

Mobile education builds resilience during shocks in five countries

<https://doi.org/10.1038/s41586-026-10990-x>

Received: 12 July 2025

Accepted: 4 August 2026

Published online: 9 September 2026

Open access

 Check for updates

Noam Angrist^{1,2,3}, Micheal Ainomugisha^{4,10}, Sai Pramod Bathena⁵, Peter Bergman⁶, Colin Crossley³, Claire Cullen^{2,3}, Thato Letsomo³, Moitshepi Matsheng³, Rene Marlon Panti⁷, Shwetlena Sabarwal⁸ & Tim Sullivan⁹

Education systems need to withstand frequent shocks, including disease and climate events, which close schools for more than 222 million children¹. During these emergencies, alternative models are needed to deliver education. However, rigorous evaluation of effective educational approaches in these settings is challenging and rare, especially across multiple countries. Here we present results from five randomized trials in India, Kenya, Nepal, the Philippines and Uganda to evaluate the provision of education in emergency settings. We test multiple scalable models of remote targeted tutoring instruction, comparing delivery by governments and non-governmental organizations (NGOs). Whereas text messages have mixed results, tutorials conducted via phone call show effectiveness across diverse settings. We find large and robust effect sizes on learning. These effects are highly cost-effective, and deliver approximately 4 years of high-quality instruction per US\$100 spent. Results show large effects when delivered by government teachers or NGO instructors. Together, our results reveal that it is possible to strengthen the resilience of education systems, enabling education provision amid disruptions, and to deliver cost-effective learning gains across contexts and with governments.

More than 2 billion people live in countries that are affected by emergencies that frequently disrupt education. Although they are frequent and disruptive, education emergencies have historically been understudied. Here we present a new database that documents just how frequent and disruptive such shocks can be. Figure 1 documents the extent of school disruption for a selected set of emergencies over the past two decades.

Schools close for long periods during these emergencies, and learning loss can be substantial^{2–5}. Education Cannot Wait, the United Nations' Global Fund for education in emergencies, estimates that 222 million children experience regular schooling disruptions and are in active need of education in emergencies programmes¹. In low-income countries, students lose on average 10% of the academic year owing to climate shocks alone⁶. Moreover, school breaks, although they are not emergencies, are another regular school disruption, and often results in substantial learning loss among low-resource households^{7,8}. These shocks exacerbate a pre-existing learning crisis, leaving many students well behind grade-level expectations⁹. Fewer than half of primary-age students in low- and middle-income countries are able to read a story or perform two-digit mathematical operations¹⁰.

Resilient education systems need to promote learning across settings and to withstand frequent disruptions. Ideally, interventions that promote learning in emergencies should be low-cost, simple to implement and quick to deploy. They should also yield high take-up across different geographies and implementation models and should

be targeted to children with diverse educational and cultural backgrounds. Understanding which approaches can be effective across these circumstances requires multi-context, multi-model studies. However, rigorous evaluation of approaches to deliver education in emergencies remains challenging and rare, especially across contexts and with governments.

Here we evaluate a set of education in emergencies programmes to promote learning during large-scale school disruptions, which affected more than 1 billion children worldwide. We conducted five randomized controlled trials, including multiple delivery models, such as NGO teacher aides and government teachers, to test scalability by government systems. Our trials were conducted across five countries: India, Kenya, Nepal, Uganda and the Philippines. In all settings, schooling was disrupted owing to the COVID-19 pandemic, and several of the countries in our study experienced some of the longest school closures in the world.

We use mobile phones to provide various educational interventions to primary school children. Mobile phones provide a platform that can reach households cheaply at scale in low-resource contexts¹¹. Whereas less than 15% of households have access to the internet in low-income countries, more than 70% have access to mobile phones¹². Moreover, mobile phones enable teachers to reach students at home even when school is disrupted, providing a resilient and flexible modality to provide education during emergencies. Mobile phones are increasingly relied on to provide aid via mobile money during humanitarian crises¹³, yet

¹Blavatnik School of Government, University of Oxford, Oxford, UK. ²What Works Hub for Global Education, Oxford, UK. ³Youth Impact, Gaborone, Botswana. ⁴Building Tomorrow, Kampala, Uganda. ⁵Alokit, Hyderabad, India. ⁶Department of Economics, University of Texas at Austin, Austin, TX, USA. ⁷Innovations for Poverty Action (IPA), Manila, Philippines. ⁸World Bank, Washington, DC, USA. ⁹NewGlobe, Nairobi, Kenya. ¹⁰Deceased: Micheal Ainomugisha. ¹¹e-mail: noam.angrist@bsg.ox.ac.uk

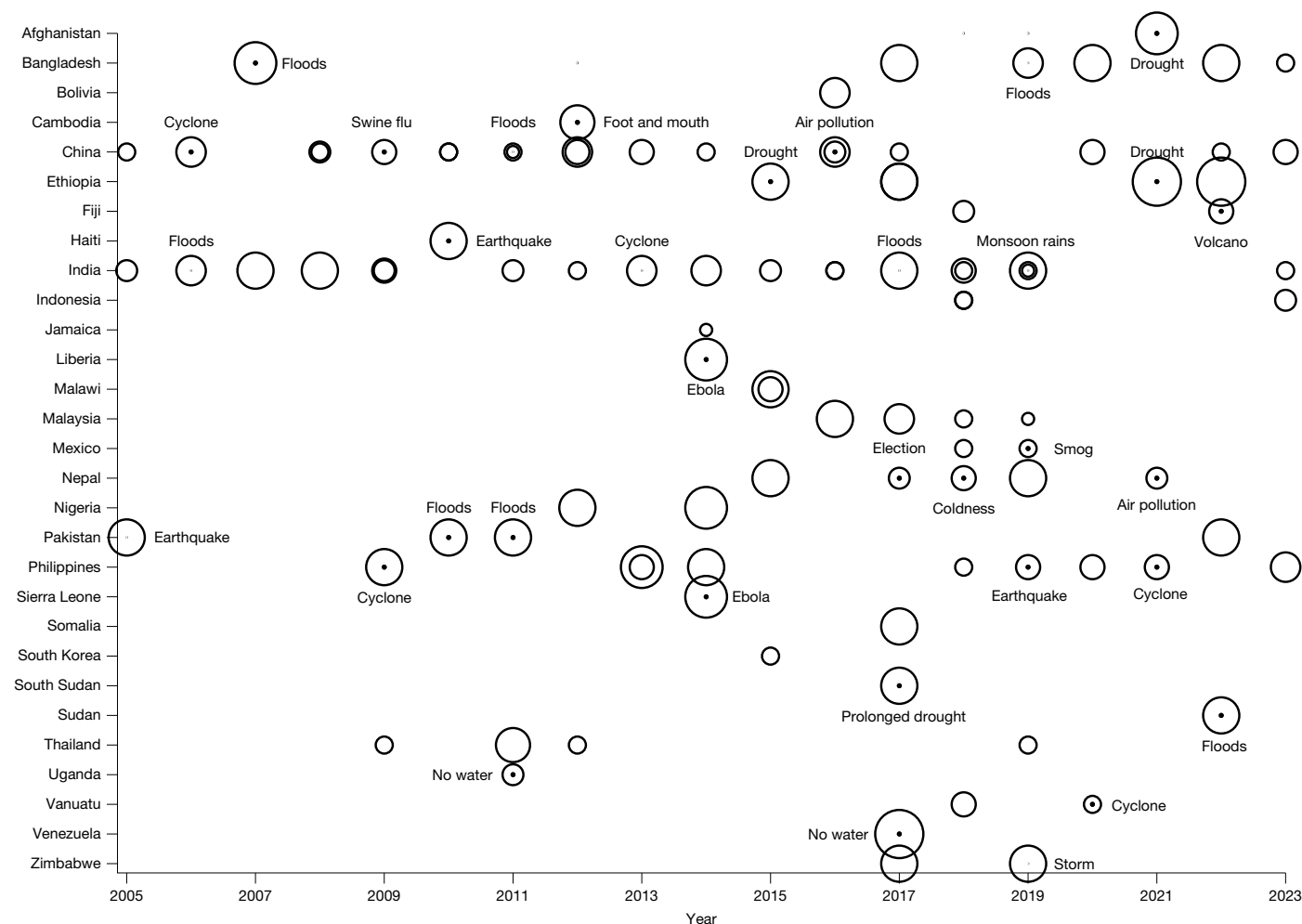


Fig. 1 | Documenting a set of education in emergencies between 2005 and 2023. The graph plots an index of the length of school closures and number of people affected by shocks that have disrupted schooling by country and year. The size of each bubble increases with the duration of school closure, the number of people affected, or both. Thicker or multiple circles represent multiple events in the same year and country. More information on the data

compilation is available in the Supplementary Information. The data comprise compiled school closure information based on press releases of the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) ReliefWeb, World Vision, United Nations Children's Fund (UNICEF), the British Broadcasting Corporation (BBC) and other local outlets. Major disruptions were also identified through The International Disaster Database (EM-DAT).

remain underutilized to provide mobile education. One treatment included a set of SMS messages, such as numeracy content provided weekly, as well as nudges to engage in educational activities. A second treatment provided additional weekly 20 min phone call tutorials for 8 weeks. The educational pedagogy was as essential as the mobile phone platform. Phone calls covered foundational numeracy and aimed to target instruction to student learning levels via regular low-cost, high-frequency assessments. This builds on both in-person and technology-based targeted instruction approaches^{14–19}.

A proof of concept in Botswana showed that phone call tutorials were effective in promoting learning during initial pandemic school disruptions²⁰. However, questions remain on whether this approach can be scaled across contexts and when delivered by governments—a pervasive challenge for social programmes^{21,22}. To ensure scientifically robust findings, we need evidence across diverse settings and delivery models²³. By conducting large randomized trials across five contexts as well as comparing models, such as NGO and government delivery, we contribute new evidence on effectiveness across diverse contexts and delivery models, key elements of scaling science²⁴.

This study contributes to the literature in several areas. First, we contribute to the nascent experimental literature on education in emergencies. Substantial research has been carried out in this area,

but much of it has been qualitative or with small samples. One exception is a randomized trial in which an NGO provided schooling in rural areas of conflict-affected regions in Afghanistan and found large effects on learning and closing of gender gaps²⁵. We expand this literature by evaluating alternative, scalable forms of education beyond traditional in-person schools, such as remote learning, which is the only option during many emergencies when schools close. There is a robust literature documenting the cost of school disruptions due to teacher strikes²⁶, earthquakes², school holidays⁷ and COVID-19 (refs. 4, 27–31). However, there is less evidence on effective approaches to stem these learning losses. We contribute evidence on scalable solutions to stem learning losses. A growing literature explores the effectiveness of remote tutoring, yet the results remain mixed^{32–35}. Without a coordinated cross-country trial it is impossible to know whether differences in effects are due to the intervention or to the setting, since both vary simultaneously. For example, a recent study³³ tested a related but different intervention in Sierra Leone—using a similar mobile phone platform but deploying a different pedagogy—and found no effect of mobile phone tutoring. In their study, the tutoring call was linked to a uniform national radio programme rather than targeting instruction based on student learning week-on-week—a key difference. To assess whether the approach works across contexts we cannot simply aggregate prior

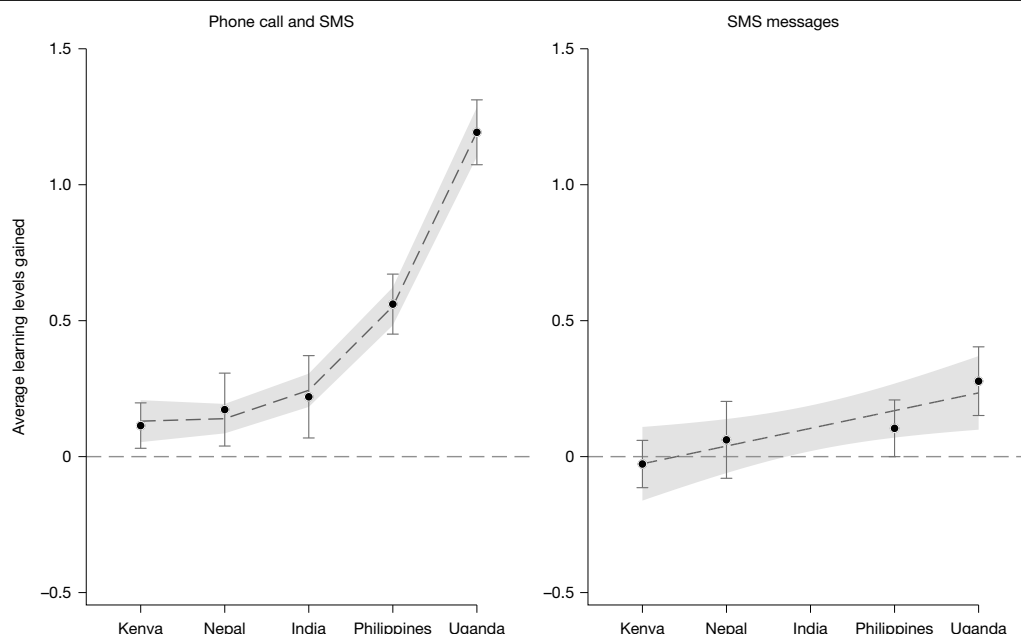


Fig. 2 | Main learning results across countries and treatments. Average raw unit operational learning levels gained in treatment groups relative to control groups. Treatment effects on learning outcomes are reported by country and treatment group with 90% confidence intervals within each country. Countries are ordered by the timing of the study, with Kenya first and Uganda last. The curve represents a fractional polynomial fit (phone arm) and a linear fit (SMS arm) of average learning levels gained against the order of the country in the

study, each with a 90% confidence band. ‘Learning levels gained’ refers to the number of operations gained, with each operation scored as an additional level gain (beginner, 0; addition, 1; subtraction, 2; multiplication, 3; division, 4). Data are presented as treatment effects relative to the control group with $\pm 90\%$ confidence intervals. Sample sizes: India ($n = 708$), Kenya ($n = 1,985$), Nepal ($n = 2,678$), the Philippines ($n = 2,469$) and Uganda ($n = 1,308$). Further details on related results are presented in Extended Data Table 1.

studies in which both intervention and context vary—we need to hold the intervention constant, while varying sites in a coordinated multi-site trial. We conduct such a coordinated five-country trial in this study.

We also contribute to a growing literature on external validity and scale. Recent examples show the extent of the scaling challenge, with many social programmes that initially worked in proof-of-concept studies no longer delivering impacts when scaled across settings or when delivered by governments^{22,24,36,37}. In contrast to some existing studies, our results identify an education approach that appears to work well across contexts and when delivered by governments. One reason might be that we leverage a particularly scalable technology: mobile phones^{11,38}. A second reason may be the generalizability of the underlying mechanisms tested in our study, which relate to other best practices in education. For example, tutoring has been shown to be one of the most effective, although expensive, approaches to improving learning in high-income settings^{39,40}. The phone call tutorials provide a cheaper version of tutoring that is applicable in low- and middle-income contexts. In addition, since the phone calls are one-on-one and have frequent learning assessments, they enable targeted instruction to every child’s level, another approach shown to consistently improve learning^{15–17,19}. Our study further advances the external validity and scale literature by conducting randomized trials across five countries in a literature where very few randomized controlled trials are multi-country studies. We also advance the scale literature by introducing experimental evaluation both of government relative to NGO delivery as well as randomizing those teachers who implement to evaluate effects on ongoing teacher practices. These experimental variations enable us to assess questions of effectiveness within government systems.

Finally, we contribute to the global education literature, with a focus on improving learning outcomes. Over the past several decades, education enrolments have improved worldwide, yet learning has barely budged^{10,41,42}. Growing evidence reveals that popular input-only reforms, such as general teacher training, provision of computers, or school grants, are not enough to improve learning. By contrast,

approaches that improve the quality of teaching, such as teaching at the right level and structured pedagogy, can generate large improvements in learning^{43–46}. The learning gains from the phone-based tutorials tested in this study can deliver up to 4 years of high-quality schooling per US\$100 (ref. 47). These types of highly cost-effective approaches can more easily scale and help address a persistent global learning crisis in low- and middle-income contexts.

Main results

We report results on a set of rich, harmonized data across five countries using both baseline and endline surveys. Questions included a learning assessment with the child and several additional child indicators, such as well-being, self-perceptions and non-cognitive skills. We also collected parental beliefs and behaviours. A portion of the survey was conducted with the parent, and learning outcomes were collected by directly assessing the child over the phone. A set of common questions were included in the baseline and endline surveys and across countries for comparability.

We report results from large-scale randomized controlled trials across five countries: India, Kenya, Nepal, the Philippines and Uganda. In Nepal and the Philippines, we include government and NGO delivery treatment arms. All countries except India also included a comparison SMS-only treatment relative to a phone call plus SMS treatment. Randomization enables identification of causal effects of the education in emergency interventions studied. Further details on the data and experimental methods are provided in the Methods section. Further details on the contexts and interventions are provided in the Supplementary Information.

Our main results are shown in Fig. 2 and show positive effects for the combined phone and SMS treatment in all countries, with increasing effectiveness over time as the trials progressed; for the SMS-only treatment, effects are more mixed across countries. Effects are further elaborated on in Extended Data Table 1, where we report effects

across contexts on the number of operations gained and the highest proficiency attained. We also show pooled average effects across all five countries. Column 1 in Extended Data Table 2 shows that for the combined phone and SMS group, there was a 0.321 s.d. ($P < 0.001$) increase in the average numerical operation across all 5 countries. SMS effects are smaller on average, with a 0.078 s.d. effect ($P = 0.006$). As shown in Fig. 2 and Extended Data Table 1, SMS messages only have a detectable statistically significant effect in a subset of contexts (Uganda and the Philippines). These results suggest that SMS messages can work in contexts with the largest need. However, live phone calls may be necessary to strike the balance of being intensive enough to deliver impact that can be sustained and scaled across diverse contexts while remaining cheap and scalable.

In addition, although effects of phone call tutorials are consistently positive, they vary in magnitude. Results are largest in later trials and in countries that experienced the longest school closures: Uganda and the Philippines. As shown in Extended Data Table 1, students in Uganda gain more than a full operation level (1.193 levels, $P < 0.001$). This translates to substantial proficiency gains, with 34.4 percentage point gains ($P < 0.001$) on the highest proficiency level. In the Philippines phone tutorials yield 0.56 level gains ($P < 0.001$) and a 13 percentage point increase ($P < 0.001$) in students reaching the highest operation level. In both countries, students were out of school for almost 2 years and the counterfactual for the control group was very limited schooling. These countries were also the later trials, with implementation fidelity improving over time. We discuss both of these mechanisms in more depth later on. Results are robust to the inclusion of a suite of control variables.

Mobile phone calls saw high take-up and high engagement throughout the programme, consistent with high impact. Detailed monitoring data shows high week-on-week engagement across sites ranging from 70% to 80%, as shown in Extended Data Fig. 1, even amid disruptions and across diverse low-resource settings. This highlights that the programme reached households using a platform that they found easy and convenient to access on a regular basis. This contrasts with the low rates of take-up of other platforms, with few households typically taking up alternative delivery modes such as online, television or radio educational resources.

For both the SMS group and the combined phone and SMS group, the magnitude of the impact and statistical significance are robust to different estimation approaches. In Extended Data Table 2, we include baseline controls and country and grade fixed effects (column 2), and weight results by country arm (columns 3 and 4). We also adjust for multiple hypothesis testing in Supplementary Table A5, including sharpened q values to address the false discovery rate⁴⁸. Since we have one outcome and two treatment arms, we consider two joint null hypotheses simultaneously. Estimates remain consistent.

Extended Data Table 3 and Supplementary Table A7 show learning outcomes across several proficiencies. Results show that the combined phone and SMS arm increases the share of students who get division problems correct by 13.5 percentage points ($P < 0.001$)—a 92% increase in division (from a control mean of 14.6%). These results by proficiency demonstrate that learning improved substantially in absolute terms, in addition to s.d. gains, and across a range of proficiencies. We also find increases in the share of students who are able to perform higher-order competencies, such as fractions (6.4 percentage point gains, $P < 0.001$). Since these competencies were not directly taught during the intervention, this reveals learning extended beyond familiarity with the content taught. These results further reveal potential dynamic complementarities in skill formation, showing that benefits of learning basic numeracy accrue to learning additional higher-order skills⁴⁹.

These results are large relative to the effectiveness of a typical education intervention. A recent review found that the median effective intervention yielded 0.10 s.d. gains in learning⁵⁰. Moreover, a review of 150 interventions found that more than half of education interventions do not work at all⁴⁷. These reviews put our results in context, with phone

Table 1 | Learning outcomes for government delivery

	All government arms	Nepal	Philippines
	(1) Operations gained	(2) Operations gained	(3) Operations gained
SMS messages	0.048 (0.049) [0.326]	0.072 (0.080) [0.367]	0.060 (0.061) [0.324]
Phone call and SMS: NGO	0.296*** (0.057) [0.000]	0.132 (0.088) [0.134]	0.510*** (0.073) [0.000]
Phone call and SMS: government	0.399*** (0.061) [0.000]	0.213** (0.087) [0.015]	0.629*** (0.086) [0.000]
Observations	5,147	2,678	2,469
Control mean	1.527	1.689	1.388
P value: NGO versus government	0.110	0.356	0.209
Country fixed effects	Yes	No	No
Countries included	Nepal, Philippines	Nepal	Philippines

Learning outcomes comparing implementation delivery models (NGO and government). Learning refers to how a child scores on four basic numeracy options: no operations correct, addition, subtraction, multiplication and division (measured on a scale of 0–4). This variable takes the value of 0 for students who got zero operations correct, 1 if students got addition correct, 2 if subtraction was correct, 3 if multiplication was correct, and 4 if students got division correct. Column (1) pools both countries in which we tested implementation scale-up in government systems. Column (2) shows the results for Nepal alone, and column (3) shows results for the Philippines alone. Standard errors are in parentheses and P values are in square brackets. * $P < 0.10$; ** $P < 0.05$; *** $P < 0.01$.

call tutorials being three times as effective as the median education intervention. Moreover, effects are even larger in these multi-country studies than in the Botswana proof-of-concept study, where learning improved by 0.12 s.d. (ref. 20). This result contrasts with prior literature showing that proof-of-concept studies rarely scale successfully to new contexts²¹ or experience diminishing returns³⁷. Rather than finding diminishing returns, we find that results improve as the approach is adapted, scaled and tested across contexts. We explore potential explanations later in the paper, including learning from experience as well as higher need for the intervention over time.

In Table 1, we show results for the subset of countries (Nepal and the Philippines) where we randomized delivery models to test scalability within government systems. Column (1) shows the pooled results: both government teachers and NGO instructors are effective at improving student learning, with government teachers improving learning by 0.40 operation levels across both contexts, and NGO teacher aides improving learning by 0.30 operation levels ($P < 0.001$ for both). These results demonstrate that education in emergencies programmes can be effectively implemented by government teachers, and not only by NGO implementers.

These results show substantial effectiveness when delivered by governments. Prior literature has found that programmes often work when delivered by NGOs but are not replicated when scaled by governments. For example, in Kenya, contract teachers improved learning, but when delivered by the government, the programme effectiveness waned³⁶. A large-scale government teacher training programme in Nepal similarly found no effect, largely owing to poor implementation⁵¹. Our results build on this literature, providing an alternative view: government delivery can achieve large learning gains. This is consistent with other studies that have refined interventions over time to facilitate government adoption^{15,17}.

Heterogeneous treatment effects, shown in Table 2, reveal that the programme is broadly effective for students across the distribution of

Table 2 | Heterogeneity

	Caregiver education	Gender	Baseline level
	(1) Operations gained	(2) Operations gained	(3) Operations gained
SMS messages	−0.013 (0.148) [0.933]	0.136*** (0.052) [0.008]	0.277*** (0.077) [0.000]
Phone call and SMS	0.288** (0.139) [0.038]	0.427*** (0.047) [0.000]	0.219** (0.095) [0.021]
Primary	0.026 (0.110) [0.812]		
Secondary plus	0.446*** (0.103) [0.000]		
SMS messages×primary	0.079 (0.169) [0.643]		
SMS messages×secondary plus	0.114 (0.154) [0.462]		
Phone call and SMS×primary	0.275* (0.160) [0.085]		
Phone call and SMS×secondary plus	0.112 (0.145) [0.439]		
Is female		0.093* (0.048) [0.052]	
SMS messages×is female		−0.070 (0.070) [0.321]	
Phone call and SMS×is female		−0.024 (0.064) [0.703]	
Baseline level			0.360*** (0.050) [0.000]
SMS messages×baseline level			0.007 (0.038) [0.848]
Phone call and SMS×baseline level			0.006 (0.035) [0.857]
Observations	8,028	9,148	9,148
Control mean	1.726	1.726	1.726
<i>P</i> value: SMS versus phone call plus SMS	0.048	0.000	0.627
Countries included	All 5	All 5	All 5
Country fixed effects	Yes	Yes	Yes

Heterogeneous treatment effects on learning. Column (1) shows the results between having caregivers who reached primary education or less and caregivers who reached secondary education or more (secondary plus). Column (2) shows the results between boys and girls. Column (3) shows results across baseline learning levels including country-specific baseline level interactions. To enable estimation for the full sample of endline learning results, baseline learning is imputed for the random sample, and not assessed based on the average. Standard errors are in parentheses and *P* values are in square brackets. **P* < 0.10; ***P* < 0.05; ****P* < 0.01.

baseline learning levels. This relatively even impact across students' starting conditions, or characteristics such as baseline learning levels or gender, is consistent with the mechanism of teaching at the right level. Targeted teaching is designed so that instruction meets children where they are regardless of their grade or age or other characteristics. Thus, broad-based learning gains are likely to be due to the fact that the programme was highly and effectively targeted, benefiting children across the distribution. We also find slight heterogeneity by parental education. Column (1) in Table 2 suggests that the programme worked particularly well for students where the caregiver had lower levels of formal education (primary education or less). This suggests that results are strongest when there are fewer alternative education support systems at home. This also reveals that even in low-literacy contexts, parents can be effective conduits for quality instruction.

We embedded an additional randomized study allocating a random subset of teachers to deliver the phone call tutorials in Nepal out of a list of hundreds of eligible government teachers. This additional experimental variation enables detection of the impact of delivering the programme on teachers' beliefs and practices—a measure of potential spillover effects into the education system that can persist through teachers beyond the intervention. In addition, teacher randomization enables generalization to the broader set of eligible government teachers. Thus, results capture government delivery impacts which do not hinge on implementers being highly selected, and effects are likely to translate to typical government teachers in the education system.

Table 3 shows that teacher practices shift; teachers are 9.3 percentage points more likely to target their feedback to students' learning level (*P* < 0.05). Teachers are also more likely to get parents involved in education (*P* < 0.1). We further find large effects on teacher perceptions that they were able to help students learn, as well as their desire to teach, with a 15.8 percentage point gain in wanting to be a teacher if they could make the choice again (*P* < 0.1 and *P* < 0.01, respectively). Although these results are reported by teachers and should be interpreted with caution, they suggest that delivering effective programmes could unlock a virtuous cycle within government education systems, in turn motivating teachers to want to teach and to improve their teaching practice further.

We examine counterfactual learning trajectories in Extended Data Figs. 2 and 3 to contextualize gains in the treatment groups. In Uganda, where effects are largest, about 17% of grade 4 students can divide at baseline in the control group. At endline, only 10% can divide, showing substantial learning loss. In the treatment group, 48% of grade 4 students can divide at endline, fully recovering learning loss due to school closures. Moreover, only 21% of grade 5 students in the control group at baseline can divide, revealing that grade 4 students in the treatment group surpass grade 5 students in the status quo. Thus, not only does the intervention facilitate full recovery of learning loss, it exceeds typical learning trajectories by nearly an order of magnitude. In other countries, learning does not deteriorate, but simply increases slowly. The intervention accelerates learning in these contexts relative to the counterfactual learning trajectory.

In the Supplementary Information, we include several additional results. We analyse effectiveness during another education emergency—a typhoon that struck the Philippines during the study. We explore additional margins of impact, including parent beliefs, student beliefs and non-cognitive skills. Finally, we explore the potential for crowd-out, to test whether tutoring time displaces alternative productive activities. Overall, we find continued effectiveness during an additional education emergency, several margins of impact beyond learning outcomes, and limited crowd-out.

Mechanisms across contexts

We explore variation in effects of phone call tutorials, which worked consistently but heterogeneously, across diverse contexts. Analysing mechanisms to explain effectiveness across contexts is challenging, and

Table 3 | Changing teacher beliefs and practices

	Teacher practices		Teacher beliefs	
	(1)	(2)	(3)	(4)
	Get parents involved	Targets to student level	Would be teacher again	Helped students' learning
Teacher implemented call	0.088*	0.093**	0.158***	0.106*
	(0.050)	(0.046)	(0.045)	(0.058)
	[0.076]	[0.043]	[0.000]	[0.067]
Observations	290	290	290	290
Control mean	0.721	0.769	0.735	0.408

Treatment effects on teacher practices and beliefs. This includes government teachers in Nepal who were randomly assigned to implement the phone call programme. We successfully spoke to 83% of the government-employed mathematics teachers who were in the schools identified by local governments. Of the eligible grade 3–5 mathematics teachers that we spoke to, 81% expressed interest in participating in the study. Of this eligible pool of 301 government teachers, 50% were randomly selected to implement the programme. The endline response rate was 96% and was balanced across the treatment and control groups. Column (1) shows effects on teachers' beliefs that they can get parents involved in their child's education. Column (2) shows effects on teachers saying that they tailor student feedback based on students' individual understanding and skill level. Column (3) shows effects on teachers saying if they had to choose their profession again, they would still be a teacher. Column (4) shows effects on teachers saying they believe they helped their students with their mathematical skills. These are all dummy variables that are coded as 1 if respondents strongly agreed on a Likert scale. Standard errors are in parentheses and *P* values are in square brackets. **P* < 0.10; ** *P* < 0.05; ****P* < 0.01.

should be done with caution, since multiple factors vary simultaneously. Table 4 includes a summary of key facts per country trial. We conduct exploratory analysis using suggestive descriptive data and observable systematic patterns. We focus on two mechanisms with particular variation across contexts: rates of targeted instruction and counterfactual conditions. Various additional margins vary across contexts, but do not vary systematically or consistently enough to facilitate credible comparisons. For example, in Kenya, younger students were taught but this varies only once, rather than systematically across settings.

Platform and pedagogy

In addition to leveraging a scalable platform—mobile phones—the programme pedagogy of targeted educational instruction, often referred to as ‘teaching at the right level’, was a core dimension of implementation fidelity. Phone call tutorials were designed to target instruction to children’s learning levels. We find that variation in effect across trials is consistent with variation in targeted instruction. Although results show that phone calls were consistently effective across countries, the magnitude of effects increases in tandem with the order of the trial: Kenya, followed by Nepal, India, the Philippines and finally Uganda. This order of effectiveness tracks the degree of targeted instruction. Detailed monitoring data provides evidence of increasing implementation fidelity as the trials progress. For example, the accuracy of targeted educational instruction increases from a starting point of 50.9% of students in Nepal to 81.5% on average in Uganda⁵² (Extended Data Fig. 4). These data reinforce the importance of targeted instruction. As the targeting mechanism improved study after study, this coincided with the programme becoming more effective. These results suggests the ‘voltage drop’ in effectiveness over time²¹—all too common when scaling—is not inevitable. Equipped with monitoring data and coordinating mechanisms to learn across trials, these results show that programmes have the potential to experience a ‘learning curve’ with improved fidelity (such as having ever more targeted instruction) and thus higher impact over time as they are scaled and implemented in new contexts. In addition to the trial-by-trial improvement in targeting accuracy, we find high rates of targeting accuracy through both NGO and government delivery. This consistent targeting accuracy across implementer types may explain the high effectiveness of both NGO and

government teachers. It also adds evidence regarding the importance of targeting accuracy as a mechanism for impact³². These data suggest that effective targeting pedagogy has an important role in programme effectiveness. This mechanism may explain why the approach tested worked across diverse contexts. Consistent with the importance of these mechanisms, a phone education programme that did not rely on these principles had less impact: phone calls in Sierra Leone that delivered non-targeted lessons based on a mass radio programme to students did not improve learning³³. In the Sierra Leone study, the approach used an accessible platform without targeted pedagogy. To successfully scale across different contexts, ensuring that targeted instruction is in place is likely to be key to improving learning.

Counterfactual conditions

Extended Data Table 4 shows several statistics characterizing counterfactual conditions. We observe clear patterns that show more severe counterfactual conditions coinciding with larger impacts. In Uganda and the Philippines, the settings with the highest impacts, we also see the most severe conditions, with more disruption and fewer alternative educational instruction options. Uganda and the Philippines have the highest degree of disruption in general and during the study, experiencing up to 2 years of disruption, with disruption occurring during the entire study period. They also have the largest share of lagging students, with 9% or less of students reaching the top operation level in the status quo. Similarly, we see that Uganda and the Philippines have low rates of outside options, with less than 22% of households saying they received any calls from educators in the past few weeks. Although these counterfactual conditions do not follow an exact rank order with effect size, the broad patterns are consistent—more severe conditions map broadly to larger effect sizes in Uganda and the Philippines. This suggests the interventions studied are likely to add value particularly in the most disrupted education in emergency settings.

Cost-effectiveness

An important feature of the education in emergencies approaches tested in this study is that they have low cost. The primary tool required to implement the programme is a mobile phone, a device that most households have access to, even in low-resource settings⁵³. Since the approach builds on existing household infrastructure, the main costs are related to content delivery and connecting with families, which are often marginal, such as airtime for phone calls. In addition to being low cost, the approach has low procurement needs, an attractive feature for governments. We carefully collected cost data in each trial. Our estimates suggest an average cost per child of the phone call and SMS tutorials of about US\$11 per child. This estimate is even cheaper than estimates in the proof-of-concept study in Botswana owing to economies of scale as the approach has scaled up, including cost savings such as use of existing pedagogical material, shorter and more efficient training, and streamlined data monitoring systems. We benchmark the impacts of the programme against other education programmes using a variety of approaches. First, we compare raw estimates of effectiveness and cost with similar programmes such as tutoring and targeted instruction. Phone call tutorials simulate the benefits of one-on-one tutoring, which has been shown to be one of the most effective educational approaches³⁹. However, many tutoring programmes have high cost. For example, a prominent tutoring programme yielded 0.19 to 0.31 s.d. learning gains at a cost of US\$2,500 per child. In low- and middle-income settings, phone calls could provide similar or larger impacts two orders of magnitude more cheaply, enabling scale-up across diverse settings. The SMS-only treatment, when effective, can be extremely cost-effective, at 41.1 learning-adjusted years of schooling (LAYS) per US\$100. However, whereas phone and SMS effects were consistent across all countries, the effects in the SMS-only arm varied across contexts. Given this, it seems

Table 4 | Trial description

	India	Kenya	Nepal	Philippines	Uganda
Sample size	850	6,724	3,732	3,492	2,138
Randomly selected endline sample size	765	3,556	3,351	3,164	1,495
Student grades	3–5	1–2	3–5	3–4	3–5
NGO delivery	✓	✓	✓	✓	✓
Government delivery	—	—	✓	✓	—
Unit of randomization	Household	School-grade cluster	Household	NGO: household; government: school-grade cluster	Household
Stratification variables used in block randomization	Student baseline level, school, female	School	Local government, parent perception of student baseline level	NGO: region, student baseline level; government: school, grade, student baseline level	Student baseline level, attends government school, previous education programme participation
Included combined phone call and SMS treatment	✓	✓	✓	✓	✓
Included SMS-only treatment	X	✓	✓	✓	✓
Implementer type	NGO (Alokit)	NGO public–private partnership (NewGlobe)	Government (Ministry of Education), World Bank, NGOs (Teach for Nepal, Street Child)	Government (Department of Education), research NGO (Innovations for Poverty)	NGO (Building Tomorrow)
Same teachers as school instruction	—	✓	—	✓ – government delivery	—
Number of weeks of implementation	8	8–12	8	8	8
Dates	Apr 2021–Jul 2021	Dec 2020–Apr 2021	Jan 2021–Jul 2021	Aug 2021–Jul 2022	Oct 2021–Jan 2022
Administrative units in study	1 state	30 counties	All 7 provinces	3 regions	9 districts

Summary of key features of the intervention across countries. Implementers of the interventions included government education ministries and NGOs.

plausible that the SMS-only intervention represents a cost-effective option in extreme education emergencies where, for example, calls are not an option or schooling is disrupted for such an extended period of time such that any provision of content is substantially better than the status quo.

Second, we use LAYS, a cost-effectiveness measure in education that has been estimated for over 150 impact evaluations in low- and middle-income countries; each LAYS can be interpreted as a high-quality year of schooling gained. We find the programme yields 3.6 LAYS per US\$100, ranking in the top 10 out of 150 education interventions reviewed⁴⁷. This result highlights the potential for phone call tutorials to deliver value to education systems and students in a broad array of contexts, both during and potentially even outside education emergencies. Even before the COVID-19 pandemic, there was high demand for cost-effective education programmes to help address the global learning crisis. Our results suggest phone-based targeted tutoring programmes, such as the one tested in this study, have the potential to deliver cost-effective learning gains across contexts and with governments.

We also contribute new evidence on the robustness of a cost-effective approach to measure learning—high-frequency phone call assessments. We leverage data across five countries and a series of randomized validity tests. Extended Data Table 5 shows the results of five checks we conducted, described in more depth in the Methods section. Results suggest that measurement of learning by phone can be highly robust, in addition to being cost-effective.

Conclusion

Here we present results from large-scale randomized trials evaluating the provision of education in emergency programmes across five countries: India, Kenya, Nepal, the Philippines and Uganda. We test multiple scalable models, including government delivery, of remote instruction

for primary school children during COVID-19, which disrupted education for more than 1 billion schoolchildren worldwide, as well as during an additional emergency—a typhoon—which further disrupted schooling.

Despite heterogeneous contexts, the results show that the effectiveness of phone call tutorials can scale across contexts; we find consistently large and robust effect sizes on learning, with average effects of 0.30–0.35 s.d. In the subset of trials in which we randomized whether the intervention was provided by NGO instructors or government teachers, we find large effects for both delivery models, indicating effectiveness when delivered within government systems (Supplementary Fig. A1).

These results have relevance to global efforts to support education in emergencies including conflict, disease, natural disasters and climate shocks, which routinely shut schools down. During these shocks, alternative models are needed to deliver education. The results presented here identify approaches that can scale effectively during disruptions across contexts and cost-effectively improve learning for students. This study further demonstrates that rigorous testing of programmes in humanitarian settings is possible.

Even outside education emergencies, millions of children worldwide learn very little in school, either because they are taught curricula beyond their learning level, or because they are unable to access quality instruction since they live in remote areas. Given widespread mobile phone ownership rates globally, phone-based tutoring programmes such as the one studied here have the potential to maintain schooling continuity and accelerate learning even outside emergency settings or after the initial emergency has eased. The low cost, high access and ease of implementation of mobile phone tutoring could build more resilience into education systems, enabling systems to better withstand frequent shocks, and to more generally utilize cost-effective approaches to address a persistent global learning crisis. One high profile example includes Karnataka, India, which is scaling this approach—now called ConnectEd globally and referred to in Karnataka as Ganitha

Ganaka—during school and to prepare for school disruptions due to monsoon seasons. Another example includes follow-on trials that built on this study in Latin America⁵⁴.

This study has substantial policy implications. Multiple efforts to scale the intervention studied in this paper are underway worldwide. This paper provides critical evidence to inform effectiveness across settings and delivery models. Future research could examine effects at even larger scale. As Extended Data Fig. 5 shows, the trials covered substantial portions of the geographies studied. However, there is still value in exploring effects at national scale in future work. Future research could examine refinements to the delivery model to engage parents further⁵⁵ and could evaluate potential spillovers on teacher instruction and student learning in the classroom. Moreover, additional work could explore long-term benefits following observed gains during the study period, with learning potentially compounding over time as skills beget more skills.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-026-10990-x>.

- Valenza, M., Stoff, C. & Yao, H. *Global Estimates: Number of Crisis-Affected Children and Adolescents in Need of Education Support* <https://go.nature.com/4xVCdEH> (Education Cannot Wait, 2022).
- Andrabi, T., Daniels, B. & Das, J. Human capital accumulation and disasters: evidence from the Pakistan earthquake of 2005. *J. Hum. Resour.* **58**, 1057–1096 (2023).
- Lichand, G., Doria, C. A., Leal-Neto, O. & Fernandes, J. P. C. The impacts of remote learning in secondary education during the pandemic in Brazil. *Nat. Hum. Behav.* **6**, 1079–1086 (2022).
- Carlana, M., La Ferrara, E. & Lopez, C. Exacerbated inequalities: the learning loss from COVID-19 in Italy. *AEA Pap. Proc.* **113**, 489–493 (2023).
- Asim, S., Bashir, S. & Gera, R. C. *Learning Loss as a Result of COVID-19* (World Bank, 2024).
- Sabarwal, S., Venegas Marin, S., Spivack, M. H. & Ambasz, D. *Choosing Our Future—Education for Climate Action* (World Bank, 2024).
- Cooper, H., Nye, B., Charlton, K., Lindsay, J. & Greathouse, S. The effects of summer vacation on achievement test scores: a narrative and meta-analytic review. *Rev. Educ. Res.* **66**, 227–268 (1996).
- Slade, T. S., Piper, B., Kaunda, Z., King, S. & Ibrahim, H. Is ‘summer’ reading loss universal? Using ongoing literacy assessment in Malawi to estimate the loss from grade-transition breaks. *Res. Comp. Int. Educ.* **12**, 461–485 (2017).
- Ganimian, A. J. & Djaker, S. *How can Developing Countries Address Heterogeneity in Students’ Preparation for School? A Review of the Challenge and Potential Solutions* (New York Univ., 2022).
- World Bank. *World Development Report 2018: Learning to Realize Education’s Promise* (World Bank, 2018).
- Aker, J. C. & Mbiti, I. M. Mobile phones and economic development in Africa. *J. Econ. Perspect.* **24**, 207–232 (2010).
- Carvalho, S. & Crawford, L. School’s out: now what? *Center for Global Development* <https://www.cgdev.org/blog/schools-out-now-what> (2020).
- Callen, M., Fajardo-Steinhäuser, M., Findley, M. G. & Ghani, T. Can digital aid deliver during humanitarian crises? *Manag. Sci.* **71**, 8995–9868 (2025).
- Banerjee, A. V., Cole, S., Duflo, E. & Linden, L. Remedying education: evidence from two randomized experiments in India. *Q. J. Econ.* **122**, 1235–1264 (2007).
- Banerjee, A. et al. From proof of concept to scalable policies: challenges and solutions, with an application. *J. Econ. Perspect.* **31**, 73–102 (2017).
- Muralidharan, K., Singh, A. & Ganimian, A. J. Disrupting education? Experimental evidence on technology-aided instruction in India. *Am. Econ. Rev.* **109**, 1426–1460 (2019).
- Duflo, A., Kiessell, J. & Lucas, A. *External Validity: Four Models of Improving Student Achievement* Working Paper w27298 (National Bureau of Economic Research, 2020).
- Ganimian, A. J., Vegas, E. & Hess, F. M. *Realizing the Promise: How Can Education Technology Improve Learning for All?* (The Brookings Institution, 2020).
- Angrist, N. & Meager, R. Implementation matters: generalizing treatment effects in education. Preprint at SSRN <https://doi.org/10.2139/ssrn.4487496> (2023).
- Angrist, N., Bergman, P. & Matsheng, M. Experimental evidence on learning using low-tech when school is out. *Nat. Hum. Behav.* **6**, 941–950 (2022).
- List, J. A. *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale* (Crown Currency, 2022).
- Mobarak, A. M. Assessing social aid: the scale-up process needs evidence, too. *Nature* **609**, 892–894 (2022).
- Maniadis, A., Tufano, F. & List, J. A. To replicate or not to replicate? Exploring reproducibility in economics through the lens of a model and a pilot study. *Econ. J.* **127**, F209–F235 (2017).
- List, J. A. Optimally generate policy-based evidence before scaling. *Nature* **626**, 491–499 (2024).
- Burde, D. & Linden, L. L. Bringing education to Afghan girls: A randomized controlled trial of village-based schools. *Am. Econ. J.* **5**, 27–40 (2013).
- Jaume, D. & Willén, A. The long-run effects of teacher strikes: evidence from Argentina. *J. Labor Econ.* **37**, 1097–1139 (2019).
- Bacher-Hicks, A., Goodman, J. & Mulhern, C. Inequality in household adaptation to schooling shocks: Covid-induced online learning engagement in real time. *J. Public Econ.* **193**, 104345 (2021).
- Jack, R., Halloran, C., Okun, J. & Oster, E. Pandemic schooling mode and student test scores: evidence from US school districts. *Am. Econ. Rev. Insights* **5**, 173–190 (2021).
- Patrinos, H. A., Vegas, E. & Carter-Rau, R. *An Analysis of COVID-19 Student Learning Loss* (World Bank, 2022).
- Moscoviz, L. & Evans, D. K. *Learning Loss and Student Dropouts During the Covid-19 Pandemic: A Review of the Evidence Two Years After Schools Shut Down* Working Paper 609 (Center for Global Development, 2022).
- Hevia, F. J., Székely, M., Vinacur, T. & Zoido, P. *Tutorías Remotas: Revisión de la Literatura* (Inter-American Development Bank, 2022).
- Schueler, B. E. & Rodríguez-Segura, D. A *Cautionary Tale of Tutoring Hard-to-Reach Students in Kenya* Working Paper 21-432 (Annenberg Institute for School Reform, 2021).
- Crawford, L., Evans, D. K., Hares, S. & Sandefur, J. *Teaching and Testing by Phone in a Pandemic* Working Paper 591 (Center for Global Development, 2021).
- Carlana, M. & La Ferrara, E. Apart but connected: online tutoring, cognitive outcomes, and soft skills. *Am. Econ. Rev.* **115**, 3487–3513 (2025).
- Gortazar, L., Hupkau, C. & Roldán-Monés, A. Online tutoring works: experimental evidence from a program with vulnerable children. *J. Public Econ.* **232**, 105082 (2024).
- Bold, T., Kimenyi, M., Mwangi, G. & Sandefur, J. Experimental evidence on scaling up education reforms in Kenya. *J. Public Econ.* **168**, 1–20 (2018).
- Araujo, M. C., Rubio-Codina, M. & Schady, N. in *The Scale-Up Effect in Early Childhood and Public Policy* (eds List, J. et al.) 211–232 (Routledge, 2021).
- Aker, J. C., Ksoll, C. & Lybbert, T. J. Can mobile phones improve learning? Evidence from a field experiment in Niger. *Am. Econ. J.* **4**, 94–120 (2012).
- Nickow, A., Oreopoulos, P. & Quan, V. *The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence* Working Paper 27476 <https://doi.org/10.3386/w27476> (National Bureau of Economic Research, 2020).
- Robinson, C. D. & Loeb, S. *High-Impact Tutoring: State of the Research and Priorities for Future Learning* Report 21 (National Student Support Accelerator, 2021).
- Pritchett, L. *The Rebirth of Education: Schooling Ain’t Learning* (CGD Books, 2013).
- Angrist, M., Djankov, S., Goldberg, P. K. & Patrinos, H. A. Measuring human capital using global learning data. *Nature* **592**, 403–408 (2021).
- Kremer, M., Brannen, C. & Glennerster, R. The challenge of education and learning in the developing world. *Science* **340**, 297–300 (2013).
- Ganimian, A. J. & Murnane, R. J. Improving education in developing countries: Lessons from rigorous impact evaluations. *Rev. Educ. Res.* **86**, 719–755 (2016).
- Glewwe, P. & Muralidharan, K. in *Handbook of the Economics of Education* (eds Hanushek, E. A. et al.) 653–743 (Elsevier, 2016).
- Eble, A. et al. How much can we remedy very low learning levels in rural parts of low-income countries? Impact and generalizability of a multi-pronged para-teacher intervention from a cluster-randomized trial in the Gambia. *J. Dev. Econ.* **148**, 102539 (2021).
- Angrist, N. et al. How to improve education outcomes most efficiently? A review of the evidence using a unified metric. *J. Dev. Econ.* **172**, 103382 (2024).
- Anderson, M. L. Multiple inference and gender differences in the effects of early intervention: a reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *J. Am. Stat. Assoc.* **103**, 1481–1495 (2008).
- Cunha, F. & Heckman, J. The technology of skill formation. *Am. Econ. Rev.* **97**, 31–47 (2007).
- Evans, D. K. & Yuan, F. How big are effect sizes in international education studies? *Educ. Eval. Policy Anal.* **44**, 532–540 (2022).
- Schaffner, J., Glewwe, P. & Sharma, U. Why programs fail: lessons for improving public service quality from a mixed-methods evaluation of an unsuccessful teacher training program in Nepal. *World Bank Econ. Rev.* **39**, 473–496 (2025).
- Angrist, N. et al. Learning curve: progress in the replication crisis. *AEA Pap. Proc.* **113**, 482–488 (2023).
- World Bank. Mobile cellular subscriptions (per 100 people). *World Bank Group* <https://data.worldbank.org/indicator/IT.CEL.SETS.P2> (2021).
- Albornoz, F., Torres, G. A., Lombardi, M., Oubiña, V. & Lobaton, P. Z. Remote tutoring in Latin America. *J. Dev. Econ.* **179**, 103687 (2025).
- Angrist, N., Cullen, C. & Magat, J. Cheaper (and more effective) by the dozen: evidence from 12 randomised A/B tests optimising tutoring for scale. Preprint at SSRN <https://doi.org/10.2139/ssrn.5755162> (2025).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026

Methods

Data

In each country, we have two waves of data: baseline and endline. In Nepal, we conducted a baseline survey with a random 50% of the sample, and in Kenya, we relied on administrative data instead of baseline survey data. The endline surveys took approximately 30 min to administer and included approximately 20 questions. These questions included a learning assessment, child wellbeing, parental engagement in educational activities, and parental perceptions of their child's learning. A portion of the survey was conducted with the parent, and learning outcomes were collected by directly assessing the child over the phone. Endline surveys were conducted a few months after the programme ended. A set of common core questions in the baseline and endline surveys are included in the Supplementary Information.

The learning assessment was adapted from the ASER (Annual Status of Education Report) test, which has been used frequently in the literature to measure learning outcomes^{14,15} and is used routinely across 14 countries. The test consists of multiple numeracy items, including two-digit addition, subtraction, multiplication and division problems. In addition, we asked students to solve a word problem and a fraction problem to capture learning outcomes beyond a core set of mathematical operations. Results are presented in several main results figures and tables. Distributions of raw baseline and endline learning levels are shown in Supplementary Fig. A3.

To maximize the reliability of the phone-based assessment, we introduced a series of quality-assurance measures. To minimize family members in the household assisting the child, students had a time cap of 2 minutes per question and we asked each child to explain their work. We only marked a problem correct if the child correctly explained how they solved the problem and enumerators were confident parents were not assisting their child. We also conducted a battery of validity checks to ensure the reliability of the learning outcomes. These results contribute new evidence on the robustness of remote learning assessment data across five countries. Phone assessment has emerged as a common strategy for large-scale household surveys such as the World Bank Living Standards Measurement Survey (LSMS). A growing literature has started to explore the validity of phone-based assessments to measure learning outcomes^{20,56–58}, with emerging evidence suggesting phone assessment can capture meaningful information at high frequency and low cost.

Extended Data Table 5 shows the results of five checks we conducted on the validity of our main learning outcomes of learning assessments via phone. Column 1 shows the first robustness check where we compare in-person to phone-based assessment for a representative sample of the same students in Kenya. An additional test included back-checks, with a random subset of students tested twice on the same competencies. We find a strong relationship as expected, with large positive coefficients and *t*-statistics ranging from 5 to more than 20. We further randomize students to receive different problems of the same proficiency (for example, four different questions to measure two-digit addition with carryover). Finally, column 5 shows results from a real-effort question to disentangle effects of the intervention on effort on the test versus cognitive skills. Students were asked to answer several effort tasks, for example figuring out the day of the week or counting zeros and ones. Altogether, results suggest measuring learning by phone can be robust, similar to findings from the World Bank LSMS for household measures of consumption.

The survey also included questions on caregiver engagement in their child's education and beliefs. We measured engagement by asking caregivers how often they spent helping their child with their schoolwork over the previous weeks. We also included a measure of a caregiver's confidence in their child having made progress in learning over the previous months, and their perception of their child's numeracy level. Additional questions included information on whether the caregiver

has returned to work. We also asked about parents' demand for remote learning services in the future, and whether they would be willing to pay for such a programme. For students, we asked about child's mental wellbeing, how much they enjoy school, and the child's own belief about what mathematics problems they will be able to answer. We also measured non-cognitive skills, such as perseverance and ambition. Results are shown in Supplementary Tables A2 and A3. Finally, we included demographic questions, recording the child's age, grade and gender.

The overall sample size, pooling all sites, is 16,936 households. The flow of participants from enrolment through randomization to endline follow-up is summarized in a CONSORT diagram (Supplementary Fig. A5). For endline surveys, we randomly sampled households from the full sample to interview for a total sample of 12,331. This was due to time and cost constraints. In Supplementary Table A8, we show that those randomly selected for endline interview are statistically equivalent to the full sample at baseline along a series of indicators such as gender, student grade and baseline learning level.

Supplementary Table A9 presents the response rate to the endline surveys and an analysis of survey attrition for those randomly selected to be part of the endline sample. The follow-up rate was around 76% of respondents at endline. Supplementary Table A9 also presents a test of whether response rates differed by treatment assignment. This provides evidence that our sample has a high and unbiased response rate.

Finally, we also include a survey for teachers to assess their beliefs and instructional practices. These questions include their desire to be a teacher, and teachers' view that the phone call tutorial programme was helpful for student learning. Questions also include instructional practices, such as involving parents in education further and better targeting feedback to students' actual learning level. These questions can capture potentially persistent effects on educational systems through teachers changing their beliefs and behaviours beyond the lifecycle of the programme.

Experimental design

We ran randomized controlled trials across five countries. Every trial had a control group and a combined phone call and SMS treatment arm. All studies except the trial in India also included an SMS-only treatment arm. In two countries, Nepal and the Philippines, we further randomized delivery of the phone call by NGO instructors or government teachers to assess scalability within government systems. We exploit random assignment to identify causal effects and estimate the impact of these education in emergencies interventions and various scalable delivery models.

We first estimate intent-to-treat effects per country *c* as follows:

$$Y_{i,j,c} = \alpha + \beta_1 \text{PhoneCalls}_{j,c} + \beta_2 \text{SMS}_{j,c} + \gamma X_{j,c} + \delta_s + \varepsilon_{i,j,c} \quad (1)$$

where $Y_{i,j,c}$ is a learning outcome for individual *i* in household *j* in country *c*. $\text{PhoneCalls}_{j,c}$ and $\text{SMS}_{j,c}$ take on the value 1 in their respective treatment arms and 0 otherwise. $X_{j,c}$ denotes a vector of baseline control variables, applied consistently across trials to enhance statistical power and precision: student grade, student gender and baseline learning level. For baseline learning, we have the same assessment as the endline in all countries except Kenya, where we use administrative exam scores, and in Nepal, where we use parent perceptions of baseline learning. Since random subsamples are drawn at baseline in a few cases, we complete the sample with average estimates to enable analysis to consider the full sample at endline when including controls. δ_s denotes the geographic strata fixed effects—the randomization stratum relevant to each country (for example, region, district or constituency). We assess impact on operations gained as well as the highest proficiency achieved in a given country. We estimate the regression using ordinary least squares. Our main specifications are

conducted per country. Given the speed and timing of the studies conducted during a historic crisis, rather than register new analysis plans, we rely on specifications published in prior related studies to guide our analysis²⁰. Our specifications are straightforward, with one main learning outcome measure. In the extended data, we conduct corollary analysis where we pool across studies and we include country fixed effects η_c . We run additional regressions using alternative weights by country as well as country-by-country estimations. Our primary analysis focuses on learning outcomes expressed as operation levels gained in raw units as well as the proportion of students who make it to the top proficiency level. We analyse standardized learning outcomes only as a corollary analysis (standardized relative to control group standard deviations at endline and centred at mean zero). All tests are two-sided statistical tests. All analysis was conducted in Stata/MP 18 software.

We also ran a subset of trials in which phone calls were delivered by NGOs or government teachers. We estimate government (Gov) versus NGO delivery effects per country using the following specification:

$$Y_{i,j,c} = \alpha + \beta_1 \text{PhoneCalls}_{j,c,\text{NGO}} + \beta_2 \text{PhoneCalls}_{j,c,\text{Gov}} + \gamma X_{j,c} + \delta_s + \varepsilon_{i,j,c} \quad (2)$$

All standard errors are clustered at the level of the unit of randomization. In most countries this was the household level, which was de facto the same as the student level since one student participated per household; in a subset of treatment comparisons it was the school grade level. In India, Nepal and Uganda all randomization was at the household level. In the Philippines, the NGO arm was randomized at the household level; the government arm was randomized at the school grade level, with a corresponding control group randomized at the same level for each arm. In Kenya, the phone call arm was randomized at the school grade level whereas the SMS arms were cross randomized at the household level. Of note, robustness checks find little to no difference if standard errors are clustered or not across all specifications. This is largely due to the fact that there is substantial variation in learning outcomes within a cluster, such that the intracluster correlation is very low at about 0.03 in the Philippines and Kenya. The above specifications are the core estimation strategies. Additional specifications include other outcomes of interest beyond learning, such as engagement, as well as heterogeneity analysis, and various types of pooling across countries.

Due to randomization, we expect treatment and control groups to be balanced at baseline in expectation. Supplementary Table A10 presents balance tests using baseline data for key demographics and the same learning level variable. We find no statistically significant differences across multiple dimensions, including gender, grade, caregiver type, caregiver education levels and baseline learning. This is robust to multiple specifications, including with and without country fixed effects. Similar results are found for each country, with balanced samples in each individual country. This reveals that each randomization yielded balanced arms in line with expectations.

In all sites, the experimental design had high compliance rates; nearly all randomized treatment households participated in the expected treatment group. Detailed weekly monitoring data shows high take-up and fidelity, as shown in Extended Data Fig. 1.

In a secondary analysis, we embed another randomized trial which randomly allocated a subset of government teachers to deliver the phone call tutorials out of a pool of hundreds of eligible teachers in Nepal. Supplementary Table A6 shows balance tests with no statistically significant differences among treatment groups. We assess whether participating in phone call tutorials changed teachers' subsequent beliefs and behaviours, which we estimate for teachers i at the school level j in each country c as follows:

$$S_{i,j,c} = \alpha + \beta_1 \text{TeacherCalls}_{i,c} + \gamma X_{i,c} + \delta_s + \varepsilon_{i,j,c} \quad (3)$$

where $S_{i,j,c}$ captures teacher beliefs or instructional practices. These outcomes measure potential spillovers into the education system through teacher beliefs and behaviours which could persist beyond the programme.

Ethics statement

Institutional review board (IRB) approval was granted by the Health Media Lab IRB (889OXFD21) and Columbia University IRB (21-384). Informed consent was performed in accordance with all relevant guidelines and regulations and obtained from caregivers, child assent and government approvals in each country. An additional country-specific IRB approval was granted by Nepal's Health Research Council (776/2020) due to local regulations. The trials were registered in the AEA registry: AEARCTR-0009779. After the study period, control group students were later able to be enrolled in the intervention to benefit from the programme.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Anonymized replication data are available at <https://github.com/nangrist/mobile-education>.

Code availability

All replication code is available from <https://github.com/nangrist/mobile-education>.

56. World Bank. *Supporting Student Learning at Home with Phone-based Formative Assessments: Landscape Review* (World Bank, 2022).
57. Rodriguez-Segura, D. & Schueler, B. E. Can learning be measured by phone? Evidence from Kenya. *Econ. Educ. Rev.* **90**, 102309 (2022).
58. Gupta, S., Satyam, K., Gupta, N. & Ahluwalia, R. *Can Phone Assessments Generate Reliable Student Learning Data?* (Central Square Foundation, 2022).

Acknowledgements We are grateful to a coalition of partners who enabled this multi-country effort. Implementing and research partners include: Youth Impact, J-PAL, Learning Collider (supported by Schmidt Futures & Citadel), Oxford University, World Bank, Nepal's Ministry of Education, Science and Technology, Teach for Nepal, Street Child, Philippines Department of Education, IPA, Building Tomorrow, NewGlobe, Alokit and Global School Leaders. We thank N. Ahuja, A. Jung and R. Zhou, who provided excellent research assistance; implementation and research teams, including N. Baddiri, M. Bayangos, K. Basnet, J. L. Carandang, C. de Maricourt, P. Ghimire, F. Karanja, M. Karki, K. Cruz, U. Limbu, R. Masapol, E. Munyaneza, S. Poudel, K. Radhakrishnan-Nair, P. Regmi, U. Sharma, S. Shrestha and K. Subedi; conference participants from ASSA, BREAD, CSAE, NEUDC and PacDev for comments; M. Carlana, D. Karlan and M. Callen for detailed comments. For the India scale-up effort, we thank the Karnataka government, J-PAL South Asia, Alokit, the Agency Fund, the What Works Hub for Global Education, the UK Foreign Commonwealth and Development Office (FCDO) as well as S. Ochmann, R. Katepallewar, S. Prabakaran, K. Manjunath, N. Sadanand, R. Grover, S. Attavar and S. Djaker. We dedicate this paper in loving memory to Micheal Ainomugisha—his commitment to evidence and improving children's lives was unwavering.

Author contributions N.A. and C. Cullen conceived and designed the study. M.M., T.L. and C. Crossley provided technical intervention and partnership assistance and coordinated implementation across sites. Country leads led implementation and data collection in India (S.P.B.), Kenya (T.S.), Nepal (S.S.), the Philippines (R.M.P.) and Uganda (M.A.). N.A. and C. Cullen led the analysis. S.S. and P.B. contributed substantially to study design and interpretation. N.A. led the manuscript drafting and editing with input from all authors. All authors reviewed and approved the final manuscript.

Funding We thank funders who made this study possible at a time of historic need, including UBS Optimus Foundation, Mulago Foundation, Douglas B. Marshall Foundation, Echidna Giving, Stavros Niarchos Foundation, Jacobs Foundation, JPAL IGI, Northwestern's 'Economics of Nonprofits' class, the Peter Cundill Foundation, and the World Bank. C. Cullen received funding from the UKRI GCRF and Oxford's OPEN Fellowship.

Competing interests N.A. and M.M. co-founded Youth Impact, the non-governmental organization that first developed and tested the ConnectEd phone tutoring programme evaluated in this study. C. Cullen heads Youth Impact's research team. Youth Impact provided training and technical assistance to implementing partners across the five trials but did not itself deliver the intervention in the study countries and received no financial benefit contingent on the study's findings. All other authors declare no competing interests.

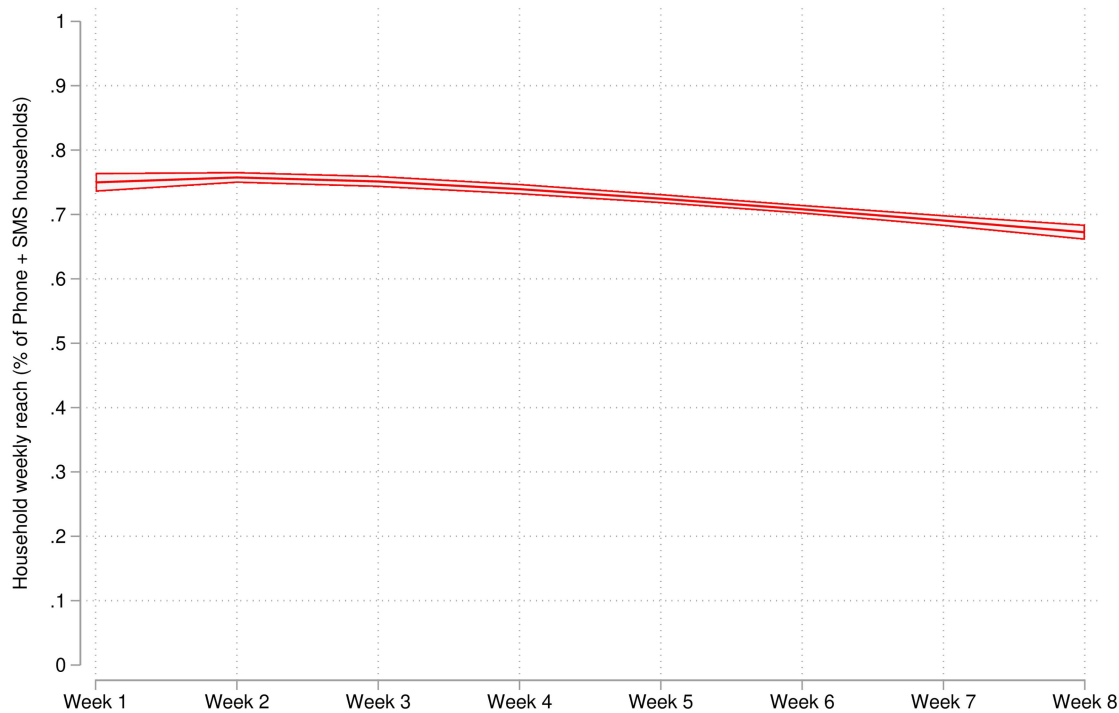
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-026-10990-x>.

Correspondence and requests for materials should be addressed to Noam Angrist.

Peer review information *Nature* thanks Salman Asim, Felipe Barrera-Osorio, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

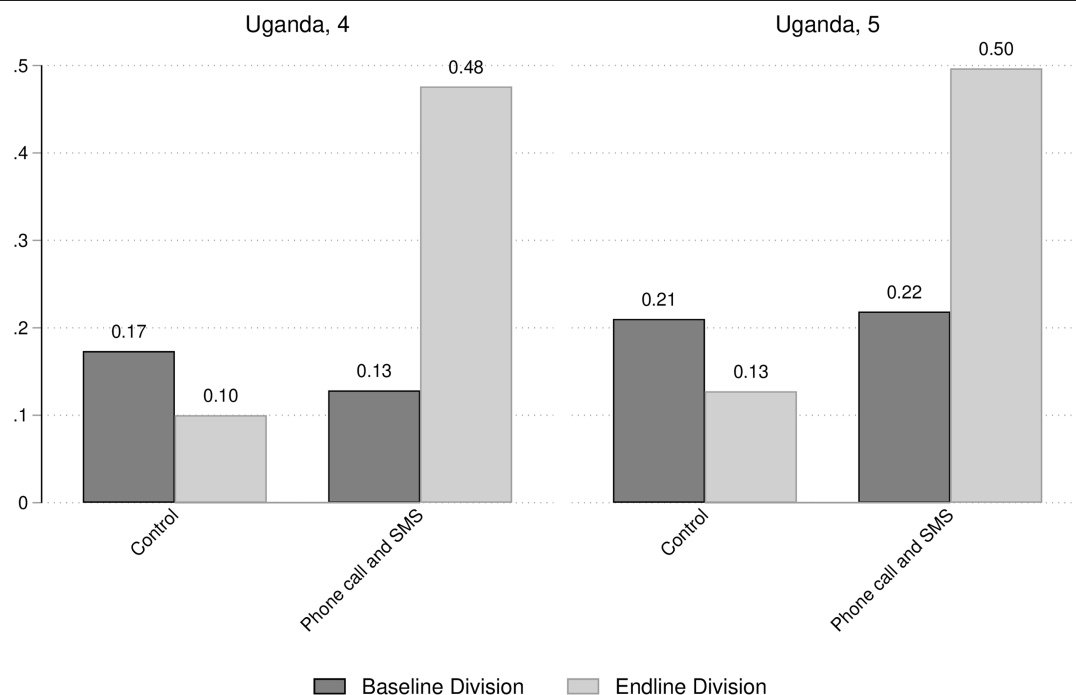
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Weekly Household Engagement in Phone Calls.

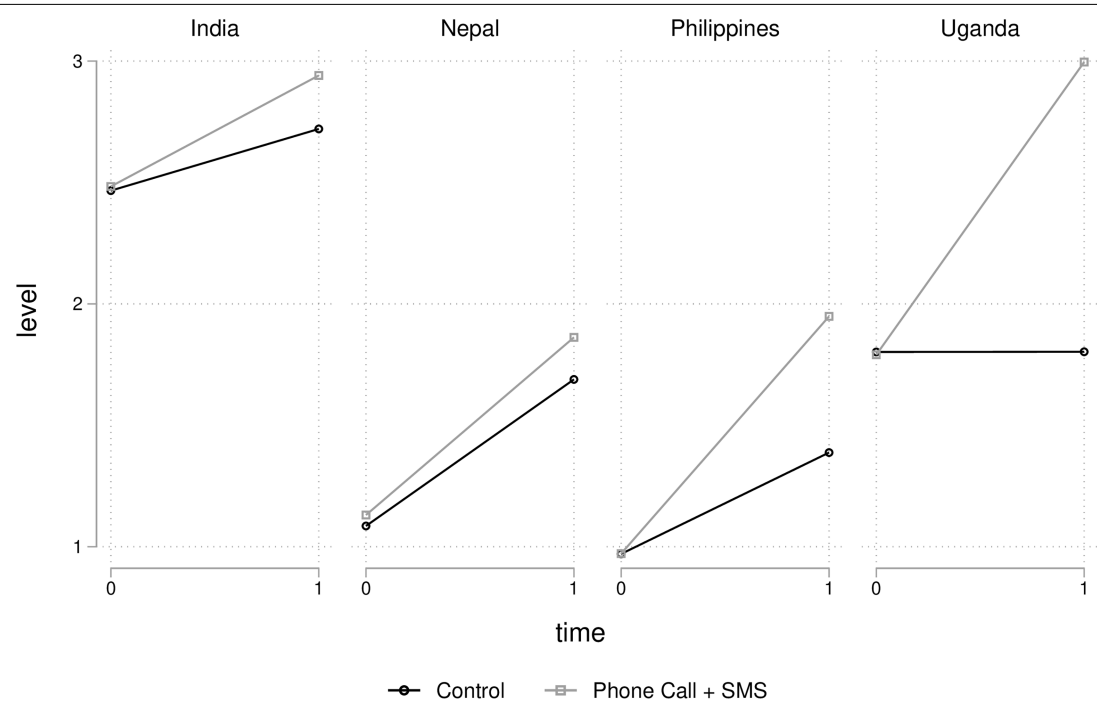
Notes: This figure shows average weekly engagement of households in the phone call and SMS arm. Each week, engagement is coded as 1 if the household

answered the call and engaged. This includes data from Nepal, Uganda and the Philippines where we have detailed and easily comparable weekly monitoring data. The shaded band represents ± 95 percent confidence intervals.



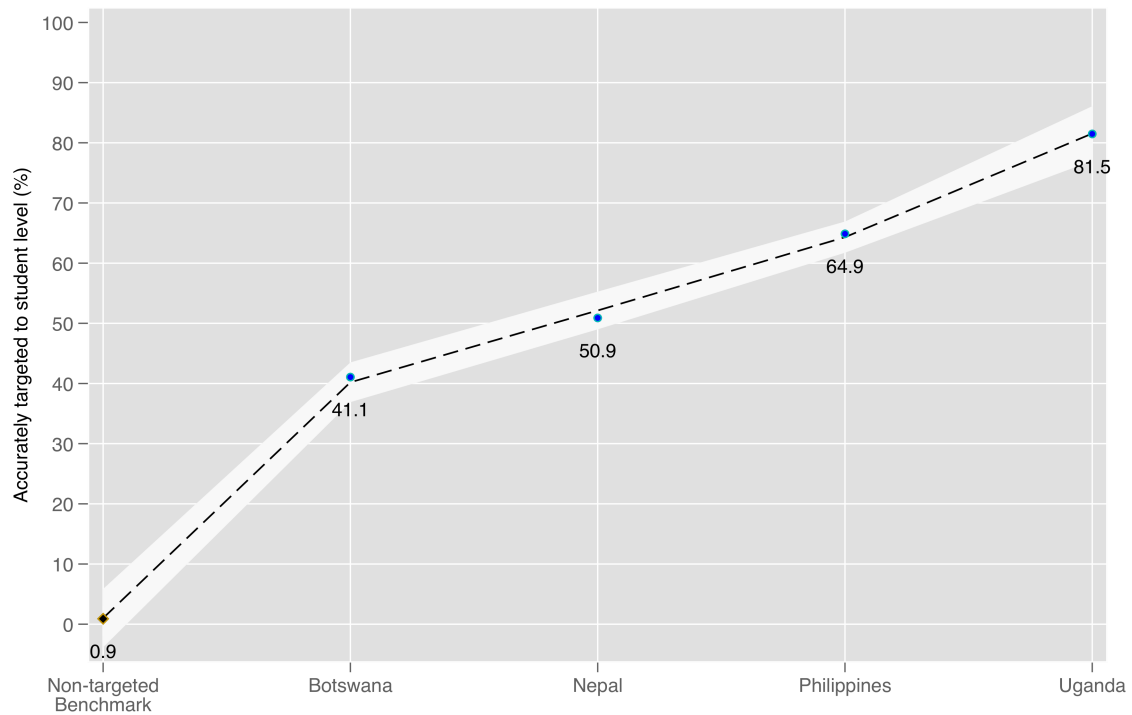
Extended Data Fig. 2 | Learning Losses & Gains - Can Divide, Uganda, by Grade. **Notes:** This figure shows the proportion of Grade 4 and Grade 5 students in Uganda that know how to correctly perform division. The shades distinguish the baseline and endline proportion. In each panel, the two bars on the left show the outcomes for students who did not receive the intervention; the two on the right show the outcomes for those who received the SMS exercises and phone call tutoring. The figure shows that there is significant

learning loss for control students between baseline and endline. It also shows that there is only a modest increase in the year-on-year share of students at division-level proficiency between grades 4 and 5 in the status quo. Finally, it shows the share of students able to perform division in the treatment group both recovers learning loss and far exceeds the year-on-year progress in division proficiency, with treated students overtaking the subsequent grade's proficiency level at baseline.



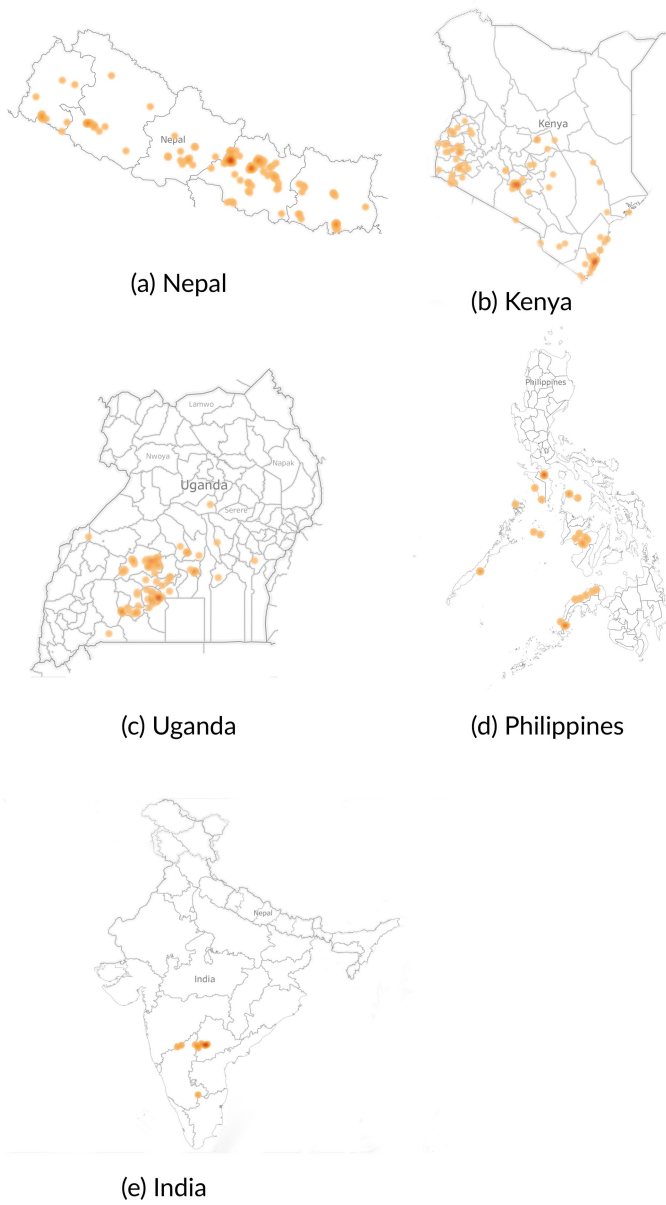
Extended Data Fig. 3 | Learning Trajectories. Notes: This figure shows baseline and endline operation levels for the phone call tutorial treatment arm by country to showcase learning gains relative to status quo learning trajectories in the control group. We include all countries except Kenya since Kenya did not use an identical learning test at baseline as at endline. All other

countries – India, Nepal, Philippines and Uganda – used the same learning test at baseline and endline. Endline sample sizes are as follows: India (n = 708), Nepal (n = 2,678), Philippines (n = 2,469), and Uganda (n = 1,308). More detail on related results can be found in Extended Data Table 1.



Extended Data Fig. 4 | Learning Curve – Improved Implementation and Targeted Instruction Across Trials. **Notes:** This figure uses monitoring data to estimate how targeted instruction was across studies where enough monitoring data existed. Targeted instruction is defined as whether instructors taught students at their level each week, and is then averaged across all eight weeks. For example, if a child did not know addition and they were taught addition, the instruction was well targeted; if a child did know addition but was still taught it rather than moving on to subtraction, instruction was not well targeted. Estimates are also included from the Botswana proof-of-concept

study. Benchmark estimates from control groups of teaching at the right level studies in India are also included to show how often instruction is targeted in the status quo. These estimates track learning outcome progress, with learning improving from Botswana (where instruction was targeted 41.1 percent of the time) to Nepal (where instruction was targeted 50.9 percent of the time) to the Philippines (where instruction was targeted 64.9 percent of the time) to Uganda (where instruction was targeted 81.5 percent of the time). The shaded band represents ± 90 percent confidence intervals. Reproduced from ref. 52; copyright 2026 American Economic Association.



Extended Data Fig. 5 | Maps of the Study Sample by Country.

Extended Data Table 1 | Learning Outcomes Gained, By Country

	(1) Operations Gained	(2) Operations Gained	(3) Highest Proficiency	(4) Highest Proficiency
<i>Panel A: India</i>				
Phone call and SMS	0.220** (0.092) [0.017]	0.210** (0.087) [0.016]	0.115*** (0.036) [0.002]	0.113*** (0.035) [0.001]
Observations	708	708	708	708
Control Mean	2.720	2.720	0.328	0.328
<i>Panel B: Kenya</i>				
SMS messages	-0.027 (0.053) [0.606]	-0.024 (0.053) [0.642]	-0.005 (0.031) [0.865]	-0.003 (0.031) [0.916]
Phone call and SMS	0.114** (0.051) [0.025]	0.092** (0.044) [0.039]	0.071** (0.028) [0.013]	0.059** (0.025) [0.020]
Observations	1985	1985	1985	1985
Control Mean	1.420	1.420	0.623	0.623
P-Val: SMS vs Phone Call + SMS	0.004	0.009	0.011	0.017
<i>Panel C: Nepal</i>				
SMS messages	0.062 (0.086) [0.472]	0.042 (0.076) [0.578]	0.006 (0.022) [0.780]	0.003 (0.020) [0.887]
Phone call and SMS	0.173** (0.081) [0.034]	0.147** (0.073) [0.045]	0.036* (0.021) [0.086]	0.031 (0.019) [0.109]
Observations	2678	2678	2678	2678
Control Mean	1.689	1.689	0.210	0.210
P-Val: SMS vs Phone Call + SMS	0.124	0.107	0.111	0.102
<i>Panel D: Philippines</i>				
SMS messages	0.104 (0.063) [0.100]	0.061 (0.061) [0.318]	0.032*** (0.010) [0.002]	0.024** (0.010) [0.019]
Phone call and SMS	0.561*** (0.067) [0.000]	0.560*** (0.063) [0.000]	0.130*** (0.015) [0.000]	0.130*** (0.014) [0.000]
Observations	2469	2469	2469	2469
Control Mean	1.388	1.388	0.030	0.030
P-Val: SMS vs Phone Call + SMS	0.000	0.000	0.000	0.000
<i>Panel E: Uganda</i>				
SMS messages	0.278*** (0.077) [0.000]	0.289*** (0.076) [0.000]	0.016 (0.021) [0.459]	0.019 (0.022) [0.376]
Phone call and SMS	1.193*** (0.073) [0.000]	1.196*** (0.071) [0.000]	0.344*** (0.026) [0.000]	0.343*** (0.026) [0.000]
Observations	1308	1308	1308	1308
Control Mean	1.803	1.803	0.093	0.093
P-Val: SMS vs Phone Call + SMS	0.000	0.000	0.000	0.000
Controls	No	Yes	No	Yes

Notes: This table reports treatment effects on learning outcomes by country. Operations gained refers to the number of operations gained, with each operation scored as an additional level gain (e.g. beginner=0, addition=1, subtraction=2, multiplication=3, and division=4). Highest proficiency refers to the highest level taught and tested in a given country, which is 2-digit division in all countries but Kenya where students were younger and taught up to 2-digit subtraction. Controls include student grade, gender, geographical strata, and baseline learning level. Standard errors are in parentheses and p-values are in square brackets.

Extended Data Table 2 | Learning Outcomes at Scale Across 5 Countries

	Pooled full sample		Weighted by country	
	(1) Learning (SD)	(2) Learning (SD)	(3) Learning (SD)	(4) Learning (SD)
SMS messages	0.078*** (0.028) [0.006]	0.069** (0.027) [0.010]	0.073** (0.028) [0.010]	0.069** (0.027) [0.011]
Phone call and SMS	0.321*** (0.026) [0.000]	0.312*** (0.025) [0.000]	0.341*** (0.028) [0.000]	0.336*** (0.027) [0.000]
Observations	9148	9148	9148	9148
Control Mean	1.726	1.726	1.726	1.726
P-Val: SMS vs Phone Call + SMS	0.000	0.000	0.000	0.000
Country Fixed Effects	Yes	Yes	Yes	Yes
Grade Fixed Effects	No	Yes	No	Yes
Countries Included	All 5	All 5	All 5	All 5

Notes: This table reports treatment effects on learning outcomes for a pooled sample as well as samples weighted by country-arm. Learning refers to how a child scores on four basic numeracy options: no operations correct, addition, subtraction, multiplication, and division (measured on a scale of 0–4 and expressed in terms of standard deviations, standardized using the control group mean and standard deviation at endline). “Pooled full sample” gives equal weight to each observation; consequently, country-treatment arms with larger samples contribute more to the pooled estimate. “Weighted by country-arm” reweights observations by country so that each observed country arm contributes equal total weight. Column (1) measures the learning differences between the treatment and control group, measured in standard deviations, with country fixed effects only. Column (2) includes baseline controls and country and grade fixed effects. Column (3) and column (4) weigh the results by country-arm. Standard errors are in parentheses and p-values are in square brackets. Given learning outcomes are standardized to have a control mean of zero, we include in the control mean row a more intuitive comparison. We include unstandardized average student learning level in the control group at endline. This variable takes the value of 0 for students who got zero operations correct, 1 if students got addition correct, 2 if subtraction was correct, 3 if multiplication was correct, and 4 if students got division correct. For example, in this table, average student levels in the control group at endline were 1.726, which is the equivalent of a student knowing between addition and subtraction level.

Extended Data Table 3 | Learning Outcomes by Multiple Proficiencies

	Division	Innumerate	Other Proficiencies		
	(1)	(2)	(3)	(4)	(5)
	Division	No operations	Place value	Word problems	Fractions
SMS messages	0.029*** (0.010) [0.006]	0.003 (0.014) [0.820]	0.007 (0.015) [0.646]	0.045*** (0.015) [0.003]	0.044*** (0.012) [0.000]
Phone call and SMS	0.135*** (0.011) [0.000]	-0.041*** (0.012) [0.000]	0.047*** (0.014) [0.000]	0.071*** (0.013) [0.000]	0.064*** (0.011) [0.000]
Observations	7163	7163	7163	7163	7163
Control Mean	0.146	0.225	0.639	0.634	0.209
P-Val: SMS vs Phone Call + SMS	0.000	0.000	0.003	0.048	0.070
Country Fixed Effects	Yes	Yes	Yes	Yes	Yes
Grade Fixed Effects	Yes	Yes	No	No	No
Countries Included	4	4	4	4	4

Notes: This table reports learning outcomes across different proficiencies. Column (1) highlights the learning gains in terms of share of students who learned division. Column (2) shows the share of students who could not correctly answer any of the basic numeracy operations (referred to as innumerate). Columns (3), (4), and (5) show the share of students who could correctly answer place-value problems, word problems, and fractions, respectively. The place-value and higher-order questions were not asked of the Grade 1 and 2 students in Kenya as these were not covered in their standard school curriculum at these ages. Standard errors are in parentheses and p-values are in square brackets.

Extended Data Table 4 | Counterfactual condition

	Nepal	Philippines	Uganda	India	Kenya
Panel A: Caregiver composition					
Mother helps most in math	0.46	0.76	0.46	0.38	0.58
Father helps most in math	0.30	0.00	0.21	0.24	0.24
Sibling helps most in math	0.17	0.09	0.16	0.27	0.10
Caregiver has no formal education	0.11	0.02	0.22	0.03	0.01
Caregiver has only a primary education	0.12	0.13	0.49	0.03	0.08
Caregiver has up to a secondary education	0.71	0.41	0.24	0.71	0.85
Panel B: Comparable intervention in status quo					
Sent SMS math problems in past weeks	0.10	0.15	0.10	0.59	0.90
Received any calls from educator in past weeks	0.14	0.22	0.07	0.43	0.65
Panel C: School disruption					
Degree of school closure	1.00	2.00	2.00	1.00	0.00
Panel D: Status quo learning level					
Achieved top operation level	0.21	0.03	0.09	0.33	0.62
Nationally representative equivalent learning level	0.23	0.05	0.15	0.31	0.68
Panel E: Education accessed by households in government delivery					
Any online instruction	0.30	0.09			
Modular digital instruction	0.03	0.11			
Any TV instruction	0.10	0.12			
Any radio instruction	0.02	0.05			
Teacher phone instruction	0.30	0.00			

Notes: This table describes status quo conditions in controls groups. Means are reported on binary variables for nearly all variables, interpreted as percents; the only exception is the degree of school closure which is coded as 0 for mostly open schools during the study, 1 for partially closed school during the study; and 2 for fully closed schools during the study period. The top learning level is simple division in most countries; Kenya is subtraction since enrolled students were younger. The nationally representative equivalent is drawn from national assessments conducted in Nepal, Uganda, and Kenya using the ICAN comparable assessment which the ASER tool and our assessment directly maps onto, and from ASER in India. The Philippines data is drawn from the RMA data for the top level of proficiency.

Extended Data Table 5 | Robustness Checks on Learning Outcomes

	Phone vs In Person	Backchecks		Random Problem	Effort
	(1)	(2)	(3)	(4)	(5)
	Learning (SD)	Addition backcheck	Division backcheck	Learning (SD)	Effort Task
Assessment Mode	0.065 (0.056) [0.244]				
Add		0.473*** (0.085) [0.000]			
Divide			0.655*** (0.030) [0.000]		
Random Order 2				0.051 (0.068) [0.458]	
Random Order 3				0.020 (0.069) [0.775]	
Random Order 4				-0.032 (0.068) [0.641]	
SMS messages					-0.004 (0.019) [0.821]
Phone call and SMS					-0.021 (0.019) [0.257]
Observations	1985	708	708	2670	5048
Control Mean	1.420	0.930	0.309	1.689	0.435
Countries Included	Kenya	India	India	Nepal	3 Countries

Notes: This table reports the robustness checks on learning outcome measurement. Learning refers to how a child scores on four basic numeracy options: no operations correct, addition, subtraction, multiplication, and division (measured on a scale of 0–4 and expressed in terms of standard deviations, standardized using the control group mean and standard deviation at endline). Column (1) shows the results on average student level for assessments via phone relative to the same assessment in-person. Column (2) and (3) show backcheck results on addition and division. Column (4) shows results of randomizing students to receive different problems of the same proficiency. Column (5) shows results from effort questions. Standard errors are in parentheses and p-values are in square brackets. In column 1, given learning outcomes are standardized to have a control mean of zero, we include in the control mean row a more intuitive comparison. We include unstandardized average student learning level in the control group at endline. This variable takes the value of 0 for students who got zero operations correct, 1 if students got addition correct, 2 if subtraction was correct, 3 if multiplication was correct and 4 if students got division correct.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study. For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a | Confirmed |
|-------------------------------------|--|
| ☐ | ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☐ | ☒ A description of all covariates tested |
| ☐ | ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Anonymized replication data is available here: <https://github.com/nangrist/mobile-education>.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	<i>Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.</i>
Data exclusions	<i>Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.</i>
Replication	<i>Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.</i>
Randomization	<i>Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.</i>
Blinding	<i>Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.</i>

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	<i>We conduct a 5-country randomized controlled trial from December 2020-July 2022. The study design and data analysis are quantitative.</i>
Research sample	<i>The study took place in India (Telangana), Kenya, Nepal, Philippines, and Uganda. The samples were drawn from regions in each country chosen by the government to be representative of high needs regions. In our five-country sample, the median student is in grade 3, 51 percent of students are female, and 46 percent did not know basic operations at baseline.</i>
Sampling strategy	<i>The overall sample size, pooling all sites, is over 16,000 households. For each wave of data collection, we randomly sampled households to interview. Samples were drawn from regions and districts chosen by government partners to be representative of high needs regions. The samples were relatively representative, covering a wide geographic distribution as shown in the Supplementary Information. Sample sizes were ~ 400-800 households per group, based on power from a proof of concept study in Botswana.</i>
Data collection	<i>In each country, we have two waves of data: baseline and endline. These questions included a learning assessment, child wellbeing, parental engagement in educational activities, and parental perceptions of their child's learning. A portion of the survey was conducted with the parent, and learning outcomes were collected by directly assessing the child over the phone. A set of common core questions in the baseline and endline surveys are included in the Supplement. The learning assessment was adapted from the ASER test. The test consists of multiple numeracy items, including two-digit addition, subtraction, multiplication, and division problems. In addition, we asked students to solve a word problem and a fraction problem. Enumerators were either randomly assigned to survey participants or blinded.</i>
Timing	<i>Across all countries, the study took place from December 2020-July 2022, starting in Kenya, then Nepal, India, then Philippines and then Uganda.</i>
Data exclusions	<i>No data was excluded.</i>
Non-participation	<i>We find a response rate at endline of approximately 76%, a high response rate in the literature. We find response rates are equivalent across treatment arms (and thus unbiased by treatment).</i>
Randomization	<i>Every country trial has a control group and a combined phone call and SMS treatment. All studies except India included an SMS-only treatment arm. In two countries, Nepal and the Philippines, we further randomized delivery of the phone call by NGO instructors or government teachers to assess scalability within government systems. In most countries the unit of randomization was the household level, which was de facto the same as the student level since one student participated per household; in a subset of treatment comparisons it was the school-grade level as described in the Methods section.</i>

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.

Validation

Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

State the source of each cell line used.

Authentication	<i>Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.</i>
Mycoplasma contamination	<i>Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.</i>
Commonly misidentified lines (See ICLAC register)	<i>Name any commonly misidentified cell lines used in the study and provide a rationale for their use.</i>

Palaeontology and Archaeology

Specimen provenance	<i>Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.</i>
Specimen deposition	<i>Indicate where the specimens have been deposited to permit free access by other researchers.</i>
Dating methods	<i>If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.</i>
☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.	
Ethics oversight	<i>Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.</i>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	<i>The study took place in India (Telangana), Kenya, Nepal, Philippines, and Uganda. The samples were drawn from regions in each country chosen by the government to be representative of high needs regions. In our five-country sample, the median student is in grade 3, median age is 9, 51 percent of students are female, and 46 percent did not know basic operations at baseline.</i>
Recruitment	<i>In total, over 16,000 households were enrolled in the multi-country evaluation. In each household a primary school student and a caregiver were identified to receive the intervention. We asked households to nominate the main caregiver who provides educational support to their child. Recruitment was done through national and local ministries of education, who nominated representative high-needs regions and schools and who contacted schools who in turn contacted households. We compare covariates to representative samples and find broadly representative sample characteristics.</i>
Ethics oversight	<i>IRB approval was granted by HML IRB (889OXFD21) and Columbia University IRB (21-384). Informed consent was performed in accordance with all relevant guidelines and regulations and obtained from caregivers, child assent, and government approvals in each country. An additional country-specific IRB approval was granted by Nepal's Health Research Council (776/2020) due to local regulations. The trials were registered in the AEA registry: AEARCTR-0009779. After the study period, control group students were later able to be enrolled in the intervention to benefit from the program.</i>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	<i>The trials were registered in the AEA registry: AEARCTR-0009779</i>
Study protocol	<i>The trials were registered in the AEA registry: AEARCTR-0009779. Given the speed and timing of the studies conducted during a historic crisis, rather than register new analysis plans, we rely on specifications published in prior related studies to guide our analysis</i>
Data collection	<i>The study took place in India (Telangana), Kenya, Nepal, Philippines, and Uganda. The samples were drawn from regions in each country chosen by the government to be representative of high needs regions. We conduct a 5-country randomized controlled trial from December 2020-July 2022.</i>
Outcomes	<i>In each country, we have two waves of data: baseline and endline. These questions included a learning assessment, child wellbeing, parental engagement in educational activities, and parental perceptions of their child's learning. A portion of the survey was conducted with the parent, and learning outcomes were collected by directly assessing the child over the phone. The learning assessment was adapted from the ASER test.</i>

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

- | No | Yes |
|--------------------------|---|
| ☐ | ☐ Public health |
| ☐ | ☐ National security |
| ☐ | ☐ Crops and/or livestock |
| ☐ | ☐ Ecosystems |
| ☐ | ☐ Any other significant area |

Experiments of concern

Does the work involve any of these experiments of concern:

- | No | Yes |
|--------------------------|--|
| ☐ | ☐ Demonstrate how to render a vaccine ineffective |
| ☐ | ☐ Confer resistance to therapeutically useful antibiotics or antiviral agents |
| ☐ | ☐ Enhance the virulence of a pathogen or render a nonpathogen virulent |
| ☐ | ☐ Increase transmissibility of a pathogen |
| ☐ | ☐ Alter the host range of a pathogen |
| ☐ | ☐ Enable evasion of diagnostic/detection modalities |
| ☐ | ☐ Enable the weaponization of a biological agent or toxin |
| ☐ | ☐ Any other potentially harmful combination of experiments and agents |

ChIP-seq

Data deposition

- ☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session

(e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

- Sample preparation *Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
- Instrument *Identify the instrument used for data collection, specifying make and model number.*
- Software *Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
- Cell population abundance *Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
- Gating strategy *Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*
- ☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

- Design type *Indicate task or resting state; event-related or block design.*
- Design specifications *Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
- Behavioral performance measures *State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

- Imaging type(s) *Specify: functional, structural, diffusion, perfusion.*
- Field strength *Specify in Tesla*
- Sequence & imaging parameters *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
- Area of acquisition *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*
- Diffusion MRI ☐ Used ☐ Not used

Preprocessing

- Preprocessing software *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
- Normalization *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
- Normalization template *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
- Noise and artifact removal *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See [Eklund et al. 2016](#))

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a | Involved in the study

- ☐ ☐ Functional and/or effective connectivity
☐ ☐ Graph analysis
☐ ☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.