in local contexts within ongoing program delivery (Andrews, Pritchett and Woolcock, 2017). The substantial returns we document suggest that investing in ongoing systematic optimization of already proven and existing programs may yield higher social returns than developing entirely new interventions, shifting conventional approaches to scaling social programs.

References

Abadie, Alberto. 2020. "Statistical nonsignificance in empirical economics." American Economic Review: Insights 2(2):193–208.

Aker, Jenny C and Isaac M Mbiti. 2010. "Mobile phones and economic development in Africa." Journal of economic Perspectives 24(3):207–232.

Al-Ubaydli, Omar, John A List and Dana L Suskind. 2017. "What can we learn from experiments? Understanding the threats to the scalability of experimental results." American Economic Review 107(5):282–286.

Andrews, Matt, Lant Pritchett and Michael Woolcock. 2017. Building state capability: Evidence, analysis, action. Oxford University Press.

Angrist, Joshua D, Guido W Imbens and Donald B Rubin. 1996. "Identification of causal effects using instrumental variables." Journal of the American Statistical Association 91(434):444–455.

Angrist, Noam, Amanda Beatty, Claire Cullen and Moitshepi Matsheng. 2024. "A/B testing in education: rapid experimentation to optimise programme cost-effectiveness." What Works Hub for Global Education .

Angrist, Noam, David K Evans, Deon Filmer, Rachel Glennerster, Halsey Rogers and Shwetlena Sabarwal. 2025. "How to improve education outcomes most efficiently? A review of the evidence using a unified metric." Journal of Development Economics 172:103382.

Angrist, Noam, Micheal Ainomugisha, Sai Pramod Bathena, Peter Bergman, Colin Crossley, Claire Cullen, Thato Letsomo, Moitshepi Matsheng, Rene Marlon Panti and Shwetlena Sabarwal. 2023. "Building resilient education systems: Evidence from large-scale randomized trials in five countries." National Bureau of Economic Research .

Angrist, Noam, Peter Bergman and Moitshepi Matsheng. 2022. "Experimental evidence on learning using low-tech when school is out." Nature Human Behaviour 6(7):941–950.

Angrist, Noam and Rachael Meager. 2023. Implementation matters: Generalizing treatment effects in education. Blavatnik School of Government, University of Oxford.

Angrist, Noam, Sarah Kabay, Dean Karlan, Lincoln Lau and Kevin Wong. 2025. "Human Capital at Home: Evidence from a Randomized Evaluation in the Philippines." National Bureau of Economic Research .

Angrist, Noam, Simeon Djankov, Pinelopi K Goldberg and Harry A Patrinos. 2021. "Measuring human capital using global learning data." Nature 592(7854):403–408.

Athey, Susan, Katy Bergstrom, Vitor Hadad, Julian C Jamison, Berk Özler, Luca Parisotto and Julius Dohbit Sama. 2023. "Can personalized digital counseling improve consumer search for modern contraceptive methods?" Science Advances 9(40):eadg4420.

Avvisati, Francesco, Marc Gurgand, Nina Guyon and Eric Maurin. 2014. "Getting parents involved: A field experiment in deprived schools." Review of Economic Studies 81(1):57–83.

Azevedo, Eduardo M, Alex Deng, José Luis Montiel Olea, Justin Rao and E Glen Weyl. 2020. "A/b testing with fat tails." Journal of Political Economy 128(12):4614–000.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland and Michael Walton. 2017. "From proof of concept to scalable policies: Challenges and solutions, with an application." Journal of Economic Perspectives 31(4):73–102.

Banerjee, Abhijit V, Shawn Cole, Esther Duflo and Leigh Linden. 2007. "Remedying education: Evidence from two randomized experiments in India." The Quarterly Journal of Economics 122(3):1235–1264.

Bergman, Peter. 2021. "Parent-child information frictions and human capital investment: Evidence from a field experiment." Journal of political economy 129(1):286–322.

Bergman, Peter and Eric W Chan. 2021. "Leveraging parents through low-cost technology: The impact of high-frequency information on student achievement." Journal of Human Resources 56(1):125–158.

Bhatt, Monica P, Jonathan Guryan, Salman A Khan, Michael LaForest-Tucker and Bhavya Mishra. 2024. "Can technology facilitate scale? Evidence from a randomized evaluation of high dosage tutoring." National Bureau of Economic Research .

Carlana, Michela and Eliana La Ferrara. 2025. "Apart but connected: Online tutoring, cognitive outcomes, and soft skills." American Economic Review 115(10):3487–3513.

Cortes, Kalena, Karen Kortecamp, Susanna Loeb and Carly Robinson. 2024. "A scalable approach to high-impact tutoring for young readers: Results of a randomized controlled trial." National Bureau of Economic Research .

Davis, Jonathan MV, Jonathan Guryan, Kelly Hallberg and Jens Ludwig. 2017. "The economics of scale-up." National Bureau of Economic Research .

DellaVigna, Stefano, Devin Pope and Eva Vivalt. 2019. "Predict science to improve science." Science 366(6464):428–429.

Dhaliwal, Iqbal, Esther Duflo, Rachel Glennerster and Caitlin Tulloch. 2012. "Comparative cost-effectiveness analysis to inform policy in developing countries." Abdul Latif Jameel Poverty Action Lab, Massachusetts Institute of Technology, Cambridge, MA .

Dizon-Ross, Rebecca. 2019. "Parents' beliefs about their children's academic ability: Implications for educational investments." American Economic Review 109(8):2728–2765.

Doepke, Matthias, Giuseppe Sorrenti and Fabrizio Zilibotti. 2019. "The economics of parenting." Annual Review of Economics 11(1):55–84.

Duflo, Annie, Jessica Kiessel and Adrienne M Lucas. 2024. "Experimental Evidence on Four Policies to Increase Learning at Scale." The Economic Journal 134(661):1985–2008.

Evans, David K and Fei Yuan. 2022. "How big are effect sizes in international education studies?" Educational Evaluation and Policy Analysis 44(3):532–540.

Fryer Jr, Roland G. 2017. The production of human capital in developed countries: Evidence from 196 randomized field experiments. In Handbook of economic field experiments. Vol. 2 Elsevier pp. 95–322.

Ganimian, Alejandro J, Emiliana Vegas and Frederick M Hess. 2020. "Realizing the promise:, How can education technology improve learning for all?" The Brookings Institution, Center for Universal Education .

Ganimian, Alejandro J and Sharnic Djaker. 2022. How can developing countries address heterogeneity in students' preparation for school? A review of the challenge and potential solutions. Technical report Unpublished manuscript. Steinhardt School of Culture, Education, and Human Development, New York University. New York.

Global Education Evidence Advisory Panel. 2023. "2023 Cost-effective Approaches to Improve Global Learning: Recommendations of the Global Education Evidence Advisory Panel (GEEAP).".

Gortazar, Lucas, Claudia Hupkau and Antonio Roldán-Monés. 2024. "Online tutoring works: Experimental evidence from a program with vulnerable children." Journal of Public Economics 232:105082.

Kasy, Maximilian and Anja Sautmann. 2021. "Adaptive treatment assignment in experiments for policy choice." Econometrica 89(1):113–132.

Kohavi, Ron, Diane Tang and Ya Xu. 2020. Trustworthy online controlled experiments: A practical guide to a/b testing. Cambridge University Press.

Kohavi, Ron, Diane Tang, Ya Xu, Lars G Hemkens and John PA Ioannidis. 2020. "Online randomized controlled experiments at scale: lessons and extensions to medicine." Trials 21:1–9.

Kohavi, Ron and Stefan Thomke. 2017. "The surprising power of online experiments." Harvard Business Review 95(5):74–82.

Koning, Rembrand, Sharique Hasan and Aaron Chatterji. 2022. "Experimentation and start-up performance: Evidence from A/B testing." Management Science 68(9):6434–6453.

Kraft, Matthew A, Beth E Schueler and Grace Falken. 2024. What Impacts Should We Expect from Tutoring at Scale? Exploring Meta-Analytic Generalizability. EdWorkingPaper No. 24-1031. Technical report Annenberg Institute for School Reform at Brown University.

Kraft, Matthew A, John A List, Jeffrey A Livingston and Sally Sadoff. 2022. Online tutoring by college volunteers: Experimental evidence from a pilot program. In AEA Papers and Proceedings. Vol. 112 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 pp. 614–618.

Kremer, Michael. 2020. "Experimentation, innovation, and economics." American Economic Review 110(7):1974–1994.

Kremer, Michael, Conner Brannen and Rachel Glennerster. 2013. "The challenge of education and learning in the developing world." Science 340(6130):297–300.

Kremer, Michael, Sasha Gallant, Olga Rostapshova and Milan Thomas. 2021. "Is Development Economics a Good Investment? Evidence on scaling rate and social returns from USAID's innovation fund." University of Chicago Working Paper .

Laster, Larry L and Mary F Johnson. 2003. "Non-inferiority trials: the 'at least as good as' criterion." Statistics in Medicine 22(2):187–200.

List, John A. 2022. The voltage effect: How to make good ideas great and great ideas scale. Crown Currency.

List, John A. 2024. "Optimally generate policy-based evidence before scaling." Nature 626(7999):491–499.

List, John A, Julie Pernaudet and Dana L Suskind. 2021. "Shifting parental beliefs about child development to foster parental investments and improve school readiness outcomes." Nature communications 12(1):5765.

Mobarak, Ahmed Mushfiq. 2022. "Assessing social aid: the scale-up process needs evidence, too." Nature 609(7929):892–894.

Muralidharan, Karthik and Abhijeet Singh. 2025. "Adapting for scale: Experimental Evidence on Technology-aided Instruction in India." National Bureau of Economic Research .

Muralidharan, Karthik, Abhijeet Singh and Alejandro J Ganimian. 2019. "Disrupting education? Experimental evidence on technology-aided instruction in India." American Economic Review 109(4):1426–1460.

Murnane, Richard J and Alejandro Ganimian. 2014. "Improving educational outcomes in developing countries: Lessons from rigorous impact evaluations." NBER working paper (w20284).

Nickow, Andre, Philip Oreopoulos and Vincent Quan. 2020. "The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence.".

Rainey, Carlisle. 2024. "Power Rules: Practical Statistical Power Calculations.".

Robinson, Carly D, Cynthia Pollard, Sarah Novicoff, Sara White and Susanna Loeb. 2024. "The effects of virtual tutoring on young readers: Results from a randomized controlled trial." Educational Evaluation and Policy Analysis p. 01623737241288845.

Siroker, Dan and Pete Koomen. 2015. A/B testing: The most powerful way to turn clicks into customers. John Wiley & Sons.

UN. 2023. Measuring Digital Development: Facts and Figures 2023. Geneva: International Telecommunication Union.
URL: *https://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx*

Vivalt, Eva and Aidan Coville. 2023. "How do policymakers update their beliefs?" Journal of Development Economics 165:103121.

Ziege, Elena and Ariel Kalil. 2025. "How Information Affects Parents' Beliefs and Behavior: Evidence from First-Time Report Cards for German School Children." University of Chicago, Becker Friedman Institute for Economics Working Paper (2025-54).

Zoido, Pablo, Iván Flores-Ceceña, Miguel Székely, Felipe J Hevia and Eleno Castro. 2024. "Remote tutoring with low-tech means to accelerate learning: Evidence for El Salvador." Economics of Education Review 98:102506.

Table 1: Description of A/B tests and program modifications

Innovation Type	Version A Status Quo Model	Version B Modification
	Panel A: Cost-reducing tests	
Dosage Distribution	Weekly calls + SMS Weekly calls + SMS	Bi-weekly + SMS Bi-weekly + SMS
Implementer Type	Bi-weekly + SMS (same tutor) Bi-weekly + SMS (same tutor)	Bi-weekly + SMS (different tutor) Bi-weekly + SMS (different tutor)
Scheduling Assignment Mechanism	Bi-weekly + SMS (facilitator-led decentralized scheduling)	Bi-weekly + SMS (call center centralized scheduling)
	Panel B: Effectiveness-enhancing tests	
Tech Package: Add-ons	Bi-weekly calls Bi-weekly + SMS Bi-weekly + SMS	Inclusion of SMS Complementary WhatsApp videos Complementary homework SMS
Motivational Nudges	Bi-weekly + SMS Bi-weekly + SMS	Caregiver involvement nudge Alumni caregiver involvement testimonials
Caregiver Engagement	Bi-weekly + SMS Bi-weekly + SMS	Caregiver co-lead tutorial via encouragement Caregiver co-lead tutorial via encouragement

Note: This table summarizes A/B tests designed to reduce costs and improve program impact. Cost-reducing tests focus on operational efficiency while maintaining learning outcomes, including dosage distribution changes (shifting from weekly to bi-weekly calls), implementer flexibility (allowing any tutor vs. the same tutor to instruct a session across weeks), and scheduling mechanisms (decentralized vs. centralized call center assignment to improve operational efficiency). Effectiveness-enhancing tests aim to improve learning outcomes. Technology enhancements include SMS math problems and complementary video lessons designed to reinforce tutoring sessions. Motivational nudges encourage greater caregiver involvement through behavioral interventions, while active caregiver engagement tests leverage evidence that direct parental participation enhances educational outcomes.

Table 2: Effects of cost-reducing modified program vs. status quo model on learning

Panel A: Dosage Distribution

	Bi-weekly call		Bi-weekly call		Pooled
	(1)	(2)	(3)	(4)	(5)
	Learning (SD)	Learning (SD)	Learning (SD)	Learning (SD)	Learning (SD)
Version B Modification	-0.031	-0.040	-0.046	-0.052	-0.044
	(0.064)	(0.064)	(0.068)	(0.067)	(0.046)
	[0.627]	[0.536]	[0.502]	[0.440]	[0.339]
Observations	727	727	733	733	1460
Status Quo Group Mean	2.860	2.860	2.867	2.867	2.863
Controls	None	Bsl level	None	Bsl level	Bsl level

Panel B: Implementer Type

	Different tutors		Different tutors		Pooled
	(1)	(2)	(3)	(4)	(5)
	Learning (SD)	Learning (SD)	Learning (SD)	Learning (SD)	Learning (SD)
Version B Modification	-0.055	-0.058	-0.020	-0.006	-0.035
	(0.059)	(0.055)	(0.063)	(0.061)	(0.041)
	[0.345]	[0.292]	[0.754]	[0.915]	[0.396]
Observations	1193	1193	1017	1017	2210
Status Quo Group Mean	2.466	2.466	2.589	2.589	2.522
Controls	None	Bsl level	None	Bsl level	Bsl level

Panel C: Scheduling Assignment Mechanism

	Call Center Centralized	
	(1)	(2)
	Learning (SD)	Learning (SD)
Version B Modification	-0.010	-0.012
	(0.065)	(0.064)
	[0.884]	[0.845]
Observations	915	915
Status Quo Group Mean	2.538	2.538
Controls	None	Bsl level

Note: This table presents learning outcomes comparing the status quo and modified versions of the program across cost-reducing A/B tests. Columns (1) to (5) show the effect on learning for students who participated in the modified program. Learning is measured on a 0-4 scale where 0 indicates no operations correct, 1 indicates addition mastery, 2 indicates subtraction mastery, 3 indicates multiplication mastery, and 4 indicates division mastery. Effects are expressed in standard deviations, standardized relative to the status quo group at endline and centered at mean zero. Columns (2) and (4) include baseline controls. Column (5) pools results from repeated tests for robustness. Standard errors are in parentheses; p-values are in square brackets.

Table 3: Effects of effectiveness-enhancing modifications vs. status quo model on learning

Panel A: Tech Package Add-ons

	Value add of SMS		WhatsApp videos		Homework SMS	
	(1)	(2)	(3)	(4)	(5)	(6)
	Learning (SD)	Learning (SD)	Learning (SD)	Learning (SD)	Learning (SD)	Learning (SD)
Version B Modification	0.116	0.118	0.022	0.028	0.082	0.084
	(0.073)	(0.073)	(0.055)	(0.053)	(0.055)	(0.053)
	[0.114]	[0.106]	[0.694]	[0.596]	[0.134]	[0.116]
Observations	738	738	1173	1173	1232	1232
Status Quo Group Mean	2.663	2.663	2.728	2.728	2.351	2.351
Controls	None	Bsl level	None	Bsl level	None	Bsl level

Panel B: Motivational Nudges

	Caregiver nudge		Testimonials	
	(1)	(2)	(3)	(4)
	Learning (SD)	Learning (SD)	Learning (SD)	Learning (SD)
Version B Modification	0.053	0.046	0.006	0.003
	(0.067)	(0.067)	(0.069)	(0.069)
	[0.428]	[0.493]	[0.934]	[0.963]
Observations	739	739	754	754
Status Quo Group Mean	3.022	3.022	2.601	2.601
Controls	None	Bsl level	None	Bsl level

Panel C: Caregiver Engagement

	Caregiver co-leads		Caregiver co-leads	
	(1)	(2)	(3)	(4)
	Learning (SD)	Learning (SD)	Learning (SD)	Learning (SD)
Version B Modification	0.249	0.242	0.200	0.191
	(0.090)	(0.090)	(0.075)	(0.075)
	[0.006]	[0.007]	[0.008]	[0.011]
Observations	454	454	611	611
Status Quo Group Mean	2.625	2.625	2.605	2.605
Controls	None	Bsl level	None	Bsl level

Note: This table shows learning outcomes comparing the status quo and modified versions of the program across effectiveness-enhancing A/B tests. Columns (1) to (6) show the effect on learning for students who participated in the modified program. Learning is measured on a 0-4 scale where 0 indicates no operations correct, 1 indicates addition mastery, 2 indicates subtraction mastery, 3 indicates multiplication mastery, and 4 indicates division mastery. Effects are expressed in standard deviations, standardized relative to the status quo group at endline and centered at mean zero. Columns (2), (4) and (6) include baseline controls. Standard errors are in parentheses; p-values are in square brackets.

Table 4: Effects of modification on learning outcomes relative to status quo

Panel A: Cost-reducing A/B Tests

	(1) Learning (SD)	(2) Learning (SD)	(3) Avg Level	(4) Avg Level
Cost-reducing modification	-0.036	-0.036	-0.051	-0.049
	(0.028)	(0.028)	(0.038)	(0.038)
	[0.200]	[0.196]	[0.173]	[0.197]
Observations	4585	4585	4585	4585
Round FE	No	Yes	No	Yes

Panel B: Effectiveness-enhancing A/B Tests

	(1) Learning (SD)	(2) Learning (SD)	(3) Avg Level	(4) Avg Level
Effectiveness-enhancing modification	0.085	0.086	0.109	0.109
	(0.025)	(0.025)	(0.032)	(0.032)
	[0.001]	[0.001]	[0.001]	[0.001]
Observations	5701	5701	5701	5701
Round FE	No	Yes	No	Yes
Total Observations	10286	10286	10286	10286

Note: This table shows the effect on learning of modifications relative to the status quo. Panel A shows that cost-reducing modifications cause no statistically significant change on learning compared to the status quo model. Panel B shows that effectiveness-enhancing modifications improve learning outcomes on average. Learning is measured on a 0-4 scale where 0 indicates no operations correct, 1 indicates addition mastery, 2 indicates subtraction mastery, 3 indicates multiplication mastery, and 4 indicates division mastery. Effects are expressed in terms of average levels as well as standard deviations. Standard errors are in parentheses; p-values are in square brackets.

Table 5: Heterogeneous treatment effects of modified program

	Cost-reducing		**Effectiveness-enhancing**	
	Learning (SD)	Learning (SD)	Learning (SD)	Learning (SD)
Modification	-0.042	-0.080	0.094	0.057
	(0.042)	(0.054)	(0.037)	(0.049)
	[0.308]	[0.144]	[0.011]	[0.246]
Female	0.139		0.063	
	(0.039)		(0.036)	
	[0.000]		[0.084]	
Treatment=1 × Female	0.006		-0.020	
	(0.056)		(0.050)	
	[0.921]		[0.683]	
Bsl learning		0.161		0.128
		(0.015)		(0.014)
		[0.000]		[0.000]
Treatment=1 × Bsl learning		0.025		0.013
		(0.021)		(0.019)
		[0.236]		[0.486]
Observations	4580	4483	5701	5530
Round Fixed Effects	Yes	Yes	Yes	Yes
Controls	Bsl level	No	Bsl level	No

Note: This table shows that modifications based on reducing costs and enhancing effectiveness do not differentially impact girls and boys (columns (1) and (3)). Impact also does not vary by baseline student level (columns (2) and (4)). Standard errors are in parentheses. P-values are in brackets.

Table 6: Effect of encouraging caregivers to co-lead calls on learning outcomes

	(1) ITT	(2) First Stage	(3) LATE
Version B Modification	0.249	0.659	
	(0.090)	(0.034)	
	[0.006]	[0.000]	
Caregiver led part of tutoring call			0.377
			(0.138)
			[0.006]
Observations	454	454	454
Status Quo Group Mean	2.625	0.067	
Controls	No	No	No

Note: This table presents the results of caregiver's self-reported engagement on learning outcomes. Students in the status quo group of the caregiver co-leading A/B test (Test 6) received bi-weekly tutoring phone calls. In the modified version, caregivers were encouraged to lead part of the tutoring call. Learning is measured on a 0-4 scale where 0 indicates no operations correct, 1 indicates addition mastery, 2 indicates subtraction mastery, 3 indicates multiplication mastery, and 4 indicates division mastery. Effects are expressed in standard deviations, standardized relative to the status quo group at endline and centered at mean zero. Standard errors are in parentheses; p-values are in square brackets. Column (1) shows the intent-to-treat (ITT) effect on learning. Columns (2) and (3) report the first-stage and Local Average Treatment Effect (LATE) effects for caregivers who co-instruct. The mean in column (1) is the average learning among students who received the status quo program. The mean in column (2) is the average share of caregivers in the status quo group who co-led at least one of the tutoring calls.

Table 7: Effect of encouraging caregivers to co-lead calls on beliefs

	Perceived caregiver engagement		Perceived child's level		Math very impt	
	(1)	(2)	(3)	(4)	(5)	(6)
Version B Modification	0.810	0.811	0.163	0.158	0.064	0.068
	(0.021)	(0.022)	(0.101)	(0.101)	(0.040)	(0.040)
	[0.000]	[0.000]	[0.106]	[0.117]	[0.106]	[0.088]
Observations	1582	1566	1121	1119	1156	1154
Status Quo Group Mean	0.000	0.000	2.700	2.700	0.719	0.719
Round FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes	No	Yes

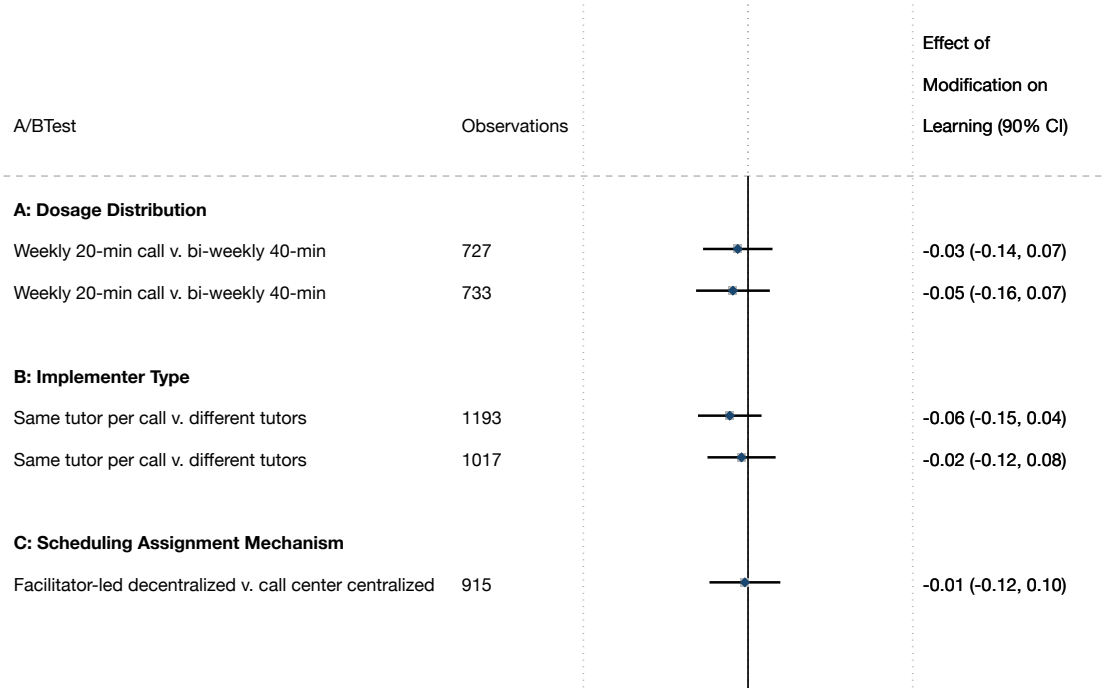
Note: This table shows the intention to treat effects of encouraging caregiver co-tutoring on three outcomes: perceived caregiver active engagement during the call, caregivers' understanding of their child's learning level, and their belief that math is very important for children to learn. The active engagement outcome is a binary variable equal to 1 if facilitators perceived that caregivers co-led at least one tutoring call during the intervention period, and 0 if they never co-led a call. This variable uses implementation data, so we have almost the full sample, as opposed to the endline-only sample for other outcomes. The perceived level variable is coded as 0 if the caregiver believes the child cannot perform any numeracy operations, 1 for addition, 2 for subtraction, 3 for multiplication, and 4 for division. The belief about math is a binary variable equal to 1 if a caregiver believes math is very important. Each regression is estimated without controls (Columns 1, 3 and 5) and with controls (Columns 2, 4, and 6). Controls include student gender, age, and baseline level.

Table 8: Efficiency Gains of A/B Test Interventions

Innovation	(1) Learning Difference (SD) across Arms	(2) Cost Difference per Child ($)	(3) Learning Gains per $100 (SD)	(4) Efficiency Gains (%)
Panel A: Cost-Reducing Tests				
Dosage Distribution				**Cost Savings**
Weekly vs. Biweekly	-0.03	-$3.57	—	0.11
	[0.627]			
Weekly vs. Biweekly	-0.05	-$3.57	—	0.11
	[0.502]			
Implementer Type				
Same vs. Different Tutor	-0.06	-$1.50	—	0.05
	[0.345]			
Same vs. Different Tutor	-0.02	-$1.50	—	0.05
	[0.754]			
Scheduling Assignment Mechanism				
Appointments vs. Call Centre	-0.01	-$2.82	—	0.10
	[0.884]			
Panel B: Effectiveness-Enhancing Tests				
Tech Package: Add-ons				**Impact Gains**
Call with SMS or not	0.116	$0.18	64.4	0.32
	[0.114]			
Whatsapp Videos	0.022	$0.87	0.0	0.00
	[0.694]			
Homework Assigned	0.082	$0.71	11.5	0.29
	[0.134]			
Motivational Nudges				
Caregiver Nudge	0.053	$0.04	0.0	0.00
	[0.428]			
Testimonials	0.006	$0.13	0.0	0.00
	[0.934]			
Caregiver Engagement				
Caregiver Co-tutors	0.249	$0.38	65.7	0.30
	[0.006]			
Caregiver Co-tutors	0.200	$0.47	42.4	0.22
	[0.008]			

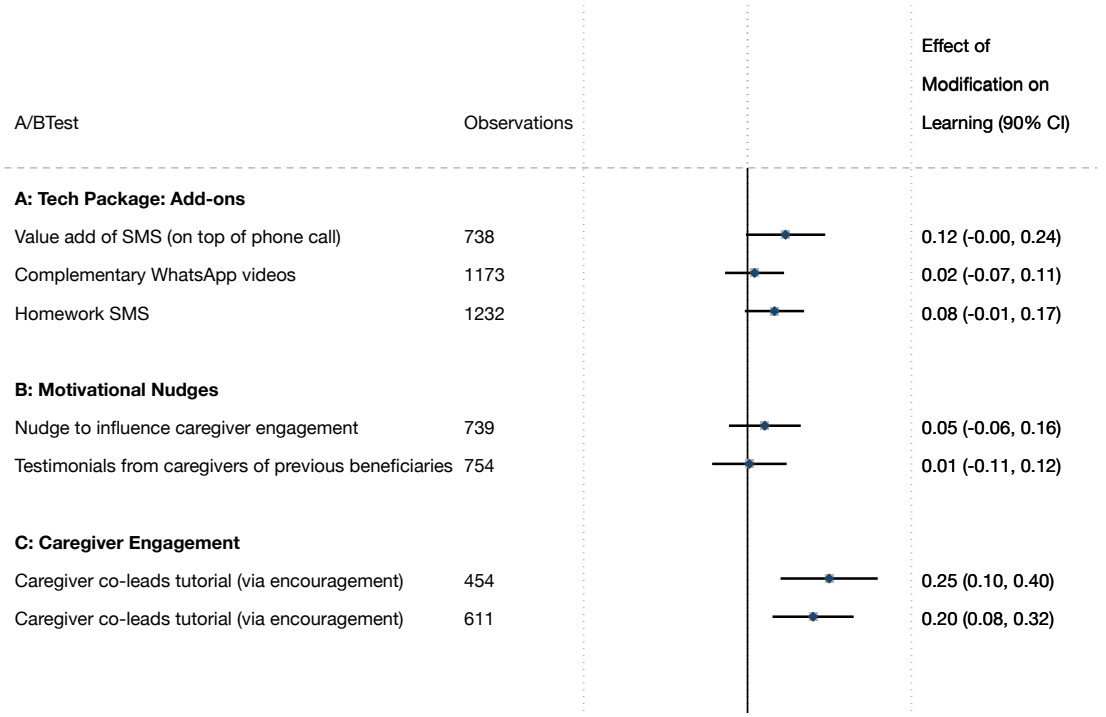
Note: This table shows cost-effectiveness analysis and efficiency gains across A/B test rounds. Cost calculations incorporate both financial costs (from program budgets) and economic costs (opportunity costs for caregivers valued at Botswana's mean hourly wage, and estimated using program data on uptake and time allocation, as well as assumptions on caregivers' opportunity costs from engaging in the call). Learning effects are expressed in standard deviations comparing modification and status quo groups at endline. Column 1 shows standardized learning differences between arms. Column 2 shows per child cost differences. Column 3 shows cost-effectiveness (additional standard deviations gained from the innovation, per $100). Column 4 shows efficiency gains between test arms. Cost-effectiveness is not applicable for cost-reducing tests since they maintain learning while reducing costs. Note that for the two effectiveness-enhancing tests with p-values between 0.1-0.15 (SMS and homework add-ons), we include cost-effectiveness calculations given their extremely low costs and potential benefits despite marginal statistical significance. Each round's cost and efficiency calculations represent marginal gains relative to that test's status quo, rather than absolute program efficiency, to avoid distortions from round-specific variation in status quo costs and impacts.

Figure 1: Effects of cost-reducing program modifications vs. status quo model



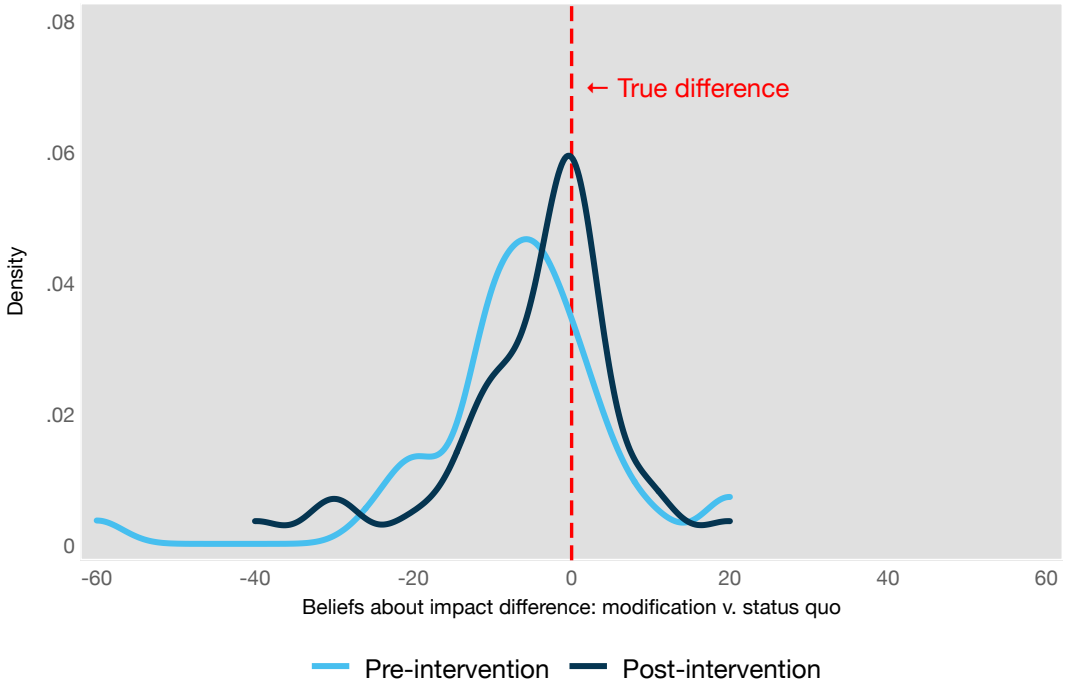
Note: This figure shows regression results from cost-reducing A/B tests. Learning outcomes are measured on a 0-4 scale representing mastery of basic numeracy operations (no operations, addition, subtraction, multiplication, division). The effect plot shows differences in learning between the status quo and modified program versions, expressed in standard deviations relative to the status quo group for each round.

Figure 2: Effects of effectiveness-enhancing program modifications vs. status quo model



Note: This figure shows regression results from effectiveness-enhancing A/B tests on the phone-based tutoring program. Learning is measured on a 0-4 scale where 0 indicates no operations correct, 1 indicates addition mastery, 2 indicates subtraction mastery, 3 indicates multiplication mastery, and 4 indicates division mastery. The effect plot shows differences in learning outcomes between the status quo and modified program versions, expressed in standard deviations, standardized relative to the status quo group at endline and centered at mean zero.

Figure 3: Teacher beliefs on program modification update more accurately post A/B test



Note: This graph shows the distribution of beliefs about the impact difference between the modified program and the status quo model. The light blue line represents implementers' beliefs before the intervention, while the dark blue line reflects their beliefs after the endline assessment. The dotted red line shows the true impact. Beliefs shift over time, aligning more closely with the actual difference in impact.

Appendices

A Appendix Figures

Figure A.1 illustrates A/B tests designed to streamline program elements to reduce costs while maintaining impact relative to the status quo. These tests specifically aimed to optimize operational efficiency, maximize instructional time, and minimize costs associated with administrative tasks such as scheduling.

"Dosage Distribution" tests compared the original weekly 20-minute call model with biweekly 40-minute sessions. The modified approach maintained identical total instruction time while reducing scheduling frequency.

The "Implementer Type" Tests evaluated learning outcomes when students received tutoring from the same tutor throughout the program versus different tutors across sessions. Both models tested whether tutor consistency was necessary for educational effectiveness or whether scheduling flexibility could be gained without learning losses.

"Scheduling Assignment Mechanism" tests compared decentralized scheduling (tutors directly coordinating with families) against centralized call center assignment (tutors provided predetermined household lists during assigned shifts). In Group A, tutors independently scheduled calls and coordinated with parents to find a mutually agreeable time through the week. Group B students were called by tutors working specific shifts, with each tutor provided a list of households to contact during their assigned shift.

Figure A.1: Cost-reducing A/B tests

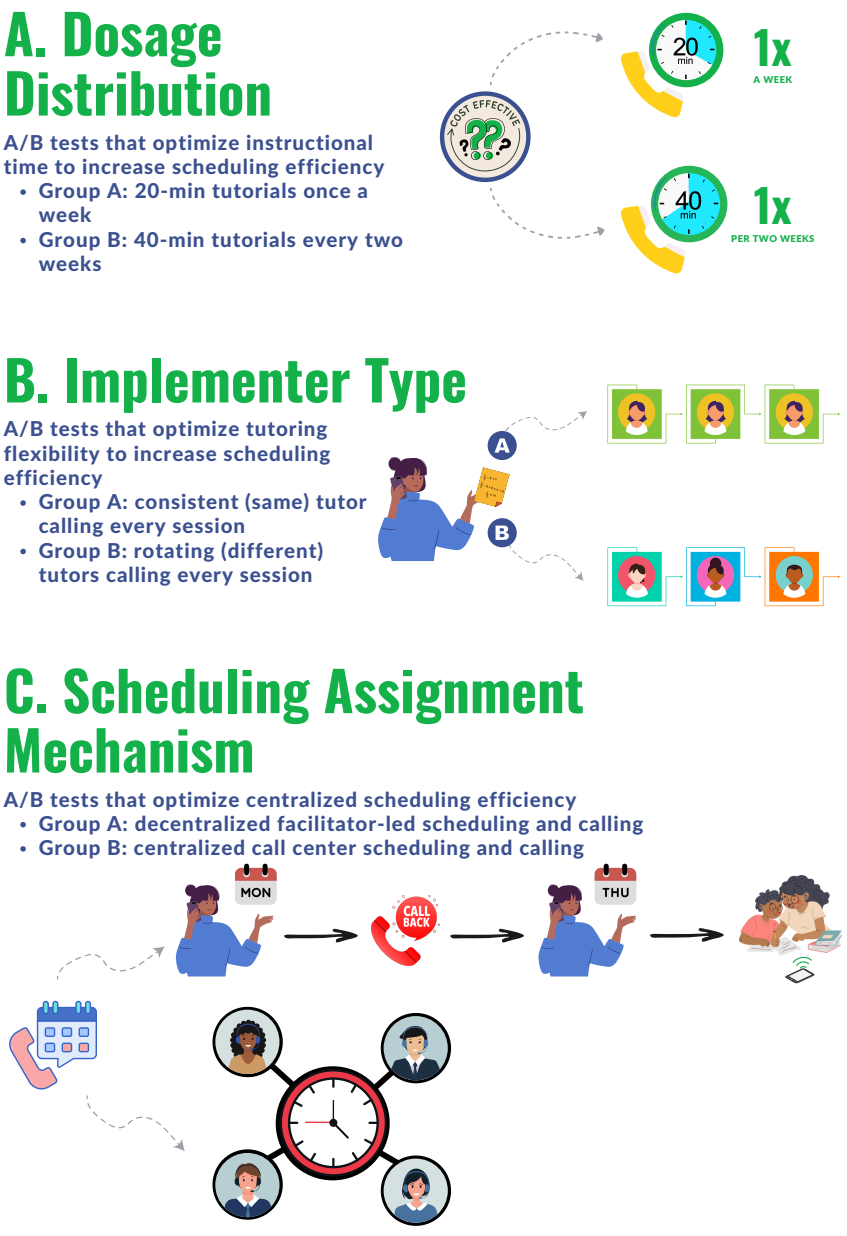


Figure A.2 illustrates A/B tests designed to optimize program effectiveness by maximizing impact at low marginal cost.

The "tech package add-on" tests examined three digital components additional to the core tutoring program. In the SMS test, Group B received the same tutoring calls as Group A plus weekly SMS messages with math problems for independent practice. In the WhatsApp test, Group B additionally received short video lessons demonstrating basic operations. In the homework test, Group A received the standard program while Group B received structured homework assignments with solutions discussed in subsequent tutoring sessions.

The "motivational nudge" tests aim to increase educational involvement through light-touch motivational messages. For example, caregivers in Group B also received testimonials from other caregivers, detailing how their involvement positively influenced their child's learning and how they continued to support their child's education using strategies learned during the tutoring calls.

The "caregiver engagement" tests encouraged active parental participation in tutoring sessions. In these tests, Group A received standard tutoring calls led by a tutor while Group B caregivers were explicitly encouraged to co-lead the second half of each call by teaching math problems to their children based on concepts and strategies introduced in the first half of the call.

Figure A.2: Effectiveness-enhancing A/B tests

A. Tech Package Add-ons

A/B tests that optimize program effectiveness by adding components to encourage additional practice

- **Test 1: Group B: status quo + SMS with math problems**
- **Test 2: Group B: status quo + complementary video lessons via WhatsApp**
- **Test 3: Group B: status quo + homework**

B. Motivational Nudges

A/B tests that optimize caregiver involvement through encouragement messages

- **Test 1: Group B: status quo + encouragement message**
- **Test 2: Group B: status quo + testimonials from caregivers of previous participants**

My child could not do operations easily. The tutor taught my child maths over the phone and asked me to also teach my child. We taught my child together and in the end, my child improved. I still help my child at home as the tutor showed me and he is getting better each day. Thank you, ConnectEd.

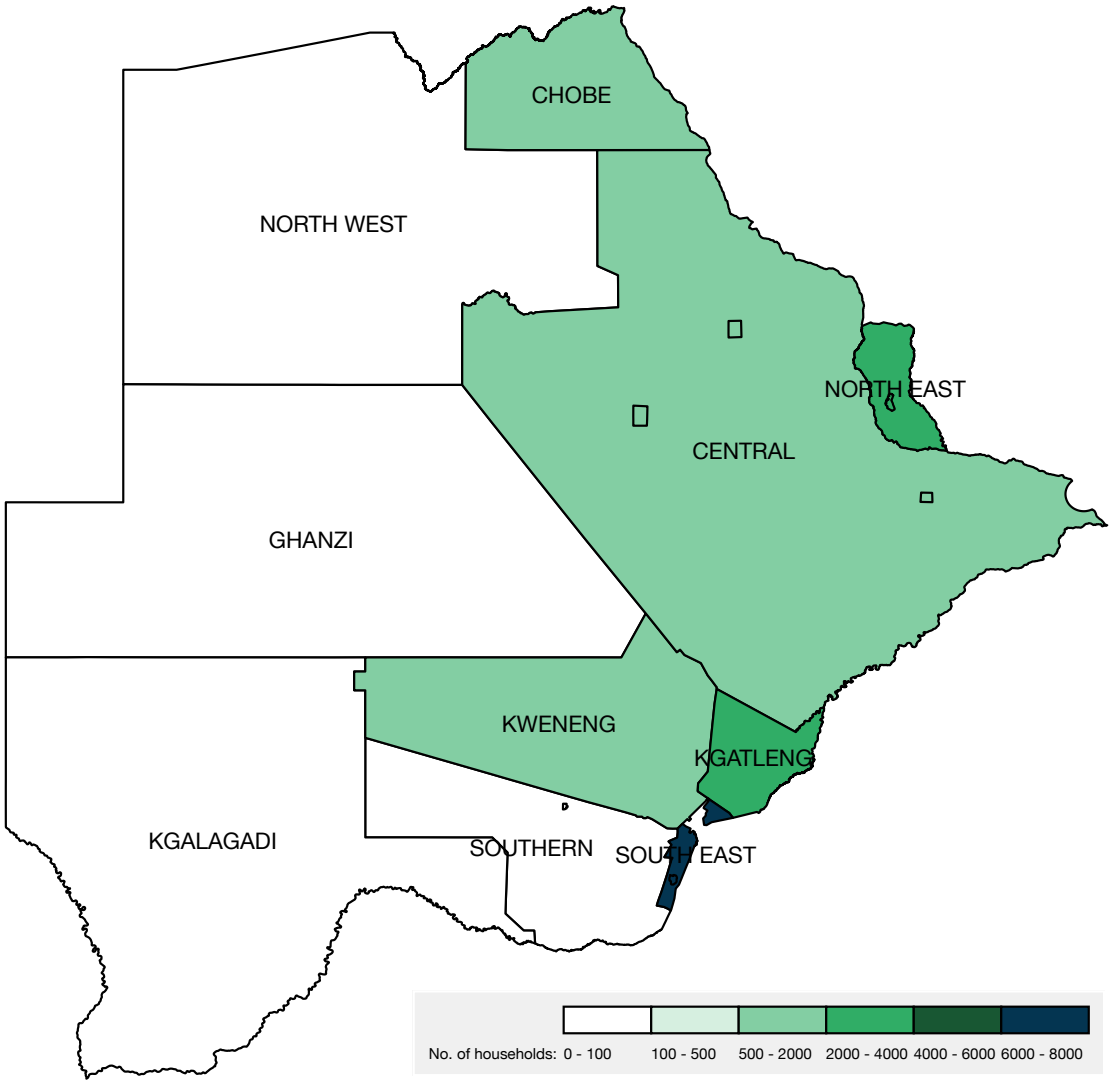
★★★★★

C. Caregiver Engagement

A/B tests that optimize active caregiver involvement through encouragement

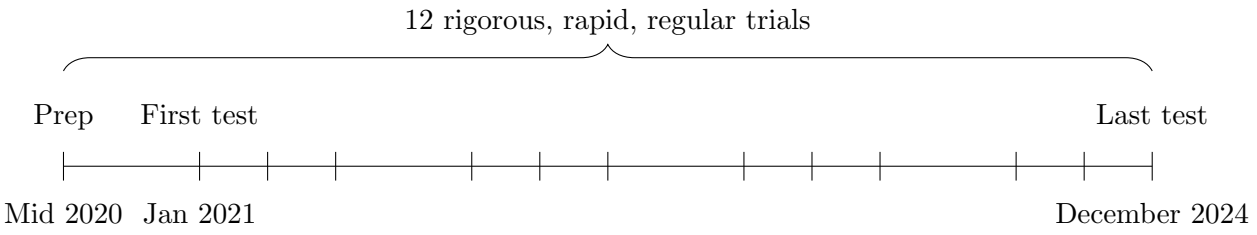
- **Test 1: Group B: status quo + encouraged caregiver to lead part of the tutorial**
- **Test 2: Group B: status quo + encouraged caregiver to lead part of the tutorial**

Figure A.3: Program coverage: households reached by region



Note: This map shows the number of households participating in the numeracy tutoring program across different regions in Botswana from 2021 to 2024. The tests cover 6 out of 10 regions and include the most populous regions in the country.

Figure A.4: Timeline of A/B test trials



Notes: This figure shows the timing of A/B tests. Each tick mark represents a term and round of A/B testing.

B Appendix Tables

Table B.1: Descriptive statistics, by A/B test

		Mean			
A/B Test	Obs	Age	Standard	Female	Bsl Level
T1 2021	1218	9.84	4.40	0.53	2.29
T2 2021	1075	10.01	4.39	0.54	2.31
T3 2021	1042	10.17	4.38	0.52	2.39
T1 2022	670	9.53	4.15	0.53	2.24
T2 2022	1095	9.61	4.14	0.48	2.02
T3 2022	912	9.42	3.70	0.50	1.92
T1 2023	1157	9.28	4.00	0.50	1.65
T2 2023	1684	9.58	4.12	0.52	2.04
T3 2023	1623	9.65	3.99	0.52	1.81
T1 2024	1319	10.18	4.95	0.54	2.07
T2 2024	1527	9.61	4.20	0.55	1.88
T3 2024	1496	9.80	4.15	0.48	2.09

Note: "Obs" refers to the number of students enrolled in the program. "Standard" indicates the average school grade level. "Female" is the proportion of girls in the sample. "Bsl Level" represents the mean level mastered at baseline. Baseline level takes the value of 0 if the student is a beginner (i.e., student who got no operations correct in the assessment, 1 if the student mastered addition, 2 for subtraction, 3 for multiplication, and 4 for division.) Across test rounds, means are similar.

Table B.2: Balance on baseline characteristics

Variable	(1) Group A Mean/SE	(2) Group B Mean/SE	(3) Difference
Standard 3	0.224	0.220	-0.004
	(0.417)	(0.414)	(0.007)
Standard 4	0.389	0.387	-0.002
	(0.488)	(0.487)	(0.008)
Standard 5	0.342	0.345	0.003
	(0.474)	(0.475)	(0.008)
Standard 6	0.044	0.048	0.004
	(0.206)	(0.215)	(0.003)
Age	9.716	9.730	0.014
	(1.149)	(1.146)	(0.019)
Sex (0=Male, 1=Female)	0.510	0.523	0.013
	(0.500)	(0.500)	(0.008)
Disability Status	0.044	0.046	0.001
	(0.206)	(0.209)	(0.005)
Baseline Learning	2.028	2.032	0.005
	(1.337)	(1.330)	(0.022)
Observations	7,256	7,229	14,818

Note: This table reports balance of baseline characteristics across the two A/B testing groups. The variables shown were collected at baseline from all 12 A/B tests. The baseline learning variable is coded as follows: 0 for students who got no operations correct in the assessment, 1 for those who got addition correct, 2 for subtraction, 3 for multiplication, and 4 for division. The covariate variable "round" is included in all estimation regressions. Significance levels are indicated by ***, **, and *, corresponding for 1%, 5%, and 10% p-value thresholds, respectively. Standard errors are in parentheses.

Table B.3: Response rate across treatments

Variable	(1) Group A Mean/SE	(2) Group B Mean/SE	(3) Difference
Reached/Student Levelled - Endline	0.705	0.716	0.011
	(0.456)	(0.451)	(0.007)
Observations	7,256	7,229	14,818

Note: This table demonstrates that attrition between baseline and endline is balanced across the two A/B testing groups. The covariate variable "round" is included in all estimation regressions. Significance levels are indicated by ***, **, and * for 1%, 5%, and 10% p-value thresholds, respectively. Standard errors are in parentheses.

What Works Hub
for Global Education

www.wwhge.org
wwhge@bsg.ox.ac.uk