Building Resilient Education Systems: Evidence from Large-Scale Randomized Trials in Five Countries

Noam Angrist, Micheal Ainomugisha, Sai Pramod Bathena, Peter Bergman, Colin Crossley, Claire Cullen, Thato Letsomo, Moitshepi Matsheng, Rene Marlon Panti, Shwetlena Sabarwal, Tim Sullivan*

Abstract

Education systems need to withstand frequent shocks, including conflict, disease, natural disasters, and climate events, all of which routinely close schools. 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. We present results from large-scale randomized trials evaluating the provision of education in emergency settings across five countries: India, Kenya, Nepal, Philippines, and Uganda. We test multiple scalable models of remote instruction for primary school children during COVID-19, which disrupted education for over 1 billion children worldwide. Despite heterogeneous contexts, 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 standard deviations. These effects are highly cost-effective, delivering up to four years of high-quality instruction per $100 spent, ranking in the top percentile of education programs and policies. In a subset of trials, we randomized whether the intervention was provided by NGO instructors or government teachers. Results show similar effects, demonstrating effectiveness within government systems. These results reveal it is possible to strengthen the resilience of education systems, enabling education provision amidst disruptions, and to deliver cost-effective learning gains across contexts and with governments.

Keywords: Human Capital, Education Systems, Education in Emergencies, Scale, COVID-19

*We are grateful to an incredible coalition of partners who enabled this multi-country response. Implementing and research partners include: Youth Impact, J-PAL, Learning Collider (supported by Schmidt Futures and Citadel), Oxford University, World Bank, Ministry of Education, Science and Technology of Nepal, Teach for Nepal, Street Child, Department of Education in the Philippines, IPA, Building Tomorrow, NewGlobe, Alokit, and Global School Leaders. Funding partners include UBS Optimus Foundation, Mulago Foundation, Douglas B. Marshall Foundation, Echidna Giving, Stavros Niarchos Foundation, Jacobs Foundation, JPAL Innovation in Government Initiative, Northwestern's 'Economics of Nonprofits' class, the Peter Cundill Foundation, and the World Bank. Cullen received funding from the UKRI GCRF & Oxford's OPEN Fellowship. We thank Natasha Ahuja, Amy Jung, and Rachel Zhou who provided excellent research assistance. Special thank you to implementation and research teams including Nassreena Baddiri, Mariel Bayangos, Kshitiz Basnet, John Lawrence Carandang, Clotilde de Maricourt, Pratik Ghimire, Faith Karanja, Manoj Karki, Karisha Cruz, Usha Limbu, Roger Masapol, Edward Munyaneza, Sunil Poudel, Karthika Radhakrishnan-Nair, Pratigya Regmi, Uttam Sharma, Swastika Shrestha, and Kishor Subedi. Thank you for thoughtful comments from participants of various seminars and conferences: ASSA, BREAD, CSAE, NEUDC, PacDev. IRB approval was granted by the Nepal Health Research Council (776/2020), HML IRB (889OXFD21), and Columbia University IRB (21-384). The trials were registered in the AEA trial registry: AEARCTR-0009779.

I Introduction

More than 2 billion people live in countries affected by emergencies that frequently disrupt education. The causes of these disruptions are numerous and far-ranging, including rainy seasons, floods, pollution, elections, teacher strikes, conflict, climate shocks, disease, and natural disasters. For example, the spread of Ebola in West Africa in 2014 disrupted both education and health systems (Christensen et al. 2021), closing school for 1.7 million children for 9 months in Sierra Leone. Monsoon rains and flooding in Bangladesh, India, and Nepal from 2017 to 2019 closed 15,000 schools. Earthquakes in Pakistan, Haiti, and Nepal destroyed tens of thousands of schools.

While frequent and disruptive, education emergencies have historically been understudied. We present a new database documenting just how frequent and disruptive such shocks can be.[1] Schools close for lengthy periods during these emergencies, and learning loss can be substantial (Andrabi, Daniels, and Das 2021; Lichand et al. 2022; Carlana, La Ferrara, and Lopez 2023). These shocks exacerbate a pre-existing learning crisis, with fewer than half of primary age students in low- and middle-income countries able to read a story or perform two-digit math operations (World Bank 2018). Our baseline data reinforces this: only 7 percent of students in our sample can do basic division. Moreover, school breaks, while not emergencies, occur routinely during holidays and for extended periods over summer months, resulting in substantial learning loss (Cooper et al. 1996). Resilient education systems need to withstand these frequent disruptions and continue to provide education. International aid organizations refer to many of these situations as "education in emergencies." Education Cannot Wait, the United Nation's Global Fund for education in emergencies, estimates that 222 million children are in active need of education in emergencies programs.

Ideally, interventions that promote learning in emergencies are 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 of diverse educational and cultural backgrounds. Understanding which approaches can be effective across these diverse 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. Most evaluations have been qualitative, with no multi-country experimental studies to date.

In this paper, we evaluate a set of education in emergencies programs to promote learning during large-scale school disruptions caused by COVID-19, which affected over 1 billion children worldwide. We conduct five randomized controlled trials, including multiple delivery models, such as NGO teacher aides as well as government teachers, to test scalability within government systems. Our trials were conducted across five countries: India, Kenya, Nepal, Uganda, and the Philippines.

[1] Figure 1 documents the extent of school disruption for a selected set of emergencies over the last two decades. Notable examples include large earthquakes in Pakistan in 2005 and in Haiti in 2010, elections in Nepal in 2017 and air pollution in 2021, Ebola in Liberia and Sierra Leone, floods in Bangladesh in 2007 and Monsoon rains in 2019, prolonged droughts in Kenya in 2017, and foot and mouth disease in 2004 in Cambodia, among others. Conflicts in countries ranging from Afghanistan, Colombia, Ethiopia, Syria, and Myanmar have also been documented with substantial negative effects on education. For example, between 2015 and 2019 alone, more than 11,000 attacks on schools were documented in at least 93 countries (GCPEA, Education under Attack 2020).

In all settings, schooling was disrupted, 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 cheaply reach students at scale in low-resource contexts. While less than 15 percent of households have access to the internet in low-income countries, over 70 percent have access to mobile phones (Carvalho and Crawford 2020). 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. 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 minute phone call tutorials for eight 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 low-cost, high-frequency assessments. This approach builds on effective targeted instruction approaches both in-person and using technology (Banerjee et al. 2007; Banerjee et al. 2017; Muralidharan, Singh, and Ganimian 2019; Duflo, Kiessel, and Lucas 2020).

A proof of concept in Botswana showed phone call tutorials were effective in promoting learning during initial COVID-19 school disruptions (Angrist, Bergman, and Matsheng 2022). However, questions remain on whether this approach can be scaled across contexts and when delivered by governments – a pervasive challenge for social programs (List 2022; Mobarak 2022). This paper addresses the critical question of which types of education in emergency approaches can improve learning across a broad array of settings by conducting large randomized trials across five contexts as well as comparing scalable models, such as NGO and government delivery.

Our results show consistently large and robust effect sizes of phone call tutorials on learning across contexts, with average effects across all five countries of 0.30-0.35 standard deviations. We find results are largest in countries that experienced the longest school closures: Uganda and the Philippines. These results translate into large learning gains in absolute terms. In Uganda, for example, less than 20 percent of grade 4 students can divide at baseline, but by endline, nearly 50 percent can. These gains fully recover learning losses in math and enable substantial progress beyond status quo learning rates. On average, across all countries, we find a 65 percent increase in the share of students who learn division. The effects are largest for students whose caregivers have only a primary education (rather than secondary and beyond), suggesting that results are strongest when there are fewer alternative educational support systems at home. Additional results show positive effects of phone call tutorials on learning higher-order competencies, such as fractions. Since fractions were not directly taught during the intervention, this provides evidence that learning extended beyond familiarity with the content taught. It further reveals dynamic complementarities, with the benefits of learning basic numeracy accruing to learning additional skills (Cunha and Heckman 2007).

We randomized whether the intervention was provided by NGO instructors or government teachers in the Philippines and Nepal. The average effect of phone call tutorials on learning when

delivered by NGOs is 0.26 standard deviations and 0.31 when delivered by government teachers. These results show similar and statistically indistinguishable effects, indicating that government systems can effectively deliver these types of education in emergency responses. We also find high engagement across sites ranging from 70 to 80 percent, revealing the accessibility and robustness of the approach even in disrupted and low-resource environments.

We further embedded a study randomly allocating a subset of government teachers to deliver phone call tutorials in Nepal. This experimental variation enables detection of the impact of delivering the program on teacher beliefs and practices – an effect that could spill over into the education system and persist beyond the intervention. Results shows teacher practices shift substantially; teachers are 9.3 percentage points more likely to target their feedback to students' learning level. Teachers are also more likely to get parents involved in education. 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. These results suggest that delivering effective programs can unlock a virtuous cycle within government education systems, in turn motivating teachers to want to teach and to improve their teaching practice. Of note, the teachers in this trial closely resemble the average primary school teacher, suggesting government delivery results are likely to translate to a broad set of teachers in the education system.[2]

In contrast to phone call tutorials, the effects of SMS messages alone are mixed. On average, we find a 0.08 standard deviation effect on learning. While these effects are positive and statistically significant when pooled across contexts, they are not consistently statistically significant by country. Average effects are driven by substantial impacts in Uganda, with a 0.20 standard deviation effect that is significant at the 99 percent level, as well as effects in the Philippines, with an effect of 0.09 that is significant at the 90 percent level; there is no effect in Kenya or Nepal. These results suggest that SMS messages can work in contexts with the largest need, such as Uganda and the Philippines, but not in all contexts. Live phone call instruction, on the other hand, appears to best strike the balance of being intensive enough to deliver sustained impact across diverse contexts while remaining cheap and scalable.

Beyond the core set of randomized trials that occurred during COVID-19 school disruptions, an additional education emergency took place during the course of the study: a devastating typhoon in the Philippines that destroyed 4,000 classrooms and disrupted learning for 2 million children (OCHA 2022). The typhoon resulted in a shock which further disrupted schooling and learning. Our initial randomization remained unbiased among groups affected by the typhoon, enabling us to assess effectiveness of phone call tutorials in this additional education emergency context. Results show that the typhoon was associated with reduced learning by approximately 0.12-0.20 standard

[2]Teachers in the study are likely representative of a broader set of teachers since all were government teachers, teachers were directly requested to participate through government protocols by the Ministry of Education, and 80 percent of teachers expressed interest in the program and were enrolled into the eligible pool for the study. Moreover, given in this trial we introduce further randomization, with teachers randomly assigned to implementation, the set of implementers are representative of the overall eligible pool of teachers.

deviations. Results also show that phone call tutorials continued to be effective, with 0.26 standard deviation gains relative to the control group, although SMS messages alone were not enough to stem learning losses. These results suggest that phone call tutorial effectiveness can persist across multiple types of education emergencies.

We also contribute to the development of remote learning assessments. High-frequency remote assessment data enabled real-time targeting of instruction to student levels as well as evaluation of program effectiveness. Even prior to the pandemic, phone calls have been used for household surveys, such as the World Bank Living Measurement Study (LSMS) and UNICEF's Multiple Indicator Cluster Survey (MICS). We conduct five validity checks on high-frequency, low cost learning assessments via phone. A first check compares in-person to phone-based assessment for the exact same set of students in Kenya. We find no statistically distinguishable difference between these two modes of assessments. 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. We further randomize various problems of the same proficiency (e.g., four different questions to measure 2-digit addition with carryover). Results show no difference by question, showing accurate estimates of latent ability. Finally, we include a real-effort question to disentangle effects of the intervention on effort on the test, which has been shown to affect test scores during in-person exams (Gneezy et al. 2019), versus cognitive skills. We find no statistically significant effects of the interventions on effort, revealing that learning gains are indeed a function of cognitive skills.

In addition to assessing impact on learning outcomes, we examine parent beliefs and demand for the intervention. We find parents and children both update their beliefs broadly in line with true learning progress. These results build on a literature exploring the ability to learn not only through being taught but also through noticing real-world progress (Hanna, Mullainathan, and Schwartzstein 2014) as well as a literature on the importance of parents knowing their child's learning level, enabling them to better support their education (Bergman 2021). In addition, parent caregivers demand the program. At endline, 97 percent of parents state they would like to receive the phone call tutorials, which increases even further in the phone call tutorial treatment group, as does willingness to pay for the program.

We also examine impacts on non-cognitive skills, such as perseverance and ambition. We adapt a questionnaire used by Carlana and La Ferrara (2021). While we don't find statistically significant effects for SMS messages on their own, we find sizable effects on both of these outcomes in the phone call tutorials, with up to a 29 percent increase in ambition. We further find positive effects on measures of well-being, such as enjoying school and worrying less. A growing literature highlights the importance of both cognitive as well as non-cognitive skills for future life outcomes (Jackson 2018). These results provide experimental evidence that some educational interventions, such as phone call tutorials, can promote both.

This study contributes to a nascent experimental literature on education in emergencies. Substantial research has taken place on this topic, however much of it has been qualitative or with small samples. One exception is a randomized trial where an NGO provided schooling in rural

areas of conflict-affected regions in Afghanistan and found large effects on learning and closing of gender gaps (Burde and Linden 2013). We build on this literature by providing evidence from randomized controlled trials across multiple contexts and for government delivery models. We also expand the 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. In addition, a large literature documents the cost of school disruptions, such as teacher strikes (Jaume and Willen 2019), earthquakes (Andrabi et al. 2021), schools holidays (Cooper et al. 1996), and COVID-19 (Bacher-Hicks et al. 2021; Jack et al. 2021; Patrinos et al. 2022; Moscoviz and Evans 2022; Carlana, La Ferrara, and Lopez 2023). However, less evidence exists on effective approaches to stem these learning losses. We contribute evidence on scalable solutions to stem learning losses across multiple contexts and through government delivery models. This relates to an emerging evidence base on education interventions tested during COVID-19 (Carlana and La Ferrara 2021; Hassan et al. 2021; Crawfurd et al. 2021; Schueler and Rodriguez-Segura 2021; Lichand et al 2022; Hevia et al. 2022; Angrist et al. 2022).

We also contribute to a growing literature on scale. Recent examples show the extent of the scaling challenge, with many social programs which initially worked in proof-of-concepts no longer delivering impact when scaled or delivered by governments (Mobarak 2022; List 2022).[3] In contrast with many existing studies, our results identify an education approach that can scale across contexts and within government systems.[4] One reason might be that we leverage a particularly scalable technology: mobile phones. Few technologies have as widespread access across so many diverse contexts (Aker and Mbiti 2010; Aker et al. 2012). 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 (Nickow, Oreopoulos, and Quan, 2020; Robinson and Loeb 2021). The phone call tutorials in our study provide a cheap and scalable version of tutoring applicable in low- and middle-income contexts. In addition, since the phone calls are one-on-one and have frequent learning assessments, they enable highly targeted instruction to every child's learning level, another educational approach shown to consistently improve learning outcomes (Banerjee et al. 2017; Muralidharan, Singh, and Ganimian 2019; Duflo, Kiessel, and Lucas 2020).

In addition to identifying a scalable approach, our study advances the scale literature by conducting randomized trials across five countries in a literature where less than 1 percent of RCTs in a set of top economics journal articles are multi-country studies.[5] Some argue that while RCTs

[3]One example includes a contract teacher program in Kenya where effects dissipated when delivered by the government (Bold et al. 2018). Another example includes the diminishing effects of early childhood programs as they were scaled up from an efficacy trial ("proof of concept") in Jamaica, to a pilot in Colombia, to an at-scale program in Peru (Araujo, Rubio-Codina, and Schady 2021).

[4]This is consistent with a set of studies which also focus on targeted educational instruction programs, and have also worked when delivered by governments (Duflo et al., 2020; Banerjee et al. 2017).

[5]Out of a set of 400 papers in development from 2019 and 2021 in a set of top economics journals, 19 percent were RCTs; of those that were RCTs, about 1 percent were multi-country studies. The set of journals considered includes the Top 5 economic journals (American Economic Review, Quarterly Journal of Economics, Econometrica, Journal of Political Economy, and Review of Economic Studies) and other top-tier general interest journals (Review of

address internal validity concerns, they might not address important external validity questions (Pritchett and Sandefur 2015). This study highlights the potential for randomized trials to be conducted and coordinated across contexts – addressing both internal and external validity challenges simultaneously. While rare, this approach is gaining traction, with another prominent example including a multi-country study on microcredit (Banerjee et al. 2015). Finally, by evaluating the effectiveness of government delivery with experimental variation relative to NGO delivery across multiple contexts, we assess questions of scalability within government systems.

Third, we contribute to the global education literature, with a focus on improving learning outcomes. Over the past few decades, education enrollments have improved worldwide yet learning outcomes have barely budged (Pritchett 2013; World Bank 2018; Angrist et al. 2021). Estimates from our baseline survey, for example, show that 36 percent of students in grades 3 to 5 could not do any basic operations, falling well behind grade-level curriculum expectations. 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. In contrast, approaches that improve the quality of teaching, such as teaching at the right level and structured pedagogy, can generate large improvements in learning (Kremer, Brannen, and Glennerster 2013; Ganimian and Murnane 2016; Glewwe and Muralidharan 2016; Eble et al. 2021). The learning gains from the phone-based tutorials tested in this study can deliver up to 4 years of high-quality schooling per $100 spent. These effects rank in the top percentile of cost-effective interventions, benchmarked relative to 150 education policies and programs (Angrist et al. 2020). This reveals the potential of the approach to cost-effectively improve learning across low-and middle-income contexts and to help address a persistent global learning crisis.

The rest of this paper is structured as follows. Section II provides context and presents new data documenting how frequently education emergencies occur. Section III describes the data collected across studies and sites. Section IV describes the experimental design and empirical strategy. Section V includes the results and cost-effectiveness analysis, and Section VI concludes.

II Conceptual Framework

In this section, we outline a framework with two key components underlying the effectiveness of phone-based tutoring: platform and pedagogy. In terms of platform, the program tested reaches students on an accessible, widely available platform: mobile phones. In terms of pedagogy, the program adapts a proven pedagogical approach: targeting teaching to a student's learning level rather than their age or grade. We describe the framework further below, and prior evidence which informs each component, and we again refer to the conceptual framework in the mechanisms section to explore program effectiveness across contexts.

Economics and Statistics, Economic Journal, Journal of the European Economic Association, and all four American Economic Journal journals), and a top field journal (the Journal of Development Economics). Other prominent multi-country RCT efforts include the evaluations of Graduation programs (Banerjee et al. 2015) and Teaching at the Right Level (Banerjee et al. 2017) and some early grade reading interventions (Lucas et al. 2014).

II.A Platform: reach at the right level

The phone-based tutoring program studied leverages the scalable and affordable technology of mobile phones. The program was designed to reach people 'at the right level' on a platform most households have easy access to.[6] Almost 80 percent of households in low- and middle-income countries own a mobile phone, and almost everyone knows someone who has a phone they could borrow if needed. In our five study countries, average mobile phone penetration was 108 mobile phone subscriptions per 100 people (ITU 2021). Coverage is also high, with at least 97% of the world's population covered by at least a 2G mobile network, which is often sufficient to make and receive calls, and a minimum of 93% of households in our study countries.

Given high access to mobile phones, even in low resource settings, this platform presents a scalable way to deliver educational content to households (Aker and Mbiti 2010; Aker et al. 2012). While other technology platforms also present some opportunities for educational instruction, such as television, radio, and online media, access is often lower than through mobile phones. In low income countries, less than 30 percent of households report owning a television, 50 percent a radio, and 20 percent having internet access, suggesting content delivered over these platforms is less likely to have high uptake, particularly amongst the most disadvantaged households (CGD 2020). This is confirmed by households in our sample who report low usage of educational resources delivered via non-phone platforms (see Section III.D for details by country).

Phones also reduce the frictions that may be associated with some other platforms in emergency and low resource settings. For example, given phones are smaller than televisions, radios, and computers, they are often easily transportable in an emergency. Phones can also charge faster and last longer than computers and tablets. Finally, phone call interventions can also offer needed flexibility. While radio and TV programs often require caregivers to find dedicated time in their schedules, tuning into a pre-scheduled television or radio program, phone call programs can provide more options, increasing take-up. Households receive calls from a tutor who solicits their interest in participating, then initiates the tutoring calls and schedules according to a given household's availability, imposing minimal coordination costs on participating households. Moreover, since the tutor calls the household, there is minimal to no cost to the household to participate.

In summary, a key mechanism through which phone call programs may improve learning outcomes is by reaching households 'at the right level' on a platform most households can easily access.

II.B Pedagogy: teach at the right level

There is now a wealth of literature across contexts on the effectiveness of targeting teaching instruction to a student's level rather than their age or grade (Banerjee et al., 2017). The principle of targeted instruction focuses on identifying each student's individual learning level through a light-touch learning assessment. Instructors then use this information to target their instruction and 'teach at the right level.' Targeting instruction stands in stark contrast with business-as-usual

[6] We thank Rukmini Banerji, CEO of Pratham, who inspired the use of the term 'reach at the right level.'

teaching in most education systems. In most systems, teachers typically teach to a one-size-fits-all grade-level curriculum. Yet most children are not at grade level. For example, while grade-level curricula often expect students in grades 3 to 5 to be able to do two-digit division, in our study sample, only 7 percent of students can do so. This phenomenon is pervasive. A survey in Kenya, Tanzania, and Uganda showed that three-quarters of students in grade 3 could not even read a simple sentence, falling well below grade-level expectations (World Bank 2018). In this context, students being taught grade-level material will be left behind and stay behind.

Targeted instruction approaches have received growing attention as highly cost-effective relative to status quo teaching and have been tested using various in-person and in-school models (Banerjee et al. 2017; Muralidharan, Singh, and Ganimian 2019; Duflo, Kiessel, and Lucas 2020; Angrist and Meager 2023). The phone call tutoring program tested in this study adapts the targeted instruction pedagogy for a remote setting, with frequent assessments conducted weekly to enable ongoing, flexible targeted instruction even in disrupted settings. In addition, phone call tutorials mimic some of the targeting benefits of tutoring programs, which are conducted in smaller groups, thus also enabling more targeted instruction to each child's level. Tutoring is an approach most often tested to date in high-income settings (Nickow, Oreopoulos, and Quan, 2020; Robinson and Loeb 2021). We adapt tutoring approaches to low- and middle-income settings, and test a low-cost phone call tutorial model made affordable via use of mobile phones.

The two components outlined in this framework – platform and pedagogy – come together to both reach at the right level as well as teach at the right level. This framework builds on prior evidence, and provides conceptual clarity on the mechanisms underpinning the potential effectiveness of phone call tutorials across diverse and disrupted contexts.

III Context

III.A A global learning crisis, exacerbated by COVID-19

While schooling rates have increased worldwide, students face a global learning crisis, with many students in school but learning very little (World Bank 2018). In our baseline survey, for example, 36 percent of students in grades 3 to 5 could not do any basic operations, falling well behind curriculum expectations. COVID-19 school closures exacerbated this global learning crisis, with school closures forcing over 1.6 billion learners out of classrooms. These types of school closures result in large learning losses which have been documented in North America, Western Europe, Asia, and Sub-Saharan Africa. The lasting economic toll of learning loss is enormous, estimated at over 20 trillion dollars (Azevedo et al. 2020). This pre-existing and exacerbated learning crisis necessitates effective educational approaches that can rapidly improve learning outcomes during and beyond emergencies.

III.B Education in Emergencies

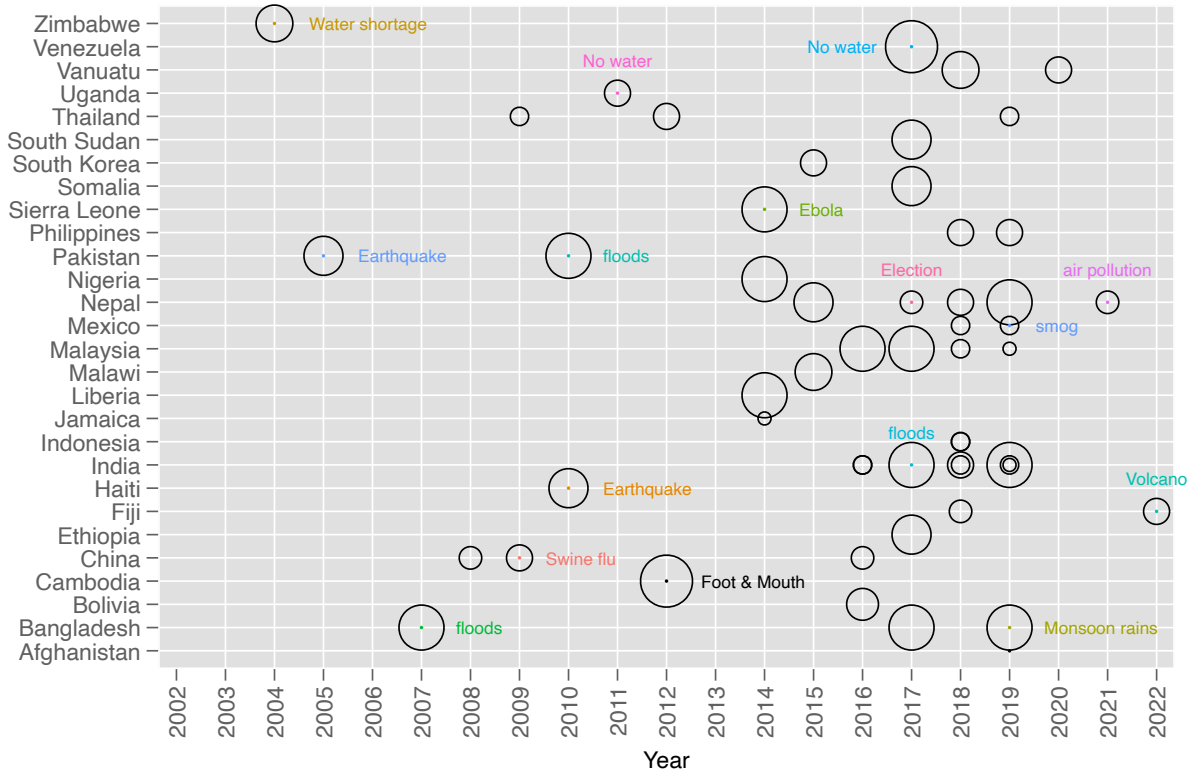
COVID-19 school closures are a recent large-scale example of disruptions to schooling. However, emergencies that affect education are all too common, and affect many people every year. We present novel data documenting the extent of school closures over the past two decades for a subset of countries. We codify information from various sources, including reports from international aid organizations such as Save the Children and UNICEF, the United Nations Office for the Coordination of Humanitarian Affairs (OCHA), as well as national and international news reports.

Figure 1 plots an index of the length of school closures and number of people affected by shocks that have disrupted schooling by country and year. We observe that large-scale disasters affect many people, occur in many countries, and happen at high frequency. The spread of multiple diseases has shuttered schools, including Swine Flu in China in 2009, Foot and Mouth in Cambodia in 2012, and Ebola in Sierra Leone in 2014. In 2019, smog and air pollution from wildfires closed schools around Mexico City. In 2020, a cyclone in Vanuatu destroyed and damaged 885 schools. Extreme climate events often disrupt schooling, such as water shortages in Zimbabwe in 2004 as well as in Venezuela in 2017. Floods and rainy seasons perhaps most routinely cause widespread school disruption. For example, in 2017, monsoons and flooding closed and damaged over 10,000 schools across Malaysia, India, Bangladesh and Nepal. In Pakistan, nearly 8,000 schools were damaged in 2010. A more recent crisis in Pakistan in 2023, which will be included in an extension of this dataset, disrupted 34,000 schools. As climate change-related shocks increase, school disruptions are likely to occur with ever great frequency and consequence, stunting students' education.

In recognition of the importance of addressing shocks to education, the United Nations established a billion-dollar Global Fund called "Education Cannot Wait" in 2016 at the World Humanitarian Summit to support learning in emergencies and protracted crises. Multiple additional international aid structures also exist to support education in emergencies. For example, the Interagency Network for Education in Emergencies (INEE) was established to connect and share best practices amongst more than 18,000 members from 4,000 institutions globally working in education in emergencies, including the United Nations, governments, NGOs, teachers, and students.

A prominent review by Burde et al. (2015) on education in emergencies notes that most evaluations in the sector have been qualitative. The review highlights a major gap in experimental studies and the need for interventions which focus on improving education quality, in addition to education access, in emergency settings. They identify the use of mobile phone technology to deliver educational instruction as a particularly promising yet underutilized approach, and recommend future experimental research in this area.

Figure 1: Documenting a Set of Education in Emergencies (2002-2022)



Notes: This figure 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 larger the bubble the larger either the length of school closure or the number of people affected, or both. More information on the database compilation is available in the Appendix.
Source: Compiled school closure information based on press releases of the United Nation's Office for the Coordination of Humanitarian Affairs (OCHA) ReliefWeb, World Vision, UNICEF, the BBC, and other local outlets.

III.C An Overview of the Study: Randomized Trials Across Five Countries

Our evaluation of education in emergencies interventions took place in the context of COVID-19 school closures across five countries: India, Kenya, Nepal, the Philippines, and Uganda. In total, over 16,000 households were enrolled in this 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: 53.3 percent of nominated caregivers were mothers, 17.4 percent were fathers, 17.3 percent were siblings, and 3.2 percent were grandparents. 24 percent of nominated caregivers had completed primary education or less. Students were enrolled in grades 3 to 5, with one exception in Kenya where students were in grades 1 and 2. The trials took place between December 2020 and July 2022. Appendix Table A9 includes a summary of key facts per country trial, and Figure A3 in the Appendix shows a detailed timeline. Appendix Figure A6 maps the location of the children's schools within several of the study

countries; it shows study participants are distributed across wide areas in most countries, and in some cases, are relatively nationally representative. Students provided phone numbers where they could be contacted, which was most often their parent's phone number, and in some rare cases, the phone of a neighbour or other community member.

Our evaluation includes multiple education in emergencies programs and delivery models. We leverage mobile phones – a high-access, low cost and scalable way to reach students and their caregivers when school is out of session – to provide various educational interventions over the course of eight weeks.[7] One treatment includes a set of SMS messages, such as numeracy content provided weekly, as well as nudges to engage in educational activity. A second treatment added weekly one-on-one 20-minute phone call tutorials. Phone calls covered foundational numeracy and aimed to target instruction to student learning levels via light-touch, high-frequency assessments conducted on a weekly basis. Using the assessment, instructors led students through foundational numeracy practice problems to help them master addition, subtraction, multiplication, and division. For example, students who did not know addition would be taught addition, while students who knew addition but not subtraction would be taught subtraction. Calls were made to caregivers who invited the child in their household to join the program, and would then hand them the phone or use speakerphone for the tutorial session. The implementation ratio of students to teachers was approximately 20:1 for teachers implementing the program full-time, and 5:1 for teachers implementing the program part-time. The phone call program is called "ConnectEd", highlighting phone calls' ability to connect students to quality education even during school disruption.

In addition, we tested several scalable delivery models, including NGO and government teacher delivery, which were randomized within a country. The government delivery arms were conducted with government teachers. They further included headquarter ministry sign off, as well as regional government engagement, including joining training, engagement with routine monitoring data, and periodic mid and high-level steering committee meetings. Some delivery models hired teacher aides to deliver instruction, while others used teachers.[8]

We include a few descriptive statistics to describe our samples. In our five-country sample, the median student is in grade 3 and 51 percent of students are female. On baseline student learning proficiency, 36 percent of students did not know any basic operations, 27 percent were at addition level (meaning they had 'mastered' addition and were being taught the next level), 12 percent were at subtraction level, 17 percent at multiplication level, and 7 percent at division.[9] These low learning levels are well below grade-level curriculum expectations. Moreover, detailed monitoring data shows high week-on-week engagement across sites ranging from 70 to 80 percent, revealing

[7] These types of mobile phone-based interventions have shown to be able to close parent-child information frictions and improve learning in high-income settings (Bergman 2021) as well as low and middle-income settings (Angrist, Bergman, and Matsheng 2022).

[8] Note: we use the terms 'teacher' and 'teacher aide' to reflect several types of trained teaching instructors. Those referred to as 'teachers' are currently employed primary school teachers. 'Teacher aides' include community education volunteers as well as instructors trained to be teachers but currently employed by NGOs. Details of implementer types are discussed in each country section below.

[9] This excludes Kenya, which is the only country where we do not have baseline student level data identical to the other countries broken down by specific proficiency, though we have baseline test score data on a 100-point scale.

the scalability and robustness of the approach even in disrupted and low-resource environments, as shown in Figure A1. We also find sustained rates of engagement across all eight weeks of phone call tutorials staying within a narrow range of high engagement.

This set of five trials sought to scale an approach first tested in Botswana in 2020 (Angrist, Bergman, and Matsheng 2022). The original Botswana trial tested the impact of a phone based tutoring program that had been developed during the first few months of COVID-19. The NGO Youth Impact, one of the largest in the country, developed a phone call that targeted tutoring for remote settings when schools were disrupted. The Botswana study – a tightly controlled initial proof-of-concept – found the program to be cost-effective, resulting in a 0.12 standard deviation improvement in learning outcomes. The present study sought to test the scalability of the approach across five country contexts and a variety of implementer types, including scalability through government delivery models. A rich array of data was collected with response rates above 80 percent, enabling both high-quality evaluation and detailed exploration of outcomes and mechanisms.

In all contexts, the NGO that first developed the program, Youth Impact, provided support to each of the implementing partners in new countries to help train their staff in the program, as well as provide monitoring tools, training, and technical assistance. Support typically involved providing the curriculum and advice if any adaptations were needed for the local context (for example, these ranged from careful language translation to adapting material if place value was taught differently in new contexts); a several-hour training-of-trainers session; monitoring surveys and advice on how to use the monitoring data tools; and occasional meetings during implementation to give advice and troubleshoot. All interactions between the research team and implementing partners took place remotely over video or voice calls. Training of trainers was led by a Youth Impact master trainer for each local implementing partner, who in turn directly trained teachers and tutors in their context and language. Training was practical and interactive, with breakout groups to practice tutoring delivery, followed by live phone calls to practice with real households.

Once implementation began, each week, program managers in each implementing organisation collected and analysed weekly monitoring data submitted by all tutors. This allowed the program coordinators in each implementing organisation to monitor and follow up on the weekly phone call success rate and program delivery. Implementing organisations used this data to provide support to help teachers and tutors troubleshoot, reach more students, improve targeting accuracy, and share lessons and tips across tutors.[10] Youth Impact supported partners across countries, providing a coordinating mechanism to share lessons across trials. However, Youth Impact had no prior presence in these countries, showcasing a replicable model for scale-up in new settings.

A typical phone call involved the following format. The tutor would call the household after arranging a mutually agreeable time. Phone calls were directed toward the student, although caregivers were encouraged to place the phone on speaker and be available to support. Working at the student's identified learning level from their previous week's performance, tutors guided students through a simple set of 3-4 steps on how they should solve the operation being taught that week

[10]See details in Appendix D for additional information on training and implementation.

(addition, subtraction, multiplication, or division). Once the student had walked through a series of problems with the tutor and they had solved a few problems together, then the call concluded with a 'checkpoint question'. This consisted of one math question at the level taught that week. This question enabled tutors to evaluate and update the child's learning level to ensure they could target instruction to the child's correct level the following week. For example, a student who could not do the week's addition checkpoint question would receive instruction in addition in the following week's call; if they got addition correct, the following week they would be taught subtraction.[11]

III.D Specific Country Context for Each Trial

For each country context, we describe the status quo learning levels and COVID-19 education context, the implementing partner and sample description, and the instructor type. Figure A6 characterizes school disruptions in all five countries. In each setting, learning gaps are large both before and during COVID-19, showing a need for more high-quality and scalable educational instruction. In addition, the typical government responses to school disruption, such as radio and TV, have engagement rates below 30 percent.[12] This highlights the need for remote approaches which can yield higher engagement, such as phone call tutorials. Moreover, this puts in context the significance of the high engagement we find of the phone call tutorials in our study, consistently reaching over 70 to 80 percent. Finally, the diverse settings and implementation models in the study, spanning five countries as well as NGO and government delivery, enables us to assess the scalability of targeted phone tutoring across a wide array of contexts and delivery models.

III.D.1 India

Prior to COVID-19, India placed near the bottom of all countries taking the Program for International Student Assessment (PISA), ranking 72nd out of of 73 countries. In our baseline survey, only 24 percent of the Grade 3-5 students in the sample could do division, showing a substantial need for additional foundational numeracy instruction.

During COVID-19, schooling was significantly disrupted in India. In our sample in Telangana state, primary schools were partially closed for the duration of the program (UNESCO 2023). In response to the crisis, the government encouraged households to access educational materials shared on TV, online, and radio.

The implementing NGO in India was Alokit, a non-governmental organisation based in Hyderabad, which is part of an international network of organizations called Global School Leaders. Alokit provides education and school leadership-related programming in schools. The program

[11] The content of the calls and SMSs was focused on teaching the four foundational numeracy skills of addition, subtraction, multiplication and division. This is broadly aligned to a core set of competencies expected to be taught in school in the study countries for students in grades 3 to 5 grades. i.e. by grade 4, most curricula expect students to know two-digit division.

[12] In all contexts, support by governments and school districts during disruption was mixed, with provision of some public education support such as television and radio programming, and encouragement of teachers to support their students remotely, though evidence suggested limited uptake of these services. In this section we outline what services were available and taken up in the countries where we have data.

supported government efforts to ensure all primary school students achieve foundational numeracy skills.

Alokit had pre-existing relationships with government officials and study schools, nearly all of which were government residential schools. The majority of the students were from rural areas and belong to India's most marginalised caste communities. Between two to four teachers were selected from each school to tutor the students. The phone calls were delivered by school teachers based in the same schools as the students in the study. Calls were conducted in either of two languages, English or Telugu, the most common language in the state.

III.D.2 Kenya

Although Kenyan students have higher foundational learning levels than most other countries in Sub-Saharan Africa, learning outcomes still lag far behind those in higher-income countries (World Bank 2020). In third grade, only 47 percent of Kenyan students could solve a second grade mathematics problem (Uwezo, 2016), and 36 percent of grade 2 and 3 students achieved a minimum proficiency level in mathematics (UNESCO, 2016).

By December 2020, when the study launched, schools had been fully closed for 9 months. Schools re-opened from early 2021 onwards, partway through the intervention. The main distance learning programs available to students spanned TV, radio, and online and phone-based educational offerings.

The study took place in 30 of Kenya's 47 counties, in the universe of 112 schools the partner operated. NewGlobe, the partner NGO of the Kenya study, operates one of the largest, low-cost private school networks across the country. In Kenya 33 percent of pre-primary students are enrolled in private schools, and 16 percent of primary students (UNESCO 2022). Kenya was the only context where grades 1 and 2 were included in the program, rather than grades 3 through 5. This was done because the baseline level of foundational numeracy skills was higher than in other contexts, and thus earlier grades were most comparable in terms of baseline learning levels.

The phone call program was implemented by students' normal class teacher, who undertook the phone calls during school closures, as well as after school as part of their teaching duties once schools re-opened part-way through the study. Randomization was at the teacher level, equivalent to the school-grade level. Teachers were advised to deliver tutoring sessions in English, and also conducted some parent interactions in Kiswahili.

III.D.3 Nepal

In 2018, only 28 percent of Grade 5 students in Nepal demonstrated grade-level proficiency in mathematics (NASA 2018; Radhakrishnan et al., 2021). In our baseline survey, only 5 percent of the Grade 3 to 5 students in the sample could do division.

Nepal's schools were closed when the trial started in January 2021. While partial re-openings occurred mid-way through implementation, school disruptions (e.g. frequent closures and reopenings in response to COVID-19 infection waves) were common throughout the study. During this

time, the government rolled out learning programs using radio, television, and online sources, and disseminated printed learning materials to students. However, few students were able to access existing remote educational support materials during school closures. Baseline data showed that in the status quo, only 31 percent of students had teacher interactions during school closure and less than five percent accessed radio or online education (Radhakrishnan et al., 2021). These gaps in remote learning access highlight the need to provide additional support, such as ongoing teacher phone calls to tutor their students.

The study sample included 10 local governments selected by the Ministry of Education, Science and Technology (MoEST) and World Bank, and covered all 7 provinces, with broad geographic spread in the country. The sample included public school students in grades 3 through 5.

To assess scalability, the phone call tutorials were implemented through both a government delivery treatment arm and an NGO delivery arm. Students were randomly assigned to either arm or a control group. The program was implemented by a coalition of partners including MoEST, local governments, the World Bank, Teach for Nepal, and Street Child. In the government arm, public school teachers implemented the program, with teachers teaching students from outside of their own region.[13] In the NGO arm, facilitators trained as teachers delivered the phone call tutoring program. Calls were conducted in over 12 languages, with students matched to teachers who spoke their language.

III.D.4 Philippines

The Philippines scored second-lowest of 79 countries in the 2018 PISA mathematics assessment. During COVID-19, Philippines students also faced one of the world's longest school closures, with schools closed for over two years (UNESCO 2023). Only 2 percent of students in our baseline sample could do division, placing them far behind grade-level expectations; these baseline learning levels are some of the lowest in our study.

At the start of our Philippines study in August 2021, schools had been closed for almost 1.5 years, and they remained closed throughout the trial. During school closures, the government encouraged households to access educational content provided on radio, television, online, and printed materials. However our data reveals limited use of distance educational materials, with fewer than 5 percent accessing radio and TV resources. These gaps, like in other countries, reveal the need for additional high-quality educational support.

The program was implemented in 3 of the country's 17 regions. These regions were selected by the government and represent some of the most marginalized communities in the country. Students from grades 3 and 4 were enrolled in the study.

In the Philippines, like Nepal, the phone call tutoring program was implemented through a government delivery treatment arm and an NGO delivery arm. In the government arm, public

[13] In an additional randomised trial discussed below, we used the over-subscribed list of eligible public school teachers in Nepal to randomly assign teachers to an implementation 'treatment' arm, or control arm. We then also measured causal effects of being an instructor on teacher beliefs and practices.

school teachers employed by the Philippines Department of Education implemented the program. Working with the government, Innovations for Poverty Action, a research NGO, trained government teachers from participating schools on the intervention. The teachers were assigned to call students at their school in the grade they taught. In the NGO-led teacher-aide arm, Innovations for Poverty Action in coordination with the central and regional government offices, hired and trained tutors from the pool of government teacher applicants to implement the program. Calls were conducted in the 5 most commonly spoken languages in the study regions.

III.D.5 Uganda

Ugandan students are well below grade-level expectations in foundational skills. Recent data shows that just 6 percent of grade 4 students are able to read a paragraph and only 2 percent can solve a simple math problem (World Bank 2019). On our baseline survey, only 14 percent of the Grade 3-5 students could do division.

The learning crisis was compounded when, like the Philippines, Uganda entered one of the world's longest COVID-induced school closures, with schools closed for nearly two years from March 2020 to January 2022 (UNESCO 2023). While the government provided distance learning support to households, engagement with these materials was limited (CESS 2021). Only 29 percent of households engaged in radio lessons, 22 percent in printed self-study materials, and under 12 percent in TV and online (Uwezo Uganda, 2021). Like in other contexts, this highlights the need for higher-engagement approaches, such as live phone calls.

The study took place in 9 out of Uganda's 135 Districts, which covered some of the most rural part of the country. The sample included students in grades 3 to 5.

In this study, the implementing organisation was Building Tomorrow, an education-focused NGO in Uganda. Phone calls were made by Building Tomorrow's Community Education Volunteers, resident members of communities who are recruited and trained to serve as grassroots education 'extension agents'. These volunteers encourage out-of-school children to enroll in school, and lead literacy and numeracy lessons in community settings. In Uganda, calls were conducted in one of three languages: English, Luganda, and Runyankore.

IV Data

In each country, we have two waves of data: baseline and endline.[14] The endline surveys took approximately 30 minutes 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.

[14] In Nepal, we conducted a baseline survey with a random 50 percent of the sample, and in Kenya, we relied on administrative data instead of baseline survey data.

Endline surveys were conducted a few months after the program ended. A set of common core questions in the baseline and endline surveys are included in Appendix C.

The learning assessment was adapted from the ASER test, which has been used frequently in the literature to measure learning outcomes (Banerjee et al. 2007; Banerjee et al. 2017) 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 to the ASER test, we asked students to solve a place value word problem, a fraction problem, and an addition word problem to capture learning outcomes beyond a core set of mathematical operations.

To maximize the reliability of the phone-based assessment, we introduced a series of quality-assurance measures. To minimize the likelihood of family members in the household assisting the child, students had a time cap of two 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. A first robustness check compares in-person to phone-based assessment for the exact same set of students.[15] An additional test included back-checks, with a random subset of students tested twice on the same competencies. We further randomized various problems of the same proficiency (e.g. four different questions to measure 2-digit addition with carryover). Finally, we include a real-effort question to disentangle effects of the intervention on effort on the test versus cognitive skills. Prior research suggests effort on the test can affect test scores during in-person exams (Gneezy et al. 2019). Thus, by providing a measure of effort, we can disentangle effort effects versus cognitive skills gains.

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 program. For students, we asked about child's mental wellbeing, how much they enjoy school, and the child's own belief about what math problems they will be able to answer. We also measured non-cognitive skills, such as perseverance and ambition. Finally, we included demographic questions, recording the child's age, grade, and gender.

The overall sample size, pooling all sites, is 16,936 households. For endline surveys, we randomly sampled households from the full sample to interview. This was due to time and cost constraints. In Appendix Table A2, 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,

[15]This quality assurance test was conducted in Kenya. All study students first completed the phone-based endline assessment, followed by the in-person endline assessment. Tests asked identical questions, assessing performance on the same concepts, question types, and difficulty level. This enables us to assess how test administration method affects scores and to include standard in-person assessments in the analysis.

and baseline learning level. The samples randomly selected to be interviewed for endline yield a total endline sub-sample of 12,707.

Appendix Table A3 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 very high at around 80 percent of respondents at endline. Table A3 also presents a test of whether response rates differred by treatment assignment. We find no evidence of differential response rates between treatment and control groups. We also find no evidence of differential attrition in any individual country. 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 program 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 potentially capture persistent effects on educational systems through teachers changing their beliefs and behaviors beyond the lifecycle of the program.

V Experimental Design and Empirical Strategy

We ran randomized controlled trials (RCTs) 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 as follows:

$$Y_{ij} = \alpha + \beta_1 PhoneCalls_j + \beta_2 SMS_j + \gamma X_j + \delta_s + \eta_c + \epsilon_{ij} \quad (1)$$

where Y_{ij} is a learning outcome for individual i in household j. $PhoneCalls_j$ and SMS_j take on the value 1 in their respective treatment arms and 0 otherwise. X_j denotes a vector of baseline control variables to enhance statistical power and precision.[16] δ_s refers to relevant strata in each study, which includes baseline learning, gender, school, and region fixed effects.[17] We both assess impact on specific learning proficiencies as well as on standardized outcomes relative to control group variation. Distributions of baseline and endline learning are shown in Appendix Table A1. Our main specifications pool across studies and we include country fixed effects η_c. We estimate the regression using ordinary least squares (OLS). We run secondary regressions using alternative weights by country as well as country-by-country estimations.

[16] These variables include: student grade and baseline learning levels

[17] The strata used in each trial were as follows. India: baseline level, gender, school; Kenya: school; Nepal: region, baseline level; Philippines: baseline level; Uganda: baseline level, school type; previous education program participation.

We also ran a subset of trials where phone calls were delivered by NGOs or government teachers. We estimate government versus NGO delivery effects using the following specification:

$$Y_{ij} = \alpha + \beta_1 PhoneCalls_j NGO + \beta_1 PhoneCalls_j Gov + \gamma X_j + \delta_s + \eta_c + \epsilon_{ij} \quad (2)$$

In the main analysis of the pooled sample, no adjustments are made to reflect the differences in sample sizes between countries, so every observation is weighted equally. This follows standard practice in the analysis of multi-site RCTs. We test robustness using regressions that instead weight each country equally, and find similar results.

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.[18] The above specifications are the core estimation strategies. Additional specifications include other outcomes of interest beyond learning, such as engagement, as well as regressions run country by country, 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. Appendix Table A5 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 Figure A1.

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. Thus, we can compare control and treatment teachers in the government delivery arm, and whether participating in phone call tutorials changed teachers' subsequent beliefs and behaviors, which we estimate for teachers i at the school level j as follows:

$$S_{ij} = \alpha + \beta_1 TeacherCalls_i + \gamma X_i + \delta_s + \epsilon_{ij} \quad (3)$$

where S_{ij} captures teacher beliefs or instructional practices. These outcomes measure potential

[18] In India, Nepal, and Uganda all arms were randomized at the household level. In Kenya and the Philippines some arms were randomized at the household level (e.g. the NGO arms of the Philippines) and some at the school-grade level (e.g., the Government arm of the Philippines) with corresponding control groups randomized at the same level for each arm. In Kenya, the phone call arm was randomized at the school-grade level while 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 do the fact that there is substantial variation in learning outcomes within a cluster, such that the intracluster correlation is very low at about .03 in the Philippines and Kenya.

spillovers into the education system through teacher beliefs and behaviors which could persist beyond the program.

VI Results

VI.A Main Results: Learning Across Contexts and Scale Delivery Models

Our results show consistently large and robust effect sizes of phone call tutorials on learning across contexts, with average effects across all five countries between 0.30-0.35 standard deviations. Column 1 in Table 1 shows that for our main learning outcome of foundational numeracy skills—measured using average student level — we find large, statistically significant learning differences between treatment and control groups. For the combined phone and SMS group, there was a 0.327 standard deviation ($P < 0.001$) increase in the average numerical operation across all 5 countries. For the SMS messages group, we find smaller but still statistically significant improvements in learning, of 0.083 standard deviations (P=0.003). 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, for example including baseline controls and country and grade fixed effects (Column 2), and weighting results by country-arm (Columns 3 and 4).

These results are large relative to the typical education intervention's effectiveness. A recent review by Evans and Yuan (2022) found that the median effective intervention yielded 0.10 standard deviation gains in learning. Moreover, a review of 150 interventions by Angrist et al. (2020) found that over half of education interventions do not work at all. These reviews put our results in context, with phone call tutorials 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 standard deviations (Angrist, Bergman, and Matsheng 2022). This result contrasts with prior literature showing that proof-of-concept studies rarely scale successfully to new contexts (List 2022) or experience diminishing returns (Caridad, Rubio-Codina, and Schady 2021). Rather than finding diminishing returns, we find results improve as the approach is adapted, scaled, and tested across contexts. A potential explanation is higher need for the intervention, with longer and more disruptive school closures in the multi-country trials with up to two years of school closure.

In Table 2, we show results for the subset of countries (Nepal and the Philippines) where we randomized delivery models to test scalability within government systems. We compare government teachers with NGO delivery. Column 1 shows the pooled results: both government teachers and NGO instructors are effective at improving student learning, with government teachers improving learning by 0.314 standard deviations across both contexts, and NGO teacher aides improving learning by 0.263 standard deviations ($P < 0.001$ for both). The similarity in effect, with no statistically significant difference between models, demonstrates that these education in emergencies programs can be effectively implemented by government teachers.

These results show striking effectiveness when delivered by governments, on par with NGOs.

Prior literature has found that programs often work when delivered by NGOs but fail to replicate when scaled by governments. For example, in Kenya, contract teachers improved learning, but when delivered by the government, program effectiveness waned (Bold et al. 2018). A large-scale government teacher training program in Nepal similarly found no effect, largely due to poor implementation (Schaffner, Glewwe, and Sharma 2021). Our results build on this literature, providing an alternative view: government delivery can achieve large learning gains. Figure 2 visualizes results from Tables 1 and 2, with average effects across studies and split by government versus NGO delivery models.

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.[19] This experimental variation enables detection of the impact of delivering the program 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.

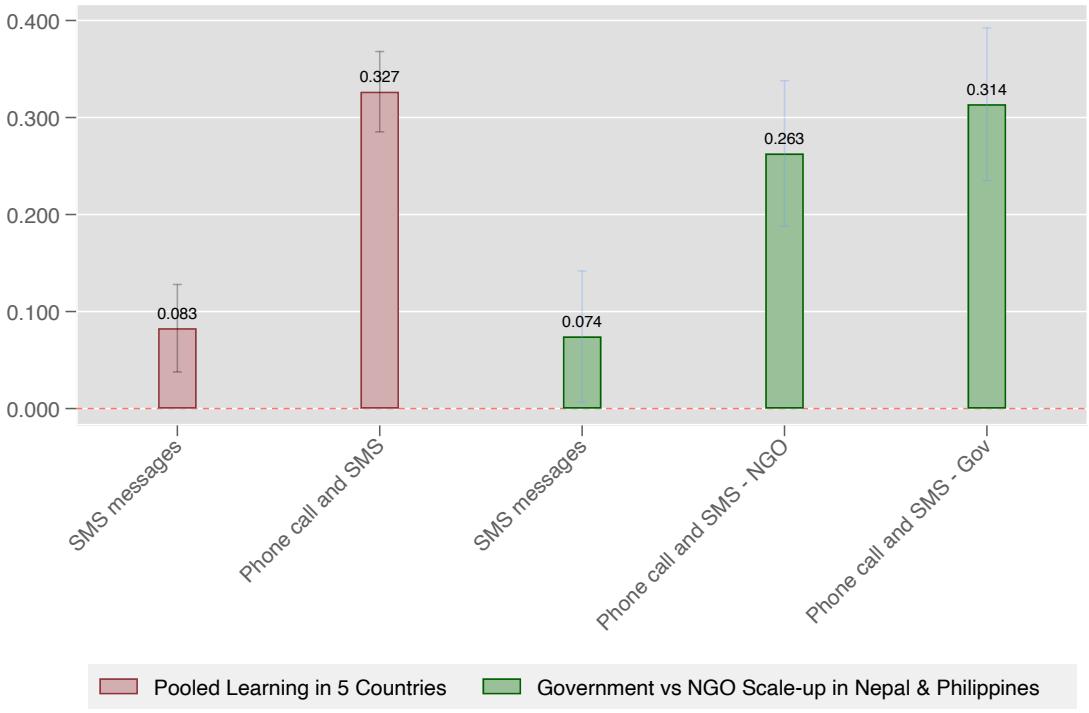
Table 3 shows teacher practices shift substantially; teachers are 9.3 percentage points more likely to target their feedback to students learning level ($P < 0.05$). Teachers are also more likely to get parents involved in education ($P < 0.1$). We further find large effects on teacher perceptions that 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). These results suggest that delivering effective programs could unlock a virtuous cycle within government education systems, in turn motivating teachers to want to teach and to improve their teaching practice.

Table 4 shows results by country. We find that phone call tutorials are effective in every country, showing scalability and effectiveness across diverse contexts. Results are largest in countries which experienced the longest school closures: Uganda and the Philippines. In both countries, where students were out of school for almost two years and the counterfactual for the control group was very limited schooling, effects are extremely large with the phone call tutorials arm producing improvements of 0.891 standard deviations in Uganda and 0.454 in the Philippines on average ($P< 0.001$ for both). These are also the only countries where there were statistically significant learning gains in the SMS-only arm, with 0.207 gains in Uganda and 0.090 in the Philippines. SMS messages have no statistically significant effect in Nepal and Kenya. 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.

.

[19] Over 80 percent of eligible teachers wanted to engage in the phone call tutorials. Of those enrolled in the study, 97 percent of both control and treatment teachers stated that they wanted to deliver the program in the future.

Figure 2: Learning Outcomes (SD) at Scale across 5 Countries and Scaled by Government



Notes: This figure shows treatment effects on learning outcomes. Data are presented as treatment effects relative to the control group ±95 percent confidence intervals. Full statistical results for the treatment effects are presented in Table 1 and 2. Effects are expressed in terms of standard deviations for comparable units. Learning refers to how a child scores on four basic numeracy options: no operations correct, addition, subtraction, multiplication, and division (for which we report the average level on a scale of 0–4). The colors distinguish between the pooled learning in 5 countries and the Government vs NGO scale-up in Nepal and Philippines.

Table A6 shows learning outcomes across specific proficiencies. Results show that the combined phone and SMS arm increases the share of students who get division problems correct by 13.9 percentage points – a 99 percent increase in division (from a control mean of 14.1 percent). These results by proficiency demonstrate that learning outcomes improved substantially in absolute terms, in addition to standardized deviation gains, and across a range of proficiencies. We also find increases in the share of students able to correctly answer place-value problems, and higher-order competencies, such as fractions and word problems. Since fractions were not directly taught during the intervention, this reveals learning extended beyond familiarity with the content taught. These results further reveal dynamic complementarities in skill formation, showing the benefits of learning basic numeracy accrues to learning additional higher-order skills (Cunha and Heckman 2007).

We further explore learning loss and learning recovery along specific proficiencies in Appendix Figure A2. In Uganda, where effects are largest, less than 17 percent of grade 4 students can divide at baseline in the control group. At endline, only 10 percent can divide, showing substantial learning loss. In the treatment group, 48 percent of grade 4 students can divide at endline, fully recovering learning loss due to school closures. Moreover, only 21 percent 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 learning loss recovery, it exceeds typical learning trajectories by nearly an order of magnitude.

In Table A8, we show results on caregivers' time usage, including on overall educational investments and time spent at work. It is possible that by inducing caregivers to support one child's education, they invest less in other children in the households' education. Yet we find evidence of the program crowding in, rather than crowding out, overall time spent on education. The program caused a net increase in the share and frequency of caregivers undertaking educational activities with their children. Moreover, some studies suggest that mothers' engagement in their child's education could crowd out their labor market participation (Evans, Jakiela, Knauer 2021). However, we find limited evidence of such a trade-off, potentially since the program is highly efficient, requiring only a short amount of time to produce large learning gains. These results suggests that highly efficient education programs can deliver large learning gains, with minimal risk of crowd-out of other productive educational and work activities.

VI.B Measuring learning by phone

In addition to the main results of the education in emergency programs, we also 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 (World Bank 2022; Angrist, Bergman, and Matsheng 2022; Gupta et al. 2022; Rodriguez-Segura and Schueler 2022), with emerging evidence suggesting phone assessment can capture meaningful information at high frequency and low cost.

Table 6 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 the exact same students in Kenya. We find no statistically significant difference between these two modes of assessment. 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 over 20. We further randomize students to receive different problems of the same proficiency (e.g., four different questions to measure 2-digit addition with carryover). Results show no difference by question, showing accurate estimates of latent ability. 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.[20] We find no statistically significant effects of the interventions

[20]There is no standard measurement of student effort in the academic literature. However, real-effort tasks range from solving mazes to adding a series of 2-digit numbers. Other proxies of effort include measuring the rate of decline in performance as the test progresses or the effort exerted while filling out an additional survey. Since we aimed to differentiate numerical ability from effort, we chose a problem that required little arithmetic skill, and required non-arithmetic effort.

on effort, revealing that learning gains are largely a function of cognitive skills.

VI.C Results from Another Education Emergency

COVID-19 disrupted education for over a billion children, causing one of the world's largest scale education emergencies. However, as Figure 1 shows, many other education emergencies occur beyond COVID-19. Yet little experimental evidence exists in these settings on how to most effectively promote learning.

In addition to the core set of randomized trials assessing impact of phone-based tutorials during COVID-19 school disruptions, an additional education emergency took place during the course of the study – a devastating typhoon in the Philippines which destroyed 4,000 classrooms and disrupted learning for 2 million children. We collected detailed student-level data on who was affected by the typhoon (called "Typhoon Rai"). The typhoon resulted in an a shock which further disrupted schooling and learning. Our initial randomization remained unbiased among groups affected by the typhoon, enabling us to assess effectiveness of phone call tutorials in this additional education emergency context.[21] The results provide evidence on the disruptive effects of the typhoon on learning and approaches to ensure continuity in learning during disruptions.

We start by estimating learning losses due to Typhoon Rai. Results in Table 7 show that the typhoon was associated with reductions in learning by approximately 0.12-0.20 standard deviations (Column 1 and 2). In addition, since randomization of treatments is orthogonal to which students were affected by the typhoon, we can estimate the impact of the phone-based tutorials during this additional emergency. Results in column (4) which includes controls show that despite the detrimental effects of the typhoon, phone call tutorials continued to be effective, with 0.26 standard deviation learning gains relative to the control group ($P < 0.01$). SMS messages alone were not enough to stem learning losses. These results reveal that the more effective education in emergencies program, phone call tutorials, persist across multiple emergency settings.

VI.D Secondary Outcomes: Parent and Child Beliefs and Non-Cognitive Skills

In addition to learning outcomes, we examine impacts on beliefs about learning. We provide new evidence on children's perceptions of their own learning in Table 8 Column 1, an outcome rarely explored to date. We find children update their beliefs substantially, by 0.21 standard deviations in the phone call and SMS treatment and by 0.04 in the SMS-only treatment - both remarkably similar to gains observed in Table 1. Column 2 examines caregivers' beliefs and shows that they observed their child's learning, updating their beliefs about their child's level of learning by 0.114 standard deviations in the phone call and SMS treatment and by 0.075 in the SMS only treatment – corresponding broadly to gains observed in Table 1. This reveals parents can learn through noticing (Hanna, Mullainathan, and Schwartzstein 2014) although imperfectly and less accurately

[21] There is no statistically significant relationship between randomization to treatment groups and being affected by the typhoon, with p-values of 0.65 and 0.62 for the Phone Call as well as the SMS only treatment, respectively. A full balance table is available on request.

than their children. Moreover, caregivers explicitly state that they think their child's learning has progressed (column 3), with large effects for phone call tutorials, in line with where treatment effects are largest. These results build on a literature exploring parents knowledge of their child's learning level, enabling them to better support their education (Bergman 2021).

Finally, Table 9 examines impacts on non-cognitive skills. A growing literature highlights the importance of both cognitive as well as non-cognitive skills for future life outcomes, such as graduation from college and labor market outcomes (Jackson 2018). We assess impacts on a series of non-cognitive skills, such as perseverance and ambition, in line with Carlana and La Ferrara (2021). We asked students who had just completed a riddle if they wanted to complete another difficult riddle question (perseverance), and if so, whether they wanted an easier or harder one (ambition). We find sizable effects on these outcomes, with 6.2 percentage point gains in ambition, a 29 percent increase relative to the control mean. We further find positive effects on measures of child well-being, such as enjoying school and worrying less, statistically significant at the 95 percent level and above. There was no effect on these non-cognitive skill outcomes for the SMS-only treatment.

These results reveal that while there is often a debate between whether education should focus on cognitive or non-cognitive skill acquisition, they are not mutually exclusive. Indeed, some educational interventions, such as the phone call tutorials tested in this study, can promote both. It also shows that phone-based tutoring models can impart the types of non-cognitive skills that are often viewed as a benefit of brick-and-mortar schools. This suggests that when emergencies do disrupt education, children can still make progress on both cognitive and non-cognitive skills from targeted phone tutoring programs.

VI.E Mechanisms

The effectiveness of the phone call tutorials on learning across diverse settings is striking. In this section, we bring data to bear on two mechanisms underpinning the effectiveness of the approach across contexts, as outlined in the conceptual framework: platform and pedagogy.

VI.E.1 Platform: reach at the right level

As described in the conceptual framework, the program's delivery platform of widely available mobile phones served to enable high take-up and high engagement throughout the program. Data show extremely high consent rates to participate: over 95 percent in most settings. Moreover, engagement in the program was high, with over 95 percent of treatment households in the phone call arm reached during at least one call.

Detailed monitoring data also shows high week-on-week engagement across sites ranging from 70 to 80 percent, as shown in Figure A1. This highlights that the program reached households using a platform they found easy and convenient to access on a regular basis. This contrasts with the low rates of take-up of other platforms, with well below a third of households typically taking up online, television or radio educational resources. In our study contexts, less than 5 percent of the sample had accessed radio or online education resources in Nepal or the Philippines. A study

in Uganda found that 29 percent of students engaged in radio lessons, 22 percent printed self-study materials, and 12 percent accessed online and TV resources (Uwezo Uganda 2021).

High engagement with mobile phone platforms is also demonstrated by caregivers having very high demand for the phone call tutorials even in low resource and disrupted settings. The control means are striking, with 97 percent of parents stating they would like the phone call tutorials, an unusually high level of interest. The phone call tutorials induce even greater interest, bridging the gap to 100 percent interest. Willingness to Pay (WTP) for the program also increases by 4 to 6 percentage points, revealing the potential to stimulate further demand.[22]

This high household consent, engagement, and demand is consistent with the notion that phones provide a widely accessible and convenient platform for households to engage with educational content, enabling households to be 'reached at the right level'.

VI.E.2 Pedagogy: teach at the right level

The pedagogy – teaching at the right level – was as critical as the platform. Phone call tutorials were designed to target instruction to children's learning level. In this section, we examine heterogeneous treatment effects, both across and within contexts, to see how much targeted instruction predicts program effectiveness.

We start by exploring conventional heterogeneous treatment effects in Table A7. As column 3 of Table A7 shows, the program is similarly effective for students across the distribution of baseline learning levels. The consistency of impact regardless of students' starting conditions or characteristics such as baseeline learning levels or gender, also shown in A7, 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, minimal heterogeneity is likely due to the fact that the program was highly targeted, benefiting each child similarly.[23]

Next we explore heterogeneous treatment effects across contexts. We find that effect variation across trials is consistent with variation in targeted instruction. While results show that phone calls were consistently effective across countries, the magnitude of effects increase in tandem with the order of the trial: Kenya, then Nepal, then India, then 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 percent of students in Nepal to 81.5 percent on average in Uganda, shown in Appendix Figure A4. These data reinforce the importance of targeted instruction. As the targeting mechanism improved study after study, this

[22]WTP was measured by asking if a math tutoring program were hypothetically offered to the household in the future, how much would they be willing to pay for it. This outcome is constructed as an indicator variable coded 1 if they would be willing to pay.

[23]The only margin where we find slight heterogeneity is on parental education. Column 1 in Table A7 suggests the program 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.

coincided with the program becoming ever more effective.[24] Equipped with monitoring data and coordinating mechanisms to learn across trials, these results show that programs have the potential to have improved fidelity (e.g. having ever more targeted instruction), becoming higher impact over time as they are scaled and implemented in new contexts.

Altogether, the evidence suggests that both a widely accessible mobile platform and effective targeting pedagogy played important roles in program effectiveness. These mechanisms may explain why the approach tested worked so consistently across diverse contexts while other related education programs that did not rely on these two principles had less impact. For example, Crawfurd et al. (2021) found that phone calls in Sierra Leone that delivered non-targeted lessons based on a mass radio program to students did not improve learning. In this case, it was an accessible platform without targeted pedagogy. To successfully scale across different emergency contexts, these two underlying mechanisms may be key to ensuring education programs improve learning.

VI.F Cost-effectiveness

An important feature of the education in emergencies approaches tested in this study is that they are low cost. The primary tool required to implement the program is a mobile phone, owned by nearly every household in most countries (World Bank 2021). 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, a particularly 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 $16 per child.[25] We benchmark the program's impacts against other education programs using a variety of approaches. First, we compare raw estimates of effectiveness and cost with similar programs, like tutoring and targeted instruction. Phone call tutorials simulate the benefits of one-on-one tutoring, shown to be one of the most effective educational approaches (Nickow et al. 2020). However, many tutoring programs are high cost. For example, a prominent tutoring program yielded 0.19 to 0.31 SD learning gains at a cost of $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.[26]

Second, we use a new cost-effectiveness measure in education that has been estimated for over

[24] An additional potential explanation for explaining the largest program impacts is need: Philippines and Uganda had the longest school closures and in turn had the largest effects in our trials. However, this is unlikely to be the main explanation since our studies and surveys estimate impacts over similar time intervals.

[25] This estimate is even cheaper than estimates in the proof-of-concept study in Botswana due 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.

[26] The SMS-only treatment, effective in two out of 5 countries, and marginally significant in the pooled analysis, is extremely cost-effective (when effective), at 41.1 LAYS per $100. However unlike the Phone and SMS treatment, the statistical significance of the SMS-only arm impacts varied across contexts, with the only significant results observed in two contexts with some of the world's longest school closures: Philippines and Uganda. Given this, it seems plausible that the SMS-only intervention presents 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.

150 impact evaluations in low- and middle-income countries called Learning Adjusted Years of Schooling (LAYS), interpreted as a high-quality year of schooling gained (Angrist et al. 2020). We find the program yields 3.9 LAYS per $100, ranking among the top 10 out of 150 education interventions reviewed.[27] 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 education programs to help address the global learning crisis. Our results suggest phone-based targeted tutoring programs, such as the one tested in this study, have the potential to deliver cost-effective learning gains across contexts and with governments.

VII Conclusion

In this paper, we present results from large-scale randomized trials evaluating the provision of education in emergency programs across five countries: India, Kenya, Nepal, 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 over 1 billion schoolchildren worldwide, as well as during an additional emergency, a typhoon, which further disrupted schooling.

Despite heterogeneous contexts, 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 standard deviations. In the subset of trials where we randomized whether the intervention was provided by NGO instructors or government teachers, we find similar effects, indicating effectiveness when delivered within government systems.[28]

These results have relevance to global efforts to support education in emergencies. Emergencies – including conflict, diseases, natural disasters, and climate shocks – routinely shut down schools, affecting millions of students who forget and forgo learning. During these shocks, alternative models are needed to deliver education. The results presented in this paper show that rigorous testing of programs in humanitarian settings is possible and identify approaches that can scale effectively across contexts and cost-effectively improve learning for students.

Even outside global education emergencies, millions of children worldwide learn very little in school, either because they are taught curricula beyond their learning level, or they are unable to access quality instruction since they live in remote areas. Given widespread mobile phones ownership rates globally, phone-based tutoring programs like the one studied here have the potential to maintain schooling continuity and accelerate learning even outside emergency settings. The low-cost, high-access, and ease of implementation of 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.

[27] This follows the approach used by the Global Education Evidence Advisory Panel by calculating LAYS across diverse types of education programs. This comparison assesses impact on a set of specific numeracy questions. It is not intended to be a comprehensive examination of all content students would learn in school.

[28] Further work could test other dimensions of scale, for example rollouts conducted at scale with millions of children.

VIII References

Aker, Jenny C., and Isaac M. Mbiti. "Mobile phones and economic development in Africa." Journal of economic Perspectives 24, no. 3 (2010): 207-32.

Aker, Jenny C., Christopher Ksoll, and Travis J. Lybbert. "Can mobile phones improve learning? Evidence from a field experiment in Niger." American Economic Journal: Applied Economics 4, no. 4 (2012): 94-120.

Andrabi, Tahir, Benjamin Daniels, and Jishnu Das. "Human capital accumulation and disasters: Evidence from the Pakistan earthquake of 2005." Journal of Human Resources (2021): 0520-10887R1.

Angrist, Noam, Peter Bergman, and Moitshepi Matsheng. "Experimental evidence on learning using low-tech when school is out." Nature human behaviour (2022): 1-10.

Angrist, Noam, Simeon Djankov, Pinelopi K. Goldberg, and Harry A. Patrinos. "Measuring human capital using global learning data." Nature 592, no. 7854 (2021): 403-408.

Angrist, Noam, Michael Ainomugisha, Saipramod Bathena, Peter Bergman, Colin Crossley, Claire Cullen, Thato Letsomo, Moitshepi Matsheng, Rene Marlon Panti, Shwetlena Sabarwal, and Tim Sullivan. 2023. "Learning Curve: Progress in the Replication Crisis." AEA Papers and Proceedings.

Angrist, Noam, David K. Evans, Deon Filmer, Rachel Glennerster, F. Halsey Rogers, and Shwetlena Sabarwal. "How to improve education outcomes most efficiently? A comparison of 150 interventions using the new Learning-Adjusted Years of Schooling metric." (2020).

Angrist, Noam, and Rachael Meager. "Implementation Matters: Generalizing Treatment Effects in Education." Available at SSRN 4487496 (2023).

Araujo, M. Caridad, Marta Rubio-Codina, and Norbert Schady. "70 to 700 to 70,000: Lessons from the Jamaica Experiment." In The Scale-Up Effect in Early Childhood and Public Policy, pp. 211-232. Routledge, 2021.

Azevedo, João Pedro, Amer Hasan, Diana Goldemberg, Koen Geven, and Syedah Aroob Iqbal. "Simulating the potential impacts of COVID-19 school closures on schooling and learning outcomes: A set of global estimates." The World Bank Research Observer 36, no. 1 (2021): 1-40.

Bacher-Hicks, Andrew, Joshua Goodman, and Christine Mulhern. "Inequality in household adaptation to schooling shocks: Covid-induced online learning engagement in real time." Journal of Public Economics 193 (2021): 104345.

Banerjee, Abhijit V., Shawn Cole, Esther Duflo, and Leigh Linden. "Remedying education: Evidence from two randomized experiments in India." The Quarterly Journal of Economics 122, no. 3 (2007): 1235-1264.

Banerjee, Abhijit, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. "A multifaceted program causes lasting progress for the very poor: Evidence from six countries." Science 348, no. 6236

(2015): 1260799.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton. "From proof of concept to scalable policies: Challenges and solutions, with an application." Journal of Economic Perspectives 31, no. 4 (2017): 73-102.

Bergman, Peter. "Parent-child information frictions and human capital investment: Evidence from a field experiment." Journal of Political Economy 129, no. 1 (2021): 286-322.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, and Justin Sandefur. "Experimental evidence on scaling up education reforms in Kenya." Journal of Public Economics 168 (2018): 1-20.

Burde, Dana, and Leigh L. Linden. "Bringing education to Afghan girls: A randomized controlled trial of village-based schools." American Economic Journal: Applied Economics 5, no. 3 (2013): 27-40.

Burde, Dana, Ozen Guven, Jo Kelcey, Heddy Lahmann, and Khaled Al-Abbadi. "What works to promote children's educational access, quality of learning, and wellbeing in crisis-affected contexts." Education Rigorous Literature Review, Department for International Development. London: Department for International Development (2015).

Carlana, Michela, and Eliana La Ferrara. "Apart but connected: Online tutoring and student outcomes during the COVID-19 pandemic." (2021).

Carlana, Michela, Eliana La Ferrara, and Carolina Lopez. "Exacerbated Inequalities: the Learning Loss from COVID-19 in Italy." AEA Papers and Proceedings (2023).

Carvalho, S. Crawfurd, L. School's Out: Now What? (Center for Global Development, 2020); https://www.cgdev.org/blog/schools-out-now-what

CEES. Dissemination on Mitigation and Management of COVID19 Impact on Uganda's Education System (2021). https://cees.mak.ac.ug/dissemination-mitigation-and-management-covid19-impact-ugandas-education-system/

Cooper, Harris, Barbara Nye, Kelly Charlton, James Lindsay, and Scott Greathouse. "The effects of summer vacation on achievement test scores: A narrative and meta-analytic review." Review of educational research 66, no. 3 (1996): 227-268.

Crawfurd, Lee, David K. Evans, Susannah Hares, and Justin Sandefur. Teaching and testing by phone in a pandemic. No. 591. Center for Global Development, 2021.

Christensen, Darin, Oeindrila Dube, Johannes Haushofer, Bilal Siddiqi, and Maarten Voors. "Building resilient health systems: Experimental evidence from sierra leone and the 2014 ebola outbreak." The Quarterly Journal of Economics 136, no. 2 (2021): 1145-1198.

Cunha, Flavio, and James Heckman. "The technology of skill formation." American economic review 97, no. 2 (2007): 31-47.

Duflo, Annie, Jessica Kiessel, and Adrienne Lucas. External Validity: Four Models of Improving Student Achievement. No. w27298. Cambridge: National Bureau of Economic Research, 2020.

Eble, Alex, Chris Frost, Alpha Camara, Baboucarr Bouy, Momodou Bah, Maitri Sivaraman, Pei-

Tseng Jenny Hsieh 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." Journal of Development Economics 148 (2021): 102539.

Evans, David K., Pamela Jakiela, and Heather A. Knauer. "The impact of early childhood interventions on mothers." Science 372, no. 6544 (2021): 794-796.

Evans, David K., and Fei Yuan. "How big are effect sizes in international education studies?." Educational Evaluation and Policy Analysis 44, no. 3 (2022): 532-540.

Glewwe, Paul, and Karthik Muralidharan. "Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications." In Handbook of the Economics of Education, vol. 5, pp. 653-743. Elsevier, 2016.

Gneezy, Uri, John A. List, Jeffrey A. Livingston, Xiangdong Qin, Sally Sadoff, and Yang Xu. "Measuring success in education: The role of effort on the test itself." American Economic Review: Insights 1, no. 3 (2019): 291-308.

Gupta, Saloni, Kumar Satyam, Niharika Gupta and Rahul Ahluwalia (2022). "Can Phone Assessments Generate Reliable Student Learning Data?" Central Square Foundation.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. "Learning through noticing: Theory and evidence from a field experiment." The Quarterly Journal of Economics 129, no. 3 (2014): 1311-1353.

Hassan, Hashibul, Asad Islam, Abu Siddique, and Liang Choon Wang. Telementoring and homeschooling during school closures: A randomized experiment in rural Bangladesh. No. 13. TUM School of Governance at the Technical University of Munich, 2021.

Hevia, Felipe J., Miguel Székely, Tamara Vinacur, and Pablo Zoido. "Tutorías remotas: revisión de la literatura." (2022). Inter-American Development Bank.

Jackson, C. Kirabo. "What do test scores miss? The importance of teacher effects on non–test score outcomes." Journal of Political Economy 126, no. 5 (2018): 2072-2107.

Jack, Rebecca, Clare Halloran, James Okun, and Emily Oster. "Pandemic Schooling Mode and Student Test Scores: Evidence from US School Districts." American Economic Review: Insights (2021)

Jaume, David, and Alexander Willén. "The long-run effects of teacher strikes: evidence from Argentina." Journal of Labor Economics 37, no. 4 (2019): 1097-1139.

Josephson, A., Kilic, T., Michler, J. D. Socioeconomic impacts of COVID-19 in low-income countries. Nature human behaviour (2021), 5(5), 557-565.

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

Lichand, Guilherme, Carlos Alberto Doria, Onicio Leal-Neto, and João Paulo Cossi Fernandes. "The impacts of remote learning in secondary education during the pandemic in Brazil." Nature Human Behaviour 6, no. 8 (2022): 1079-1086.

Lucas, Adrienne M., Patrick J. McEwan, Moses Ngware, and Moses Oketch. "Improving early-grade literacy in East Africa: Experimental evidence from Kenya and Uganda." Journal of

Policy Analysis and Management 33, no. 4 (2014): 950-976.

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

Moscoviz, Laura, and David K. Evans. "Learning loss and student dropouts during the covid-19 pandemic: A review of the evidence two years after schools shut down." Center for Global Development, Working Paper 609 (2022).

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. "Disrupting education? Experimental evidence on technology-aided instruction in India." American Economic Review 109, no. 4 (2019): 1426-60.

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

Patrinos, Harry Anthony, Emiliana Vegas, and Rohan Carter-Rau. "An Analysis of COVID-19 Student Learning Loss." (2022).

Pritchett, Lant. The rebirth of education: Schooling ain't learning. CGD Books, 2013.

Pritchett, Lant, and Justin Sandefur. "Learning from experiments when context matters." American Economic Review 105, no. 5 (2015): 471-475.

Radhakrishnan, Karthika; Angrist, Noam; Bergman, Peter; Cullen, Claire; Matsheng, Moitshepi; Ramakrishnan, Anusha; Sabarwal, Shwetlena; Sharma, Uttam. Learning in the Time of COVID-19 : Insights from Nepal (2021). World Bank, Washington, DC.

Robinson, Carly D., and Susanna Loeb. "High-impact tutoring: State of the research and priorities for future learning." National Student Support Accelerator 21, no. 284 (2021): 1-53.

Rodriguez-Segura, Daniel, and Beth E. Schueler. "Can learning be measured by phone? Evidence from Kenya." Economics of Education Review 90 (2022): 102309.

Schaffner, Julie, Paul Glewwe, and Uttam Sharma. Why Programs Fail: Lessons for Improving Public Service Quality from a Mixed-Methods Evaluation of an Unsuccessful Teacher Training Program in Nepal. No. 1701-2021-3437. 2021.

Schueler, Beth E., and Daniel Rodriguez-Segura. "A Cautionary Tale of Tutoring Hard-to-Reach Students in Kenya. EdWorkingPaper No. 21-432." Annenberg Institute for School Reform at Brown University (2021).

UNESCO. Sustainable Development Goal (SDG) 4 Kenya Country Profile (2016).

UNESCO. COVID-19 Education Response (2023). https://covid19.uis.unesco.org/global-monitoring-school-closures-covid19/country-dashboard/

Uwezo. Are Our Children Learning? Kenya Sixth Learning Assessment Report (2016). Nairobi.

Uwezo. Are our Children Learning? Illuminating the Covid-19 Learning Losses and Gains in Uganda. Uwezo National Learning Assessment Report, 2021 (2021). Kampala.

World Bank. 2018. World Development Report 2018: Learning to realize education's promise.

World Bank. 2019. Uganda Economic Update, 13th Edition, May 2019; Uganda Economic Update, 13th Edition, May 2019 : Economic Development and Human Capital in Uganda - A Case for Investing More in Education. World Bank, Washington, DC.

World Bank. 2021. Mobile cellular subscriptions (per 100 people). https://data.worldbank.org/indicator/IT.CEL.SETS.P2

World Bank. Supporting Student Learning at Home with Phone-based Formative Assessments: Landscape Review. World Bank, 2022.

IX Tables

Table 1: Learning Outcomes at Scale Across 5 Countries

	pooled full sample		weighted by country-arm	
	(1) Learning (SD)	(2) Learning (SD)	(3) Learning (SD)	(4) Learning (SD)
SMS messages	0.083***	0.083***	0.138***	0.139***
	(0.027)	(0.027)	(0.028)	(0.028)
	[0.003]	[0.002]	[0.000]	[0.000]
	{0.0297}	{ 0.5743}	{0.0990}	{0.0495}
Phone call and SMS	0.327***	0.321***	0.414***	0.408***
	(0.025)	(0.025)	(0.026)	(0.026)
	[0.000]	[0.000]	[0.000]	[0.000]
	{0.0099}	{0.0099}	{0.0099}	{0.0099}
Observations	8902	8902	8902	8902
Control Mean	1.318	1.318	1.318	1.318
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 converted to standard deviations, with a mean of 1 in the control group). "Pooled full sample" weighs the observations across all the countries equally. "Weighted by country-arm" accounts for the differences in sample sizes among countries. Column (1) measures the learning differences between the treatment and control group and 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. Curly brackets show the Romano-Wolf P-values to adjust for multiple hypothesis testing for the two treatment arms.

Table 2: Learning Outcomes for Government Scale Models

	All Government Arms	Nepal	Philippines
	(1)	(2)	(3)
	Learning (SD)	Learning (SD)	Learning (SD)
SMS messages	0.074*	0.056	0.090*
	(0.040)	(0.061)	(0.048)
	[0.064]	[0.363]	[0.060]
Phone call and SMS - NGO	0.263***	0.111*	0.434***
	(0.046)	(0.067)	(0.057)
	[0.000]	[0.096]	[0.000]
Phone call and SMS - Gov	0.314***	0.170**	0.482***
	(0.050)	(0.067)	(0.070)
	[0.000]	[0.011]	[0.000]
Observations	4941	2625	2316
Control Mean	1.294	1.283	1.100
P-Val: NGO vs Gov	0.333	0.381	0.534
Country Fixed Effects	Yes	No	No
Grade Fixed Effects	No	No	No
Countries Included	Nepal, Philippines	Nepal	Philippines

Notes: This table reports 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 and converted to standard deviations, with a mean of 1 in the control group). Column (1) pools both countries where we tested implementation scale up in government systems. Column (2) shows the results for Nepal alone, and column (3) shows results for Philippines alone. Standard errors are in parentheses and p-values are in square brackets.

Table 3: Spillovers to the System: 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

Notes: This table shows treatment effects on teacher practices and beliefs. This includes government teachers in Nepal who were randomly assigned to implement the phone call program. We successfully spoke to 83 percent of the government mathematics teachers who were in the schools identified by local governments. Of the eligible grade 3-5 maths teachers we spoke to, 81 percent expressed interest in participating in the study. Of this eligible pool of 301 government teachers, 50 percent were randomly selected to implement. The endline response rate was 96 percent 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 they tailor feedback to their students based on student's 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 maths skills. These are all dummy variables coded as 1 if respondents strongly agreed on a likert scale. Standard errors are in parentheses and p-values are in square brackets. Romano-Wolf P-values to adjust for multiple hypothesis testing are included in curly brackets.

Table 4: Learning Outcomes by Country

	India	Kenya	Nepal	Philippines	Uganda
	(1)	(2)	(3)	(4)	(5)
	Learning (SD)	Learning (SD)	Learning (SD)	Learning (SD)	Learning (SD)
SMS messages		-0.020	0.049	0.090*	0.207***
		(0.039)	(0.065)	(0.048)	(0.057)
		[0.606]	[0.450]	[0.060]	[0.000]
Phone call and SMS	0.212***	0.085**	0.140**	0.454***	0.891***
	(0.067)	(0.038)	(0.061)	(0.050)	(0.054)
	[0.002]	[0.025]	[0.023]	[0.000]	[0.000]
Observations	668	1985	2625	2316	1308
Control Mean	2.091	1.061	1.283	1.100	1.347
P-Val: SMS vs Phone Call + SMS		0.004	0.096	0.000	0.000
Country Fixed Effects	No	No	No	No	No

Notes: This table reports learning outcomes by country. 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 converted to standard deviations, with a mean of 1 in the control group). Standard errors are in parentheses and p-values are in square brackets.

Table 5: Learning Outcomes by Multiple Proficiencies

	Division	Innumerate	Other Proficiencies		
	(1) division	(2) no operations	(3) place value	(4) word problems	(5) fractions
SMS messages	0.030***	-0.007	0.000	0.030**	0.036***
	(0.011)	(0.013)	(0.015)	(0.015)	(0.013)
	[0.007]	[0.586]	[0.986]	[0.046]	[0.004]
Phone call and SMS	0.139***	-0.050***	0.044***	0.063***	0.058***
	(0.011)	(0.011)	(0.013)	(0.013)	(0.011)
	[0.000]	[0.000]	[0.001]	[0.000]	[0.000]
Observations	6917	6917	7163	7163	7163
Control Mean	0.141	0.211	0.629	0.644	0.206
P-Val: SMS vs Phone Call + SMS	0.000	0.001	0.002	0.022	0.058
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.

Table 6: Robustness Checks on Learning Outcomes

	Phone vs In Person	Backchecks		Random Problem	Effort
	(1)	(2)	(3)	(4)	(5)
	Learning (SD)	Add Q2	Divide Q2	Learning (SD)	Effort Task
Assessment Mode	0.063				
	(0.054)				
	[0.244]				
Add		0.473***			
		(0.085)			
		[0.000]			
Divide			0.653***		
			(0.029)		
			[0.000]		
Random Order 2				0.064	
				(0.067)	
				[0.338]	
Random Order 3				0.023	
				(0.067)	
				[0.734]	
Random Order 4				-0.035	
				(0.067)	
				[0.596]	
SMS messages					-0.004
					(0.019)
					[0.822]
Phone call and SMS					-0.021
					(0.019)
					[0.257]
Observations	1985	708	708	2617	5048
Control Mean	1.065	0.930	0.309	1.311	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 converted to standard deviations). 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.

Table 7: Learning During Other Education in Emergencies - Typhoon in the Philippines

	Typhoon Effects			
	(1) Learning (SD)	(2) Learning (SD)	(3) Learning (SD)	(4) Learning (SD)
Affected by Typhoon	-0.198***	-0.116**		
	(0.052)	(0.050)		
	[0.000]	[0.021]		
Control × Typhoon Effect			-0.228**	-0.108
			(0.114)	(0.109)
			[0.046]	[0.321]
SMS messages × No Typhoon Effect			0.071	0.073
			(0.074)	(0.070)
			[0.335]	[0.299]
SMS messages × Typhoon Effect			-0.079	0.008
			(0.086)	(0.082)
			[0.359]	[0.923]
Phone call and SMS × No Typhoon Effect			0.433***	0.439***
			(0.080)	(0.076)
			[0.000]	[0.000]
Phone call and SMS × Typhoon Effect			0.203**	0.259***
			(0.102)	(0.098)
			[0.047]	[0.008]
Constant	1.334***	0.631***	1.166***	0.463***
	(0.028)	(0.080)	(0.062)	(0.096)
	[0.000]	[0.000]	[0.000]	[0.000]
Observations	1647	1647	1647	1647
Countries Included	Philippines	Philippines	Philippines	Philippines
Controls	None	Ed and Bsl Level	None	Ed and Bsl Level

Notes: This table reports the effect of a typhoon on learning outcomes in the Philippines. 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 converted to standard deviations). Column (1) shows the effect on learning of students affected by the typhoon. Column (2) accounts for the baseline distribution and the caregiver's education. Column (3) and (4) show the results of the SMS messages alone as well as phone tutorial treatments with and without controls. Standard errors are in parentheses and p-values are in square brackets.

Table 8: Parent and Child Beliefs

	Beliefs about child's level and progress			Caregiver's program interest	
	(1)	(2)	(3)	(4)	(5)
	Child estimates level	Caregiver estimates level	Child progressed	Wants program	WTP
SMS messages	0.040*	0.075***	0.010	-0.019	0.060***
	(0.024)	(0.024)	(0.015)	(0.012)	(0.020)
	[0.097]	[0.002]	[0.522]	[0.127]	[0.003]
Phone call and SMS	0.210***	0.114***	0.104***	0.034***	0.045**
	(0.023)	(0.023)	(0.014)	(0.011)	(0.020)
	[0.000]	[0.000]	[0.000]	[0.001]	[0.025]
Observations	8798	9188	6188	8918	3777
Control mean	2.347	2.285	0.350	0.968	0.434
P-Val: SMS vs Phone Call + SMS	0.000	0.079	0.000	0.000	0.452
Country Fixed Effects	Yes	Yes	Yes	Yes	Yes
Grade Fixed Effects	Yes	Yes	Yes	Yes	Yes
Countries Included	All	All	4	All	Philippines & Uganda

Notes: This table shows treatment effects on secondary outcomes related to parental and child beliefs, and demand for the program. Column (1) measures the child's belief about their level of learning, from beginner (0) to division (4), in standard deviations. Column (2) measures parents beliefs about their child's level on the same scale. Column (3) shows treatment effects on an indicator variable for caregivers being very confident that their child has progressed in their learning. Column (4) shows effects on a dummy variable that the caregiver would like access to a phone-based maths tutoring program in the future, and Column (5) shows treatment effects on a dummy variable that caregivers would be willing to pay money to access a phone-based maths tutoring program in the future. Standard errors are in parentheses and p-values are in square brackets.

Table 9: Child Non-Cognitive Outcomes

	Non-cognitive Skills		Wellbeing	
	(1)	(2)	(3)	(4)
	Perseverence	Ambition	Enjoys school	Often worried
SMS messages	0.010	0.008	0.011	-0.001
	(0.010)	(0.011)	(0.011)	(0.011)
	[0.308]	[0.465]	[0.310]	[0.957]
Phone call and SMS	0.029***	0.062***	0.023**	-0.020**
	(0.009)	(0.011)	(0.009)	(0.010)
	[0.001]	[0.000]	[0.014]	[0.043]
Observations	8962	8962	8880	8121
Control mean	0.833	0.211	0.821	0.244
P-Val: SMS vs Phone Call + SMS	0.042	0.000	0.201	0.060
Country Fixed Effects	Yes	Yes	Yes	Yes
Grade Fixed Effects	Yes	Yes	Yes	Yes
Countries Included	All	All	All	All

Notes: This table shows treatment effects on children's non-cognitive outcomes, including perseverance and wellbeing. Column (1) measures impacts on a dummy variable for whether the child wanted to answer a second riddle question after answering an initial one, following the Carlana and La Ferrara (2021) measure of perseverance. Column (2) measures impacts on an indicator variable for whether the student wanted to answer a second more difficult riddle question, indicating ambition. This was coded 0 if they didn't want to answer a second riddle question or they wanted an easier question. Column (3) shows effects on a dummy variable for whether the student says they enjoy school very much. Column (4) shows effects on an indicator for whether the child has many worries or is often worried, from the children's "Strengths and Difficulties Questionnaire". Standard errors are in parentheses and p-values are in square brackets.

A Appendix Tables

Table A1: Distribution of learning at baseline and endline across groups

Control Group

Baseline	Freq.	Percent
Beginner	786	30.16
Addition	728	27.94
Subtraction	388	14.89
Multiplication	447	17.15
Division	257	9.862
Total	2606	100

Endline	Freq.	Percent
Beginner	612	22.13
Addition	620	22.42
Subtraction	801	28.96
Multiplication	420	15.18
Division	313	11.32
Total	2766	100

Phone Call Treatment group

Baseline	Freq.	Percent
Beginner	1112	33.93
Addition	886	27.04
Subtraction	438	13.37
Multiplication	563	17.18
Division	278	8.483
Total	3277	100

Endline	Freq.	Percent
Beginner	876	20.63
Addition	633	14.90
Subtraction	1238	29.15
Multiplication	702	16.53
Division	798	18.79
Total	4247	100

Notes: This table reports the distribution of student learning levels by Control group (top two tables) and endline (bottom two tables) at baseline (top of each set) and endline (bottom of each set). Each level is the highest level students achieved. Note that baseline was not conducted in Kenya where exam scores were used instead, and with only a randomly selected half the sample in Nepal. In Kenya, given the sample was in grades 1 and 2, the endline assessment stopped at subtraction level, where the curriculum reached in these grades. Also note that a random sub-set of each study sample was approached to be interviewed at endline.

Table A2: Representativeness of sample at endline

	Baseline Variables				
	(1) Baseline Learning	(2) Grade 3	(3) Grade 4	(4) Grade 5	(5) Student is Female
Randomized to receive endline survey	-0.009	-0.000	-0.003	0.004	-0.005
	(0.021)	(0.006)	(0.006)	(0.005)	(0.010)
	[0.679]	[0.967]	[0.558]	[0.462]	[0.634]
Observations	12707	16936	16936	16936	16936
Full sample Mean	0.00	0.21	0.22	0.17	0.51
Country Fixed Effects	Yes	Yes	Yes	Yes	Yes
Grade Fixed Effects	No	No	No	No	No
Countries Included	All 5	All 5	All 5	All 5	All 5

Notes: This table reports balance on baseline characteristics of the sample that was randomly selected to be interviewed at endline, compared to the complete study sample. Column (1) shows the representativeness by baseline learning, column (2) (3) and (4) by student grade, and column (5) by student sex. These are all of the variables collected across all 5 countries at baseline. On baseline learning, note that in Nepal, a randomly selected half of the sample received a full baseline survey that included learning level. In Kenya, school assessment grades were used in lieu of a baseline levelling assessment, and not all Kenyan students had pre-existing student grades. Standard errors are in parentheses and p-values are in square brackets.

Table A3: Reach and Attrition

	Households Reached		Correlation in Attempts
	(1)	(2)	(3)
	Reached at Endline	Reached at Endline	Baseline Learning
SMS messages	-0.001	-0.001	
	(0.010)	(0.010)	
	[0.913]	[0.912]	
Phone call and SMS	0.008	0.008	
	(0.010)	(0.010)	
	[0.445]	[0.430]	
Number of attempts to reach HH			0.001
			(0.005)
			[0.889]
Observations	12331	12331	7187
Control mean	0.799	0.799	
Country Fixed Effects	Yes	Yes	Yes
Grade Fixed Effects	No	Yes	Yes
Countries Included	All 5	All 5	All 5

This table reports the differences in the households reached. Column (1) measures the households reached at endline. Column (2) measures the same but accounts for the students' grade level. Column (3) shows the effect of the number of attempts to reach a household to the baseline learning of the student. The control means show that on average 80 percent of households were reached at endline. Standard errors are in parentheses and p-values are in square brackets.

Table A4: Balance on baseline characteristics

	Baseline Variables				
	(1) Baseline Learning	(2) Grade 3	(3) Grade 4	(4) Grade 5	(5) Female
SMS messages	0.035	-0.001	0.007	-0.006	0.016
	(0.027)	(0.013)	(0.013)	(0.008)	(0.012)
	[0.191]	[0.955]	[0.571]	[0.393]	[0.181]
Phone call and SMS	0.007	-0.001	0.008	-0.006	0.004
	(0.029)	(0.012)	(0.012)	(0.007)	(0.011)
	[0.801]	[0.906]	[0.520]	[0.352]	[0.725]
Observations	9376	12331	12331	12331	12331
Control Mean	0.07	0.25	0.26	0.24	0.51
P-Val: SMS vs Phone Call + SMS	0.36	0.95	0.96	0.98	0.29
Country Fixed Effects	Yes	Yes	Yes	Yes	Yes
Grade Fixed Effects	No	No	No	No	No
Countries Included	All 5	All 5	All 5	All 5	All 5

This table reports the balance across treatment groups at baseline for the sample randomly selected to be interviewed at endline. Column (1) shows balance on baseline learning, column (2) (3) and (4) by student grade, and column (5) by student sex. These are all of the variables collected across all 5 countries at baseline. On baseline learning, note that in Nepal, a randomly selected half of the sample received a full baseline survey that included learning level. In Kenya, school assessment grades were used in lieu of a baseline levelling assessment, and not all Kenyan students had pre-existing student grades. Standard errors are in parentheses and p-values are in square brackets.

Table A5: Balance on baseline characteristics: Nepal teacher sample

	(1) Female	(2) Years teaching	(3) High school	(4) Bachelors	(5) Masters	(6) Speaks English
Treatment Teacher	0.037	0.567	0.004	0.004	-0.034	0.044
	(0.056)	(1.167)	(0.057)	(0.056)	(0.039)	(0.056)
	[0.504]	[0.627]	[0.947]	[0.941]	[0.380]	[0.434]
Observations	301	301	301	301	301	301
Control mean	0.353	13.387	0.427	0.373	0.147	0.360

This table reports the balance across Nepal's teacher treatment and control group at baseline. Column (1) shows balance on teacher gender. Column (2) shows balance based on teacher's years of experience. Columns (3) to (5) show balance on the share of teachers whose highest level of education completed is high school, bachelors degree, or masters degree, respectively. Column (6) shows balance on the share of teachers who speak English. Standard errors are in parentheses and p-values are in square brackets.

Table A6: Learning Outcomes by All 4 Operations Taught

	(1) Addition correct	(2) Subtraction correct	(3) Multiplication correct	(4) Division correct
SMS messages	-0.005	0.039***	0.067***	0.030***
	(0.012)	(0.013)	(0.015)	(0.010)
	[0.675]	[0.003]	[0.000]	[0.004]
Phone call and SMS	0.041***	0.119***	0.193***	0.138***
	(0.010)	(0.012)	(0.014)	(0.011)
	[0.000]	[0.000]	[0.000]	[0.000]
Observations	9148	9148	7163	7163
Control Mean	0.775	0.586	0.363	0.152
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	Yes	Yes	Yes	Yes
Countries Included	5	5	4	4

Notes: This table reports learning outcomes across the 4 different proficiencies taught in the curriculum and comprising the standardized learning outcome. Column (1) highlights the learning gains in terms of share of students who got addition correct. Columns (2), (3), and (4) show the share of students who could correctly answer subtraction, multiplication, and division respectively. The 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.

Table A7: Heterogeneity

	Caregiver Education	Gender	Baseline Level
	(1)	(2)	(3)
	Learning (SD)	Learning (SD)	Learning (SD)
SMS messages	-0.042	0.110***	0.064**
	(0.111)	(0.039)	(0.028)
	[0.703]	[0.005]	[0.025]
Phone call and SMS	0.206**	0.333***	0.360***
	(0.105)	(0.035)	(0.027)
	[0.050]	[0.000]	[0.000]
Primary	-0.010		
	(0.083)		
	[0.909]		
Secondary Plus	0.304***		
	(0.078)		
	[0.000]		
SMS messages × Primary	0.097		
	(0.127)		
	[0.448]		
SMS messages × Secondary Plus	0.133		
	(0.116)		
	[0.253]		
Phone call and SMS × Primary	0.215*		
	(0.121)		
	[0.075]		
Phone call and SMS × Secondary Plus	0.120		
	(0.110)		
	[0.274]		
Female=1		0.063*	
		(0.036)	
		[0.086]	
SMS messages × Female=1		-0.053	
		(0.053)	
		[0.314]	
Phone call and SMS × Female=1		-0.014	
		(0.048)	
		[0.773]	
Baseline Level			0.234***
			(0.021)
			[0.000]
SMS messages × Baseline Level			0.037
			(0.030)
			[0.204]
Phone call and SMS × Baseline Level			0.040
			(0.027)
			[0.136]
Observations	7794	8902	7035
Control Mean	1.318	1.318	1.318
Countries Included	All 5	All 5	All 5
Country Fixed Effects	Yes	Yes	Yes

Notes: This table reports heterogeneous treatment effects on learning outcomes. 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 converted to standard deviations). Column (1) shows the results between having caregivers who reached primary education or less and caregivers who reached secondary education or more. Column (2) shows the results between males and females. Column (3) shows the results across baseline learning levels. Standard errors are in parentheses and p-values are in square brackets.

Table A8: Potential for Crowdout

	Caregiver supported child educ		Caregiver unemployment	
	(1)	(2)	(3)	(4)
	Did educ activities	Often did educ activities	Unemployed	Mother unemployed
SMS messages	0.044***	0.032**	-0.019	-0.029
	(0.009)	(0.013)	(0.021)	(0.026)
	[0.000]	[0.013]	[0.358]	[0.271]
Phone call and SMS	0.042***	0.029**	0.012	0.019
	(0.008)	(0.012)	(0.019)	(0.024)
	[0.000]	[0.015]	[0.513]	[0.415]
Observations	8899	8899	3681	2446
Control mean	0.865	0.494	0.429	0.429
P-Val: SMS vs Phone Call + SMS	0.785	0.784	0.119	0.059
Country Fixed Effects	Yes	Yes	Yes	Yes
Grade Fixed Effects	Yes	Yes	Yes	Yes
Countries Included	All	All	Philippines & Uganda	Philippines & Uganda

Notes: This table shows treatment effects on outcomes that could indicate the program crowds-out or crowds-in other education and labour market activities. Column (1) measures impacts on a dummy variable for whether the caregiver says they did any educational activities with their child over the past 3 weeks. Column (2) shows how often the parent did any educational activities with their child 3 or more times in the past week. Column (3) shows effects on caregiver unemployment, and column (4) restricts this unemployment estimate to households reporting that mothers are the primary educational caregiver in the household. Standard errors are in parentheses and p-values are in square brackets.

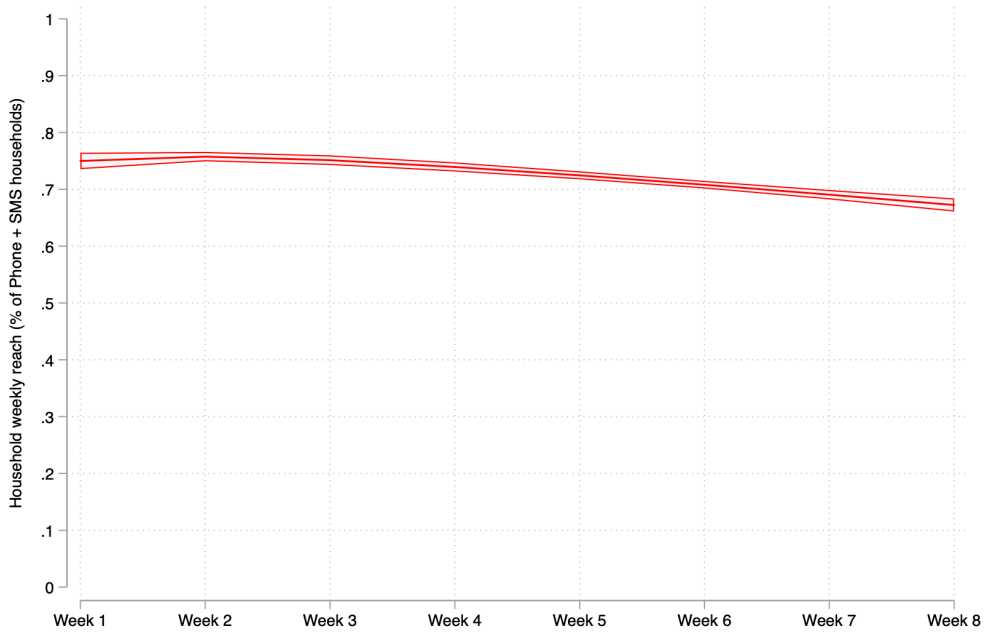
Table A9: Trial Description

	India	**Kenya**	**Nepal**	**Philippines**	**Uganda**
Sample size	850	6724	3732	3492	2138
Randomly selected endline sample size	765	3556	3351	3164	1495
Student grades	3-5	1-2	3-5	3-4	3-5
NGO Delivery	✓	✓	✓	✓	✓
Government Delivery			✓	✓	
Unit of randomization	Household	Cluster - school grade	Household	NGO: Household; Gov: Cluster - school grade:	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; Gov: school, grade, Student baseline level	Student baseline level, Attends government school, previous education program participation
Included Combined Phone Call and SMS treatment	✓	✓	✓	✓	✓
Included SMS only treatment	✗	✓	✓	✓	✓
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)
Number of weeks of implementation	8	12	16	8	8
Dates	Apr 21-Jul 21	Dec 20-Apr 21	Jan 21-Jul 21	Aug 21-Jul 22	Oct 21-Jan 22
Administrative units in study	1 state	30 counties	All 7 provinces	3 regions	9 districts
Student's own teacher delivering	Mixed but same school	✓	✗	Mixed but same school	✗

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

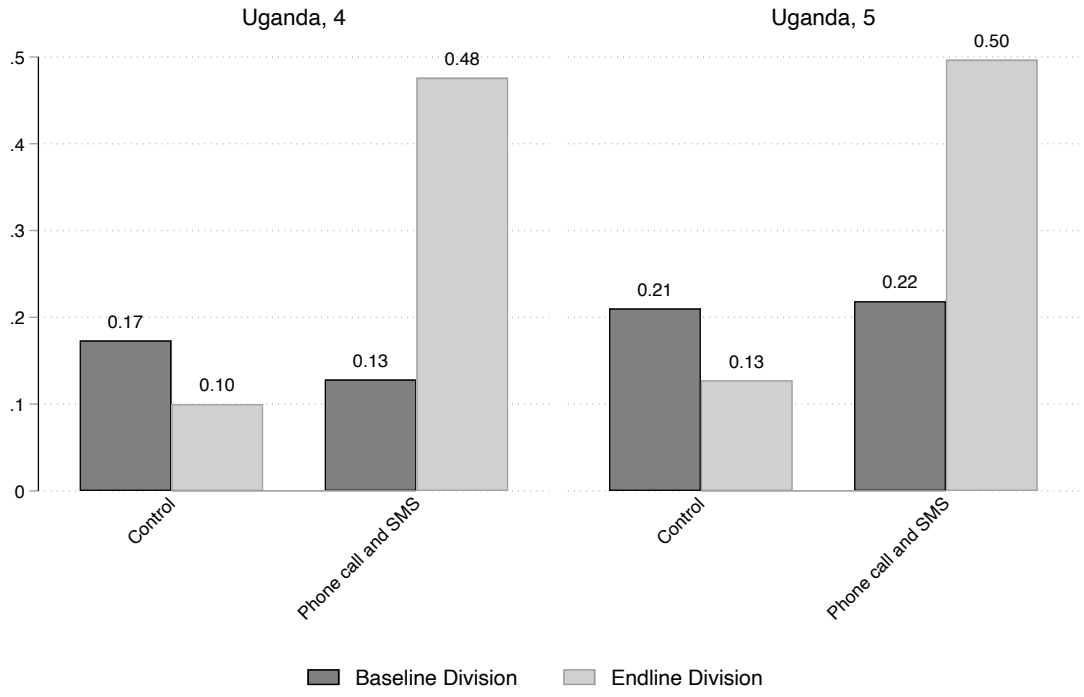
B Appendix Figures

Figure A1: Weekly Household Engagement in Phone Calls



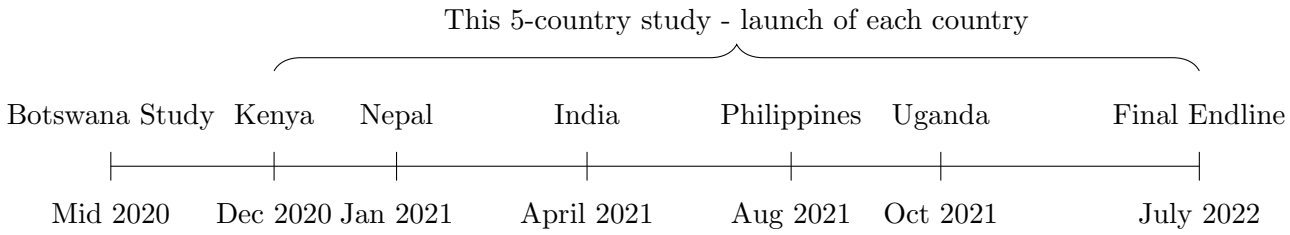
Notes: This figure shows average weekly engagement of households in the phone call and SMS arm. Each week, engagement is coded as 1 if the household answered the call and engaged. This includes data from Nepal, Uganda and the Philippines where we have detailed and easily comparable weekly monitoring data.

Figure A2: Learning Losses & Gains - Can Divide, Uganda, by Grade



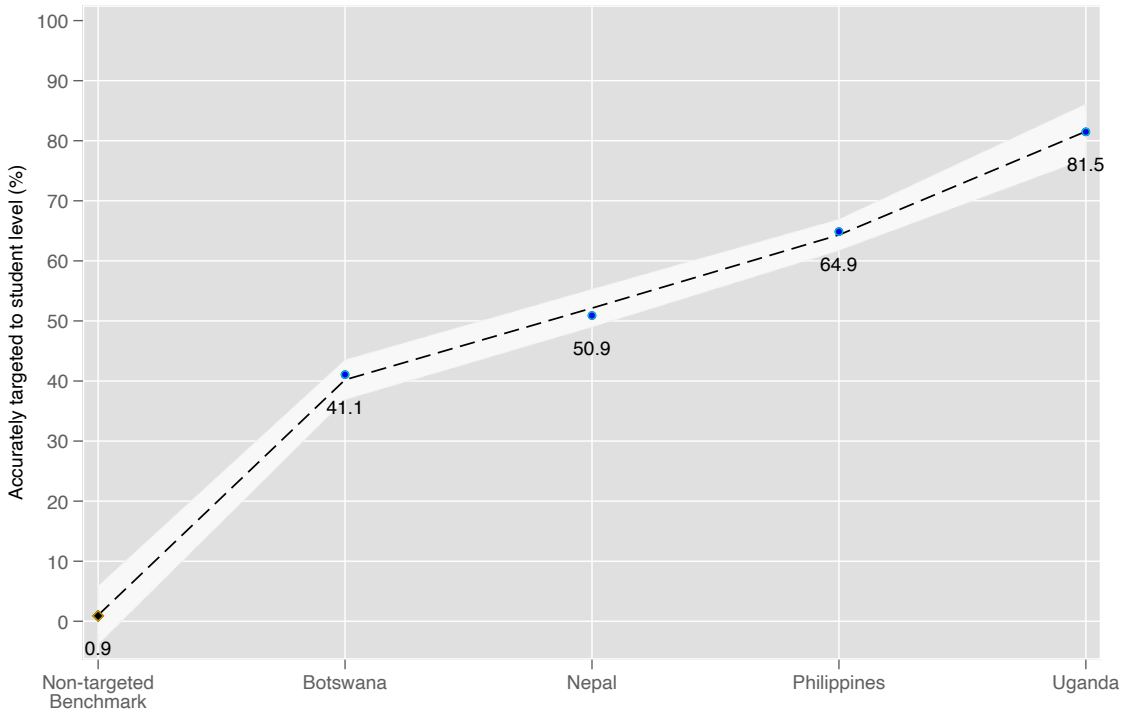
Notes: This figure shows the proportion of Grade 4 and Grade 5 students in Uganda that know how to correctly perform division. The shades distinguish the baseline and endline proportion. In each panel, the two bars on the left show the outcomes for students who did not receive the intervention; the two 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.

Figure A3: Timeline of trials



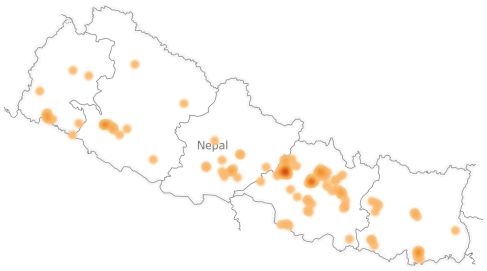
Notes: This figure shows the timing of the implementation across studies. The proof of concept study was conducted in Botswana in 2020. It was followed by five studies across five countries. An endline assessment was conducted a few months after each implementation period ended; the endline for the latest replication was in July 2022.

Figure A4: Learning Curve – Improved Implementation and Targeted Instruction Across Trials

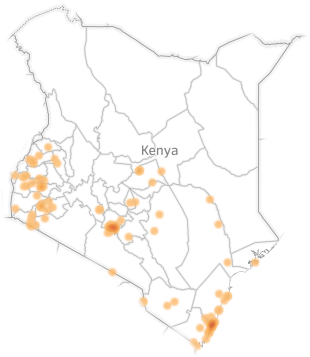


Notes: This figure is adapted from Angrist et al. (2023) in the *AEA Papers and Proceedings* which uses monitoring data to estimate how targeted instruction was across studies where enough monitoring data existed. Targeted instruction is defined as whether instructors taught students at their level each 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 till taught it rather than moving on to subtraction, instruction was not well targeted. Estimates are also included from the Botswana proof-of-concept study (Angrist, Bergman, and Matsheng 2022). Benchmark estimates from control groups of teaching at the right level studies (Banerjee et al. 2017) 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 (0.12 standard deviations when instruction was targeted 41.1 percent of the time) to Nepal (0.14 standard deviations when instruction was targeted 50.9 percent of the time) to the Philippines (0.44 standard deviations when instruction was targeted 64.9 percent of the time) to Uganda (0.88 standard deviations when instruction was targeted 81.5 percent of the time).

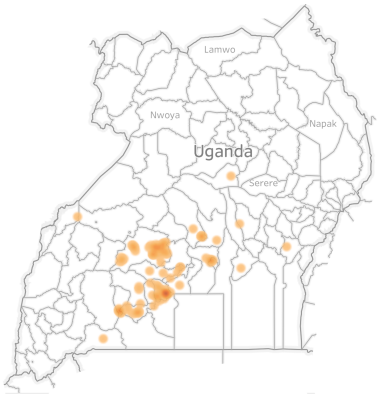
Figure A5: Maps of the Study Sample by Country



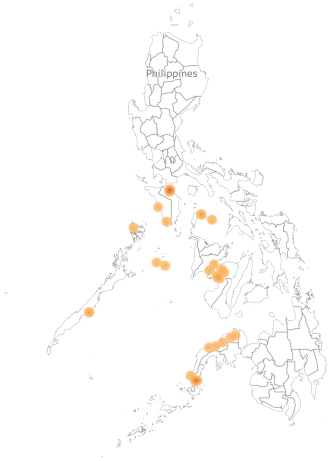
(a) Nepal



(b) Kenya



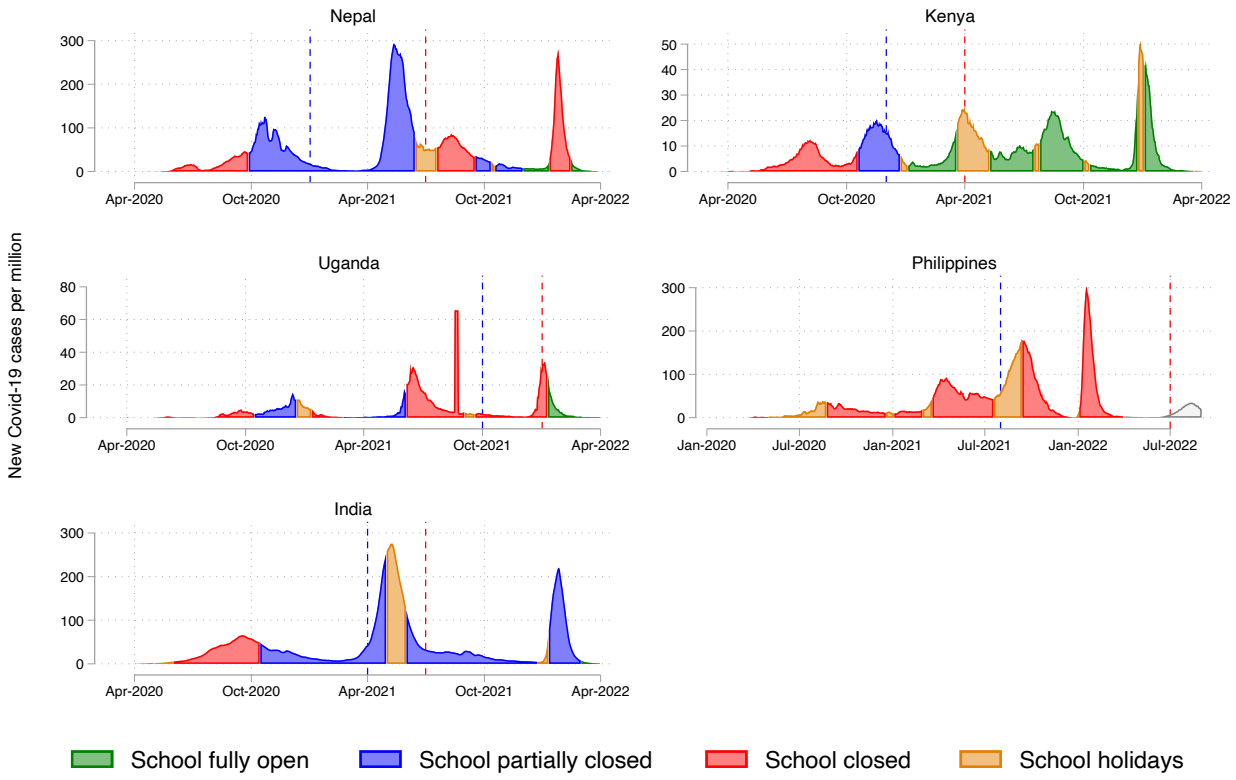
(c) Uganda



(d) Philippines

(e) India

Figure A6: COVID-19 Infections and School Disruption



Notes: This figure shows the evolution of the COVID-19 cases (in terms of new cases per million people) and resulting school closures or disruptions from April 2020 to April 2022. Each panel represents a country in our 5-country study, with the blue line indicating the approximate baseline collection date and the red line indicating the approximate endline collection date. Data sources include the Our World in Data COVID-19 dataset and UNESCO Global monitoring dataset of school closures

C Survey Tools

Building Resilient Education Systems: Baseline Questionnaire

1. **Did you speak to the household?**
   - ○ Yes
   - ○ No
2. **Did the child's caregiver provide consent to participate?**
   - ○ Yes
   - ○ No
3. **If no consent was given, why?**
4. **What is the student's name and surname?** _____
5. **In what school is the student enrolled?** _____
6. **In what district/region is the school located?** _____
7. **What class/grade is the student currently enrolled in?** _____
8. **What is the student's age?** *[Note:This question was asked in India, Uganda and the Philippines.]*
9. **What is the student's gender?** _____

MATH LEARNING MODULE: PROTOCOL

- Inform the caregiver that you would like the children to work on math problems.
- Ask the caregiver to place the call on speakerphone. They may repeat questions for the children to answer.
- Request that children answer math problems on their own on a scrap paper, including their answer. Children should work alone and not copy off anyone or work together.
- Explain that this is not an exam/test, so it's okay if the child(ren) do not get the answers correct.
- Children should take no longer than (30 seconds for Place Value) or (2 minutes for regular operation) to answer the question. If it seems someone is helping, gently ask them to refrain. If someone continues to help and/or the child takes longer than (30 seconds for place value) or (2 minutes for operations), mark that the child got this wrong.
- When finished, request that the child read out each of their answers and explain their answer to you. This is not an exam/test, so it's okay if your child(ren) does not get the answers correct. Are the child(ren) ready?
- Record all correct responses from each child and make a note of the highest operation that each child can perform.

- All students begin by answering the PLACE VALUE question and proceed with the LEARNING MODULE regardless of whether they answer with a correct response.

10. Prativa has 32 apples and organizes them by PLACE VALUE. How many TENS does she have?

- ○ The child got the question correct
- ○ The child got the question incorrect
- ○ The child gives the correct answer but is not able to convincingly explain how they got their answer/I don't believe they answered it themselves.
- ○ The parent was answering for the child/not letting the child answer, or the child used a calculator

11. The student solves: 34 + 47

- ○ The child got the question correct
- ○ The child got the question incorrect
- ○ The child gives the correct answer but is not able to convincingly explain how they got their answer/I don't believe they answered it themselves.
- ○ The parent was answering for the child/not letting the child answer, or the child used a calculator

12. The st

- ○ The
- ○ The
- ○ The thei
- ○ The calc

13. The student solves: 23*4 (28 multiplied by 3)

- ○ The child got the question correct
- ○ The child got the question incorrect
- ○ The child gives the correct answer but is not able to convincingly explain how they got their answer/I don't believe they answered it themselves.
- ○ The parent was answering for the child/not letting the child answer, or the child used a calculator

14. The student solves: 80/9 (80 divided by 9)

- ○ The child got the question correct
- ○ The child got the question incorrect
- ○ The child gives the correct answer but is not able to convincingly explain how they got their answer/I don't believe they answered it themselves.
- ○ The parent was answering for the child/not letting the child answer, or the child used a calculator

15. What languages is the learner able to speak? _____

16. **What is the main caregiver's highest level of education?** ***[Note: This question was asked at baseline in Nepal, and Philippines and endline in other countries. Answer choices were adjusted to the contexts where necessary.]***
    - ○ None/ Informal Education
    - ○ Primary Education
    - ○ High School
    - ○ Certificate/ Vocational
    - ○ College/ Bachelor's Degree
    - ○ Master's Degree or higher
    - ○ Respondent does not know
17. **What is the best phone number to reach the student?** _____
18. **Who is the owner of this phone (first name and surname)?** _____
19. **What is the relationship of the phone owner to the student?** _____
20. **Please list one alternate phone number where the child can be reached.** _____

Building Resilient Education Systems: Endline Questionnaire

1. **Did you speak to the person that typically helps the child with math problems and schoolwork?**
   - ○ Yes
   - ○ No
2. **Did they provide consent to participate in the endline survey?**
   - ○ Yes
   - ○ No
3. **Who typically provides educational instruction or support to the child outside of school?**
   - ○ Mother
   - ○ Father
   - ○ Grandparent
   - ○ Sibling
   - ○ Others not mentioned above in the household
   - ○ Teacher
   - ○ Adults from the community who do not live in the household
4. **What is the main caregiver's highest level of education?**
   - ○ None/ Informal Education
   - ○ Primary Education
   - ○ High School
   - ○ Certificate/ Vocational
   - ○ College/ Bachelor's Degree
   - ○ Master's Degree or higher
   - ○ Respondent does not know
5. **How confident do you feel that [student name] made progress in learning over the past three months?**
   - ○ The child would have significant difficulty performing any operation
   - ○ Not confident at all
   - ○ Slightly confident
   - ○ Moderately confident
   - ○ Very confident

6. **What is the highest operation that you think the child can easily perform?**
   - ○ The child would have significant difficulty performing any operation
   - ○ Addition
   - ○ Subtraction
   - ○ Multiplication
   - ○ Division
   - ○ Respondent does not know
7. **Would you be interested in receiving phone-based education support in the future?**
   - ○ Yes, through both phone calls and SMS
   - ○ Yes, only through SMS
   - ○ Yes, only through phone calls
   - ○ Yes, through either phone calls or SMS
   - ○ No, they are not interested
8. **Over the past four weeks, how often did you spend time doing educational activities (in general) with [student name]? For example, reading, practicing math problems, composition, etc.**
   - ○ Never
   - ○ Less than once per week
   - ○ 1-2 times per week
   - ○ 3-4 times per week
   - ○ 5 or more times per week
9. **Over the past four weeks, how often did you spend time doing educational activities (in general) with children in your household OTHER than [student name]? For example, reading, practicing math problems, composition, etc.**
   - ○ Never
   - ○ Less than once per week
   - ○ 1-2 times per week
   - ○ 3-4 times per week
   - ○ 5 or more times per week

10. Is the child currently in school?

- ○ Yes - in person has been in the same school for the past 3 months
- ○ Yes - in person switched to a new school in the past 3 months
- ○ No - not in school currently because school is closed but will return when school is open
- ○ No - not in school currently because school is closed and unlikely to return when schools reopen
- ○ No - not in school currently. School is open but the student has not returned.
- ○ F. Respondent refused to answer *[Note: These answer options were slightly adjusted to suit the context if necessary.]*

11. Were SMS messages that had math problems sent to you and [student name] over the past [context-specific time frame]?

- ○ Yes
- ○ No
- ○ Respondent does not know

12. Over the past [context-specific time frame], how often did the child practice math problems that were sent to you in an SMS message?

- ○ Always
- ○ Frequently
- ○ Sometimes
- ○ Rarely
- ○ Never
- ○ The respondent does not know

13. How many phone calls for math instruction did the child receive over [context-specific time frame]?

Integer

14. Is the person that typically helps the child with their math problems / school-work currently working?

- ○ Yes, full-time
- ○ Yes, part-time
- ○ No, they have retired
- ○ No, they are unemployed

15. How much does [student name] enjoy school?

- ○ Very much
- ○ Somewhat
- ○ Not much
- ○ They do not like school at all

16. Which maths problems does the child say they can easily perform?

- ○ The child would have significant difficulty performing any operation
- ○ Addition
- ○ Subtraction
- ○ Multiplication
- ○ Division
- ○ Respondent does not know

17. How true is the following statement about the child? - The child has many worries, or often seems worried.

- ○ Not true
- ○ Somewhat true
- ○ Definitely true

18. Hypothetically, how much would you be willing to pay for your child to receive weekly 1-on-1 math phone calls to support their learning?

- ○ Nothing
- ○ A little
- ○ A lot

CORE LEARNING MODULE PROTOCOL

- Inform the caregiver that you would like the children to work on math problems.
- Ask the caregiver to place the call on speakerphone. They may repeat questions for the children to answer.
- Request that children answer math problems on their own on a scrap paper, including their answer. Children should work alone and not copy off anyone or work together.
- Explain that this is not an exam/test, so it's okay if the child(ren) do not get the answers correct.
- Children should take no longer than (30 seconds for Place Value) or (2 minutes for regular operation) to answer the question. If it seems someone is helping, gently ask them to refrain. If someone continues to help and/or the child takes longer than (30 seconds for place value) or (2 minutes for operations), mark that the child got this wrong.

- When finished, request that the child read out each of their answers and explain their answer to you. This is not an exam/test, so it's okay if your child(ren) does not get the answers correct. Are the child(ren) ready?
- Record all correct responses from each child and make a note of the highest operation that each child can perform.
- All students begin by answering the PLACE VALUE question and proceed with the LEARNING MODULE regardless of whether they answer with a correct response.

19. Prativa has 47 apples and organizes them by PLACE VALUE. How many TENS does she have?

- ○ The child got the question correct
- ○ The child got the question incorrect
- ○ The child gives the correct answer but is not able to convincingly explain how they got their answer/ I don't believe they answered it themselves.
- ○ The parent was answering for the child/not letting the child answer, or child used a calculator

20. The student solves: 52 + 39

- ○ The child got the question correct
- ○ The child got the question incorrect
- ○ The child gives the correct answer but is not able to convincingly explain how they got their answer/ I don't believe they answered it themselves.
- ○ The parent was answering for the child/not letting the child answer, or child used a calculator

21. The student solves: 42 - 29

- ○ The child got the question correct
- ○ The child got the question incorrect
- ○ The child gives the correct answer but is not able to convincingly explain how they got their answer/ I don't believe they answered it themselves.
- ○ The parent was answering for the child/not letting the child answer, or child used a calculator

22. The student solves: 28*3 (28 multiplied by 3)

- ○ The child got the question correct
- ○ The child got the question incorrect
- ○ The child gives the correct answer but is not able to convincingly explain how they got their answer/ I don't believe they answered it themselves.
- ○ The parent was answering for the child/not letting the child answer, or child used a calculator

23. The student solves: 65/8 (65 divided by 8)

- ○ The child got the question correct
- ○ The child got the question incorrect
- ○ The child gives the correct answer but is not able to convincingly explain how they got their answer/ I don't believe they answered it themselves.
- ○ The parent was answering for the child/not letting the child answer, or child used a calculator

24. A man drives 24km. Then he drives 17km. How many km did he drive in total?

- ○ The child got the question correct
- ○ The child got the question incorrect
- ○ The child gives the correct answer but is not able to convincingly explain how they got their answer/ I don't believe they answered it themselves.
- ○ The parent was answering for the child/not letting the child answer, or child used a calculator

25. I will now ask you a word riddle. If the day before yesterday was Tuesday, what day is it?

- ○ The student got the answer correct
- ○ The student did not get the answer correct within 45 seconds.

26. Would you like to try to answer another logic question?

- ○ Yes, I'd like to try with a question as difficult as this one
- ○ Yes, but I'd like to try an easier question
- ○ No

27. The student solves: 3/8 + 4/8 (three eighths plus four eighths)

- ○ The child got the question correct
- ○ The child got the question incorrect
- ○ The child gives the correct answer but is not able to convincingly explain how they got their answer/ I don't believe they answered it themselves.
- ○ The parent was answering for the child/not letting the child answer, or child used a calculator

D Additional Training and Implementation Details

Trainings were all conducted largely virtually. First, a training of trainers was conducted. Program coordinators and senior project staff from partner organisations in each country (government, NGOs, and the World Bank) took part in a country-specific practical, interactive 1-day Training of Trainers workshop. This was delivered online by staff from Youth Impact. This was followed by a 1-2 day training of teachers and tutors led by each country's respective implementing organisations. These training sessions were contextualized based on country context, with training delivered in the local language and with contextually relevant insights and best practices. Implementer training was delivered online in all countries except in Kenya and Uganda, where training was in-person.

During implementation, Youth Impact's master trainer and program coordinator had periodic calls with each implementing organization to help troubleshoot and share tips and best practices learned from implementation in other countries. For example, Youth Impact reviewed data with implementing partners, shared tips about the best times to schedule sessions with households throughout the week, and shared suggestions from how other countries had paced their lessons and targeted instruction successfully.

The government-led interventions were led fully by the government. The government assigned a point person within the ministry to lead the intervention and coordinate within relevant government structures and to support teachers. Government leads met with technical support partners (NGOs and data collection firms) several times a month, and had regular phone and online communication with teachers. To the extent that there was NGO involvement, it was in conducting the training of trainers and sharing of draft tools and phone call guidelines to ensure common content and approaches were delivered across treatments.

On SMS content, the weekly messages included a set of math problems that students were encouraged to solve between calls. Each week, there was a problem provided for each level: i.e. an addition problem, a subtraction problem, a multiplication problem, and a division problem. A typical SMS message looked as follows: "Welcome to Week 2! ADDITION: 14+46=?; 18+33=?; SUBTRACTION: He picked 32 apples and gave his friend 11. How many is he left with?; MULTIPLICATION: 23x3=? 14x2=?; DIVISION: You need to divide 10 apples evenly between 2 friends. How many will each get?"

E Education in Emergencies Database References

Various sources were used to build the database for school disruptions from 2000 to 2020. This data is not meant to cover the universe of disruptions but rather to provide an illustrative snapshot of the degree to which school disruptions happen to start to capture their frequency. Our search included news articles, as well as reports from organizations such as Save the Children, UNICEF's Children's Climate Risk Index, ReliefWeb by the United Nations Office for the Coordination of Humanitarian Affairs (OCHA), and the Global Assessment Report on Disaster Risk Reduction. Below we list specific news sources documenting school disruptions.

"27 Schools Closed Due to Increased Flooding on Malaysia's Eastern Coast." TODAY,www.todayonline.com/world/asia/27-schools-closed-due-increased-flooding-malaysias-eastern-coast. Accessed 7 Nov. 2022.

"2019 Philippines Mindanao Island Earthquakes: Facts, FAQs, and How to Help." Save the Children, www.savethechildren.org/us/what-we-do/emergency-response/philippines-earthquake-mindanao-island-2019-facts. Accessed 7 Nov. 2022.

"Amid Acute Water Crisis, Schools in Shimla Shut for 5 Days." NDTV.com, 2 June 2018, www.ndtv.com/india-news/schools-in-shimla-closed-for-5-days-starting-monday-due-to-water-crisis-says-himachal-government-rep-1861557.

"Bangladesh: Over 4,000 Primary Schools Closed by Floods - Bangladesh." ReliefWeb, 21 Aug. 2007, reliefweb.int/report/bangladesh/bangladesh-over-4000-primary-schools-closed-floods.

"Bengal Schools to Close for 11 Days Over Sudden Heat Wave - Times of India." The Times of India, timesofindia.indiatimes.com/education/news/bengal-schools-to-close-for-11-days-over-sudden-heat-wave/articleshow/64639741.cms. Accessed 7 Nov. 2022.

"Bolivia Schools Close Early as Drought Empties Reservoirs." BBC News, www.bbc.com/news/world-latin-america-38073575. Accessed 7 Nov. 2022.

"Chennai Water Crisis: School Closes Down for Junior Classes, Others Declare Half-day." The News Minute, 19 June 2019, www.thenewsminute.com/article/chennai-water-crisis-school-closes-down-junior-classes-others-declare-half-day-103919.

"Children During Long Winter Vacation in Central Highlands." Children During Long Winter Vacation in Central Highlands — UNICEF Afghanistan, 17 Feb. 2019, www.unicef.org/afghanistan/stories/children-during-long-winter-vacation-central-highlands.

"China - Hong Kong: Flu Triggers Two-week School Closure - China - Hong Kong (Special Administrative Region)." ReliefWeb, 11 June 2009, reliefweb.int/report/china-hong-kong-special-administrative-region/china-hong-kong-flu-triggers-two-week-school.

"Dangerous Air Pollution in India Forces Delhi Schools to Close for 2nd Time in 2 Weeks." India Air Pollution in Delhi Spikes Again Today Forcing School and Industry Closures and Sending Residents to Hospital - CBS News, 15 Nov. 2019, www.cbsnews.com/news/air-pollution-in-india-delhi-forces-schools-industry-closed-health-problems-today-2019-11-15.

Daniele, Ushar. "Air Pollution in Malaysia Forces 400 School Closures, Sickens More Than 100 Chil-

dren — CNN." CNN, 27 June 2019, www.cnn.com/2019/06/26/health/malaysia-pollution-schools-intl-hnk/index.html.

Davison, Tamara. "Schools to Close Due to Water Shortages in Mexico City - Aztec Reports." Aztec Reports, 25 Oct. 2018, aztecreports.com/school-close-water-shortage/1801.

"Delhi Smog: Schools Closed for Three Days as Pollution Worsens." BBC News, www.bbc.com/news/world-asia-india-37887937. Accessed 7 Nov. 2022.

"Dhanusa Community Schools to Shut Down From Today." Dhanusa Community Schools to Shut Down From Today, 7 Nov. 2022, kathmandupost.com/national/2018/01/03/dhanusa-community-schools-to-shut-down-from-today.

"Dry Pipes - Water Crisis Forces Schools to Close." Dry Pipes - Water Crisis Forces Schools to Close — Lead Stories — Jamaica Gleaner, 11 Apr. 2014, jamaica-gleaner.com/gleaner/20140411/lead/lead1.html.

"Ebola Outbreak: Nigeria Closes All Schools Until October." BBC News, www.bbc.com/news/world-africa-28950347. Accessed 7 Nov. 2022.

"EC Asks Govt to Shut Schools on May 8-14." EC Asks Govt to Shut Schools on May 8-14, 7 Nov. 2022, kathmandupost.com/national/2017/05/04/ec-asks-govt-to-shut-schools-on-may-8-14.

"Flooding Forces School Closures in India's Hyderabad." Flooding Forces School Closures in India's Hyderabad — Climate Crisis — Al Jazeera, 24 Sept. 201624, www.aljazeera.com/gallery/2016/9/24/flooding-forces-school-closures-in-indias-hyderabad.

"Government Decides to Shut Schools for Four Days as Air Pollution Reaches Hazardous Levels." Government Decides to Shut Schools for Four Days as Air Pollution Reaches Hazardous Levels, 7 Nov. 2022, kathmandupost.com/national/2021/03/29/government-decides-to-shut-schools-for-four-days-as-air-pollution-reaches-hazardous-levels.

"Hand, Foot and Mouth Disease Causes 18 Bangkok School Closures – Tasty Thailand." Hand, Foot and Mouth Disease Causes 18 Bangkok School Closures – Tasty Thailand, 17 July 2012, tastythailand.com/hand-foot-and-mouth-disease-causes-18-bangkok-school-closures.

"Hand, Foot and Mouth Disease Outbreak in Malaysia: 6 Things You Need to Know About the Disease." Hand, Foot and Mouth Disease Outbreak in Malaysia: 6 Things You Need to Know About the Disease — the Straits Times, 31 July 2018, www.straitstimes.com/singapore/health/hand-foot-and-mouth-disease-outbreak-in-malaysia-6-things-you-need-to-know-about.

"Heatwave Shuts More Than 250 Malaysian Schools: Reports." Heatwave Shuts More Than 250 Malaysian Schools: Reports, phys.org/news/2016-04-heatwave-malaysian-schools.html. Accessed 7 Nov. 2022.

"Mers: South Korea Closes 700 Schools After Third Death." BBC News, www.bbc.com/news/world-asia-33002795. Accessed 7 Nov. 2022.

"Monsoon Rains Bring Severe Flooding and Landslides Across South Asia, Affecting More Than Five Million Children." Monsoon Rains Bring Severe Flooding and Landslides Across South Asia, Affecting More Than Five Million Children, 3 Nov. 2022, www.unicef.org/rosa/press-releases/monsoon-rains-bring-severe-flooding-and-landslides-across-south-asia-affecting-more.

"More Than 10,000 Schools in Sichuan Badly Damaged." More Than 10,000 Schools in Sichuan Badly Damaged, 18 May 2008, www.unicef.cn/en/press-releases/more-10000-schools-sichuan-badly-damaged.

"Pakistan: Flood Damaged Schools Lead to Education Worries - Pakistan." ReliefWeb, 26 Aug. 2010, reliefweb.int/report/pakistan/pakistan-flood-damaged-schools-lead-education-worries.

"Philippines: Mayon Volcano - Jan 2018." ReliefWeb, 3 July 2020, reliefweb.int/disaster/vo-2018-000005-phl.

Post, The Jakarta. "Moving the Capital Is an Urgent National Security Matter." The Jakarta Post, 5 Mar. 2022, www.thejakartapost.com/opinion/2022/03/04/moving-the-capital-is-an-urgent-national-security-matter.html.

Post, The Jakarta. "Pontianak Urges School Closure for Smog." The Jakarta Post, 20 Aug. 2018, www.thejakartapost.com/news/2018/08/20/pontianak-urges-school-closure-for-smog.html.

"Prolonged Drought in East Africa Forces Millions of Children Out of School." Prolonged Drought in East Africa Forces Millions of Children Out of School — Blog — Global Partnership for Education, 3 Apr. 2018, www.globalpartnership.org/blog/prolonged-drought-east-africa-forces-millions-children-out-school.

Public Company Limited, Bangkok Post. "Cambodia Shuts Schools Amid Disease Fears." www.bangkokpost.com, www.bangkokpost.com/world/303151/cambodia-shuts-schools-amid-disease-fears. Accessed 7 Nov. 2022.

"Schools Closed by Ebola Reopen in Guinea and Liberia." VOA, 12 Feb. 2015, learningenglish.voanews.com/a/schools-reopen-guinea-liberia-closed-ebola/2635017.html.

"Schools in Fiji Will Stay Closed Until Cleared." FijiTimes, 17 Jan. 2022, www.fijitimes.com/schools-in-fiji-will-stay-closed-until-cleared.

Singini, George. "Water Crisis Forces Closure of Euthini Secondary School - the Nation Online." The Nation Online, 10 Dec. 2015, mwnation.com/water-crisis-forces-closure-of-euthini-secondary-school.

Staff, Reuters. "Smog-ridden Mexico City Suspends School Classes Due to Pollution." U.S., www.reuters.com/ article/us-mexico-pollution-idUSKCN1SL2T1. Accessed 7 Nov. 2022.

Staff, Reuters. "Thai Capital to Close 435 Schools to Halt H1N1 Spread." U.S., www.reuters.com/ article/idUSBKK452436. Accessed 7 Nov. 2022.

Staff, World Vision. "Impact of Ebola on Education in Sierra Leone." Impact of Ebola on Education in Sierra Leone — World Vision, 24 Apr. 2015, www.worldvision.org/health-news-stories/impact-of-ebola-on-education-sierra-leone.

"Thousands of Schools Destroyed, Damaged or Disrupted by South Asia's Deadly Floods." Theirworld, 15 Aug. 2022, theirworld.org/news/south-asia-floods-destroy-damage-thousands-schools-india-bangladesh-nepal.

"Three Million Children Affected by Mindanao Earthquakes: Save the Children Deploys Rapid Response Team to Meet the Needs of Communities - Philippines." ReliefWeb, 1 Nov. 2019, reliefweb.int/report/philippines/three-million-children-affected-mindanao-earthquakes-save-

children-deploys-rapid.

"Unraveling the Water Crisis in Venezuela." Unraveling the Water Crisis in Venezuela — Center for Strategic and International Studies, 27 May 2021, www.csis.org/analysis/unraveling-water-crisis-venezuela.

"Vanuatu: Tackling the Impact of Natural Disasters by Building a Resilient Education System — Global Partnership for Education." Vanuatu: Tackling the Impact of Natural Disasters by Building a Resilient Education System — Stories of Change — Global Partnership for Education, www.globalpartnership.org/results/stories-of-change/vanuatu-tackling-impact-natural-disasters-building-resilient-education. Accessed 7 Nov. 2022.

"Water Crisis Forces Lyantonde Schools to Close Prematurely:: Uganda Radionetwork." Uganda Radionetwork, ugandaradionetwork.net/story/water-crisis-forces-lyantonde-schools-to-close-prematurely. Accessed 7 Nov. 2022.

"Zimbabwe: Water Shortages Force Schools to Close - Zimbabwe." ReliefWeb, 10 Sept. 2004, reliefweb.int/report/zimbabwe/zimbabwe-water-shortages-force-schools-close.