Experimental evidence on learning using low-tech when school is out

Noam Angrist [1,2,3] ✉, Peter Bergman[4,5] and Moitshepi Matsheng[1,6]

School closures occurred extensively during the COVID-19 pandemic, and occur in other settings, such as teacher strikes and natural disasters. The cost of school closures has proven to be substantial, particularly for households of lower socioeconomic status, but little evidence exists on how to mitigate these learning losses. This paper provides experimental evidence on strategies to support learning when schools close. We conduct a large-scale randomized trial testing two low-technology interventions—SMS messages and phone calls—with parents to support their child in Botswana. The combined treatment improves learning by 0.12 standard deviations, which translates to 0.89 standard deviations of learning per US$100, ranking among the most cost-effective interventions to improve learning. We develop remote assessment innovations, which show robust learning outcomes. Our findings have immediate policy relevance and long-run implications for the role of technology and parents to support education provision during school disruptions.

The COVID-19 pandemic paralyzed education systems worldwide; at one point, school closures forced over 1.6 billion learners out of classrooms[1]. Moreover, widespread school closures are not unique to COVID-19: teacher strikes, summer breaks, earthquakes, viruses (such as influenza and Ebola) and weather-related events cause schools to close. School closures result in large learning losses, which have been documented in North America, Western Europe, Asia and Sub-Saharan Africa[2–5]. These learning losses include a combination of knowledge that is forgotten over time, and forgone learning that would have occurred if schools were open. To mitigate learning loss in the absence of school, high-income families have access to alternative sources of instruction—books, computers, internet, radio, television and smart phones—that many low-income families do not[6–9].

Stemming learning loss when schools are closed, particularly in areas where learning resources are lacking in the household, requires outside-school interventions that can substitute instead of complement ongoing instruction. Doing so at scale requires cheap low-technology solutions that can reach as many families as possible. One such low-technology with high reach is mobile phones. In low- and middle-income countries, 70% to 90% of households own at least one mobile phone, while only 15% to 60% of households have internet access[10].

In this paper, we provide experimental estimates of minimizing the impact of school closures on learning in the context of the COVID-19 pandemic. Specifically, we focus on basic numeracy interventions and assessment. We evaluate two low-tech mobile phone solutions that leverage short message service (SMS) text messages and direct phone calls to support parents in educating their children. A sample of 4,500 families with primary-school-aged children across nearly all regions of Botswana were randomly assigned to either intervention arms or a control arm. The sample had characteristics that match national averages along several indicators. In one treatment arm, SMS messages provided a few basic numeracy 'problems of the week.' A second treatment arm supplemented these weekly SMS messages with a live 15–20 min phone call

walk-through of the problems. Each student in this arm received 3 h of direct instruction spread over 8 weeks.

In Botswana, the government closed schools for a planned 6 months starting 20 March 2020. Schools reopened on 17 June, were subsequently closed again after a new wave of COVID-19 cases and have since reopened. Similar waxing and waning of school closures have occurred throughout the year. Even as students returned to school (our data show that 98% of primary school students eventually returned after initial school closures), a double-shift rotation system, where half of the students attend school in the morning and the other half attend in the afternoon, drastically reduced time in school for each student. While the government launched learning programmes on national television and radio stations during this time, our data show that access to radio is relatively low, with only 20% of the control group listening to radio in the status quo.

The lack of access to education during school closures is likely to exacerbate a pre-existing learning crisis in Botswana. Analysis of data from the Southern African Consortium for Monitoring Education Quality (SACMEQ) found that 88% of grade 6 students are at grade 4 math levels or below. A census of two regions further found that only 10% of students in grade 5 could do 2-digit division and 40% could not read a simple 1-paragraph story[11].

In this context of a pressing need to improve learning and limited options to receive an education during school closures, low-tech solutions to reliably provide remote education were in high demand. Over 99% of parents reported demand for continued remote learning services even if schools reopened, probably due to uncertainty around whether schools would remain open, reduced school hours and disrupted learning.

Results of the main intervention in the trial show large statistically significant learning differences between treatment and control groups. The combined phone and SMS intervention increased learning by 0.121 standard deviations (95% CI 0.031, 0.210; $P = 0.008$), although the SMS intervention alone had no statistically significant effect on learning ($\beta = 0.024$, 95% CI −0.066, 0.114; $P = 0.602$). The improvement in learning between the combined

[1]Youth Impact, Gaborone, Botswana. [2]Department of Economics and Blavatnik School of Government, University of Oxford, Oxford, UK. [3]The World Bank, Washington, DC, USA. [4]Teachers College, Columbia University, NY, USA. [5]National Bureau of Economic Research (NBER), Cambridge, MA, USA. [6]Botswana National Youth Council, Gaborone, Botswana. ✉e-mail: nangrist@youth-impact.org

Table 1 | Treatment effects on learning outcomes

	(1)	(2)	(3)
	Avg level	Place value	Fractions
Panel A			
SMS only	0.024	0.009	0.047
	(0.046)	(0.046)	(0.046)
	[0.602]	[0.837]	[0.309]
	{−0.066,0.114}	{−0.080,0.099}	{−0.043,0.137}
Phone + SMS	0.121	0.114	0.075
	(0.046)	(0.044)	(0.046)
	[0.008]	[0.009]	[0.100]
	{0.031,0.210}	{0.028,0.200}	{−0.014,0.165}
Panel B			
Not targeted	0.070	0.026	0.029
	(0.046)	(0.045)	(0.046)
	[0.130]	[0.572]	[0.527]
	{−0.021,0.160}	{−0.063,0.114}	{−0.061,0.120}
Targeted	0.076	0.098	0.093
	(0.046)	(0.044)	(0.045)
	[0.097]	[0.026]	[0.041]
	{−0.014,0.165}	{0.012,0.185}	{0.004,0.182}
Control mean	1.974	1.774	1.605
Strata fxed effects	Yes	Yes	Yes
Observations	2,815	2,881	2,751
P value, SMS = Phone	0.033	0.017	0.528
P value, Targeted = Not targeted	0.896	0.098	0.160

Panel A in this table reports results from a regression estimating treatment effects of the SMS only, and phone and SMS intervention on learning outcomes in terms of standard deviations. Panel B reports regression estimates of the effect of targeted instruction on learning outcomes, also in terms of standard deviations. Average level 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). Place value refers to a distinct place value question. Fractions refers to a distinct question to solve a higher-order fractions problem. Each panel reports separate models that pool treatment groups by category. Standard errors are in parentheses, P values are in square brackets and 95% confidence intervals are in curly brackets. All statistical tests are two-tailed.

phone and SMS group and SMS only group equates to 0.097 standard deviations ($P = 0.033$). We find a 31% reduction in absolute innumeracy (students who cannot do any numerical operations) and an average level gain on the Annual Status of Education Report (ASER) assessment of 0.15 levels (95% CI 0.039, 0.262; $P = 0.008$). For households that participated in all sessions, instrumental variables analysis shows learning gains are 0.167 standard deviations (95% CI 0.046, 0.289; $P = 0.007$). The phone plus SMS intervention also translates to solving other foundational skill competencies, such as solving place values ($\beta = 0.114$, 95% CI 0.028, 0.200; $P = 0.009$). Lastly, our results are robust to a series of sensitivity tests, and explore how effects vary on the basis of whether instruction is targeted to the learner's learning level.

These results demonstrate that certain types of instruction through mobile phones can provide an effective and scalable method for education delivery outside of the traditional schooling system. The phone and SMS intervention is highly cost-effective, with 0.63 to 0.89 standard deviation learning gains per US$100. These results also reveal that some level of direct instruction, which can be done cheaply and virtually via phone, might be necessary;

SMS messages alone are not as effective as phone calls plus SMS messages at producing learning gains ($P = 0.033$). This is in line with existing evidence showing that SMS messages might best serve as a complement to direct instruction, as in this study, or as an accountability nudge in education systems, such as by helping families track their child's academic progress[12].

We further find that parental engagement with the interventions is high: 92% of parents report that their child attempted to solve the problems sent, with slightly higher engagement in the SMS plus phone group of 95%. Parents report 8.7% and 15.2% greater self-efficacy in supporting their child's learning because of the SMS only and phone and SMS interventions, respectively. Parents also update their beliefs about their child's learning level in tandem with their child's learning progress. This suggests that parents are involved and aware of their child's academic progress. We also find no statistically significant effects on parents' return to work post lockdown due to the interventions, which alleviates the concern that further parental engagement in their child's education might crowd out other activities, such as returning to work.

Remote instruction compelled several innovations in high-frequency, low-cost remote assessment. To measure learning, we adapted the numeric portion of a test that has consistently been used in the education literature—the ASER test—into a phone-based learning assessment[13–16]. We also incorporated time limits, and a requirement that children explain their work to accurately identify their numeracy levels. To measure the reliability of our assessment, we randomly assigned problems measuring the same proficiency to students, a version of a reliability test used in the psychometric literature[17]. We further disentangled cognitive skills gains from effort effects, which have been shown to affect test scores[18]. We tested this hypothesis with a real-effort task. We also measured whether learning gains are a matter of familiarity with the content in intervention groups that receive exposure to similar material. We tested this by including new content not covered during the intervention, but which is related, such as fractions. The familiarity hypothesis was also partially tested with randomized problems of the same proficiency. Lastly, we demonstrated the value of high-frequency, remote assessment by using a midline assessment to target content to learning levels for a cross-randomized subgroup of students.

Our work contributes to several literatures. The low-tech interventions we test relate to a growing literature on technology and education. Mobile phone SMS messages have been used to supplement adult education programmes in Niger and the USA[19–21], to help parents teach nascent literacy skills to their children in the USA[22–24], to enhance parental engagement in children's education in both Brazil[25] and Cote d'Ivoire[26], and to help parents monitor their child's effort and progress in school[27–36] (see ref. [36] for a review). We contribute to this literature by providing evidence on live, direct instruction through phone calls rather than only automated, text-message-based instruction. While mobile phone calls have been used as a medium to deliver health interventions[37], before the COVID-19 pandemic, experimental evaluation of phone calls to provide live instruction in education has been limited. Moreover, we contribute to the literature by testing low-tech approaches in a setting where these interventions operate largely as substitutes for schooling rather than as complements (see ref. [38] for a review of the role of technology as a complement or substitute for the traditional schooling system). We also contribute learning data collected via phone-based assessments.

This paper also relates to an emerging global priority to improve learning at low cost and at scale. Even before the pandemic shock to education, student learning levels were low, and progress was slow as highlighted by UNESCO and the World Bank. For example, in Kenya, Tanzania and Uganda, three-quarters of the students in grade 3 cannot read a basic sentence such as 'the name of the dog

Table 2 | Robustness check: random problem by learning proficiency

	(1)	(2)	(3)	(4)	(5)
	Addition	Subtraction	Multiplication	Division	Fractions
Random problem 2	−0.002	0.024	0.017	−0.039	0.017
	(0.020)	(0.024)	(0.028)	(0.025)	(0.026)
	[0.938]	[0.316]	[0.530]	[0.124]	[0.501]
	{−0.041,0.038}	{−0.023,0.071}	{−0.037,0.071}	{−0.088,0.011}	{−0.033,0.068}
Random problem 3	0.014	0.007	−0.004	−0.008	−0.023
	(0.021)	(0.024)	(0.028)	(0.026)	(0.027)
	[0.512]	[0.765]	[0.895]	[0.765]	[0.400]
	{−0.027,0.055}	{−0.040,0.054}	{−0.058,0.050}	{−0.059,0.043}	{−0.075,0.030}
Random problem 4	−0.011	0.036	−0.044	0.005	−0.008
	(0.020)	(0.024)	(0.027)	(0.026)	(0.026)
	[0.599]	[0.145]	[0.101]	[0.858]	[0.753]
	{−0.050,0.029}	{−0.012,0.083}	{−0.097,0.009}	{−0.047,0.056}	{−0.059,0.043}
Random problem 5	0.010	0.005	−0.011	0.002	−0.032
	(0.021)	(0.024)	(0.027)	(0.026)	(0.027)
	[0.631]	[0.849]	[0.681]	[0.951]	[0.228]
	{−0.031,0.051}	{−0.042,0.051}	{−0.065,0.042}	{−0.050,0.053}	{−0.084,0.020}
Observations	2,815	2,815	2,5	2,815	2,751
F-test: equivalence across all problems	0.715	0.458	0.139	0.307	0.498

This table reports results from a regression estimating differences in average proficiency across four randomly assigned problems relative to a base random problem for the following proficiencies: addition, subtraction, multiplication, division and fractions. For example, for a subtraction problem, a random fifth of students will receive the question '83–45', whereas another random fifth of students will receive the question '72–18' to test the subtraction with borrowing proficiency, and so on, across a total of five random problems for each proficiency. Standard errors are in parentheses, *P* values are in square brackets and 95% confidence intervals are in curly brackets. All statistical tests are two-tailed.

is Puppy'[39]. Moreover, a recent review of 150 impact evaluations in education in low- and middle-income countries found that nearly half had no effect on learning[40]. This trend of limited learning has been referred to as the 'learning crisis' by the international education community[41]. Some interventions that are effective, such as in-person tutoring programmes, can be expensive. For example, a tutoring programme which yielded 0.19 to 0.31 standard deviation learning gains cost US$2,500 per child[42]. The intervention in this trial, low-cost remote tutoring via phone calls, has similar effective sizes and is two orders of magnitude cheaper. To address learning shortfalls and gaps in education provision, which have been exacerbated by the COVID-19 pandemic[43], there is a need for approaches that cost-effectively improve learning on a global scale. In addition to the results presented in this study, other examples of cost-effective approaches that have emerged during COVID-19 include those found in refs. [44,45].

Our results have substantial implications for global policy. Recent estimates from the World Bank suggest that current school closures during the pandemic could cost up to US$10 trillion in net present value[46]. There is a pressing need to mitigate this fall-out on education worldwide. Even as schools start to reopen, this reopening is often partial—for example, with students receiving only half as many hours of instruction. Moreover, as stated earlier, school closures occur in settings beyond the current pandemic, such as in refugee settings and adverse weather events. In moments where a substitute for schooling is needed, particularly for families with fewer resources at home, the low-tech solutions tested in this study have unique potential to reach the masses. The results in this paper provide evidence that remote instruction by phone and SMS messages has the potential to improve learning for primary school children using a low-cost and scalable model when schooling is disrupted.

Results

The study took place in Botswana with 4,550 households. We compare our sample to national-level indicators and find that the final sample has characteristics that match those of a nationally representative sample as described in the sections below. Supplementary Fig. 1 shows a heat map of the location of the children's schools to demonstrate the distribution of participants across the country. Supplementary Fig. 2 provides a timeline of each step from initial phone number collection, piloting and training, programme implementation and waves of data collection. Supplementary Fig. 3 provides an overview of the experimental design. Of working phone numbers, 71% were reachable and gave consent to participate in the study.

We randomized the 4,550 phone numbers into three groups of equal size: a weekly SMS message followed by a phone call, a weekly SMS message only and a control group. We further randomly cross-randomized 2,250 numbers for a midline assessment, and approximately 1,600 of these were randomly selected to receive targeted instruction customized to their learning level using the data collected at midline. The initial randomization to SMS, phone calls and SMS, or the control group was stratified on whether at least one child in the household had previously participated in previous school-based educational programming, a proxy for having recently made substantial learning gains. Each phone number belongs to a caregiver and household.

Sample characteristics and representativeness. We include a few descriptive statistics to describe how our sample, which represents around 15% of all primary schools in Botswana, compares to characteristics of nationally representative samples. Botswana has nine regions in total and our sample covers eight of them, including the most remote and low-literacy regions.

Table 3 | Parent mechanisms: beliefs, self-efficacy and potential crowd-out

	(1) Parent reported child level	(2) Parent correct about child level	(3) Parent self-efficacy	(4) Parent perception that child learned	(5) Parent did not return to work	(6) Parent spent no time on education
Panel A						
SMS only	0.022	−0.012	0.049	0.066	−0.000	−0.001
	(0.045)	(0.023)	(0.021)	(0.022)	(0.018)	(0.009)
	[0.621]	[0.594]	[0.023]	[0.002]	[0.994]	[0.933]
	{−0.065,0.110}	{−0.058,0.033}	{0.007,0.091}	{0.024,0.109}	{−0.034,0.034}	{−0.019,0.017}
Phone + SMS	0.136	0.039	0.086	0.105	−0.029	−0.002
	(0.045)	(0.023)	(0.021)	(0.022)	(0.017)	(0.009)
	[0.002]	[0.099]	[0.000]	[0.000]	[0.092]	[0.809]
	{0.049,0.224}	{−0.007,0.085}	{0.044,0.129}	{0.062,0.148}	{−0.063,0.005}	{−0.020,0.016}
Panel B						
Not targeted	0.045	−0.001	0.050	0.071	−0.010	−0.003
	(0.045)	(0.023)	(0.022)	(0.022)	(0.018)	(0.009)
	[0.323]	[0.957]	[0.020]	[0.001]	[0.565]	[0.713]
	{−0.044,0.133}	{−0.047,0.045}	{0.008,0.093}	{0.028,0.114}	{−0.044,0.024}	{−0.021,0.015}
Targeted	0.111	0.028	0.084	0.099	−0.018	0.000
	(0.044)	(0.023)	(0.021)	(0.022)	(0.017)	(0.009)
	[0.012]	[0.239]	[0.000]	[0.000]	[0.296]	[0.973]
	{0.025,0.198}	{−0.018,0.073}	{0.042,0.125}	{0.056,0.142}	{−0.052,0.016}	{−0.018,0.019}
Control mean	2.228	0.398	0.566	0.492	0.190	0.045
Strata fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2,957	2,650	3,127	3,127	2,990	2,984
P value, SMS = Phone	0.009	0.029	0.071	0.075	0.088	0.872
P value, Targeted = not targeted	0.128	0.217	0.115	0.194	0.640	0.685

Panel A in this table reports results from a regression estimating treatment effects of the SMS only, and phone and SMS intervention on several parental mechanisms. Panel B reports regression estimates of the effect of targeted instruction on these parental mechanisms. Specifically, the table includes effects on parent accuracy of their child's learning level, their self-efficacy to support their child's learning and their belief that their child made progress in learning in general. This table also shows treatment effects on parent labour market outcomes in the form of returning to work post lockdown across treatment groups. Options for return to work included: returned to work full-time, returned to work part-time, retired or unemployed. Standard errors are in parentheses, P values are in square brackets and 95% confidence intervals are in curly brackets. All statistical tests are two-tailed.

Extended Data Fig. 1 compares study sample characteristics with national indicators for a subset of indicators. We find a similar gender split of between 50% and 52% in our sample and nationwide. We also find a similar ratio of rural students in our sample to the national average of 29%. We find similar distributions of learning: the percentage of students who score an A, B and C is 16%, 21% and 41% in study schools, respectively, and 14%, 17% and 36% for all primary schools in the nation.

In addition, we collect simple descriptive data on child age, grade and gender in surveys. Around 50% of our sample is female; the average age of students is 9.7; 28.5% of students are in grade 3, 39.1% in grade 4 and 32.4% in grade 5. The average age of caregivers participating in the randomized trial was 35, and 68% of them were female. Our data show that in the control group the median caregiver (48.5%) spends just 1–2 h on educational activities with their child per week. We asked households to nominate the best person to provide educational support to their child during school disruption: 81% of nominated caregivers were parents, 7.6% were grandparents, 7.8% aunts or uncles, and 2.8% siblings. Additional details on the primary caregiver are in Supplementary Information 'Section A: Sample Characteristics'.

For a subsample of parents (*n* = 209), we also measure parental education level and additional characteristics. This subset is not necessarily representative of the entire sample. However, they were the most responsive parents, suggesting that they probably represent an upper bound of the most literate parents. In the sample, 29% had completed schooling beyond secondary school, compared with a national average of 26% based on data from the World Bank. These measures suggest that the sample of parents have similar education levels to the national average. Moreover, the sample in the study has moderate literacy rates similar to other low- and middle-income countries. While the average secondary schooling completion rate in Europe and Central Asia is over 90%, average completion rates in lower middle-income countries are only just above 70%[47].

Primary outcomes. For our two main learning outcomes focused on foundational numeracy skills—average level and place value—Fig. 1 (see also Table 1) shows large, statistically significant learning differences between treatment and control groups. For the combined phone and SMS group, there was a 0.121 standard deviation (95% CI 0.031, 0.210; *P* = 0.008) increase in the average numerical operation. The learning gains for the combined phone and SMS intervention also translate to other foundational skill competencies, such as gains in place value of 0.114 standard deviations (95% CI 0.028, 0.200; *P* = 0.009). For households that participated in all sessions, instrumental variables analysis in Extended Data Fig. 2
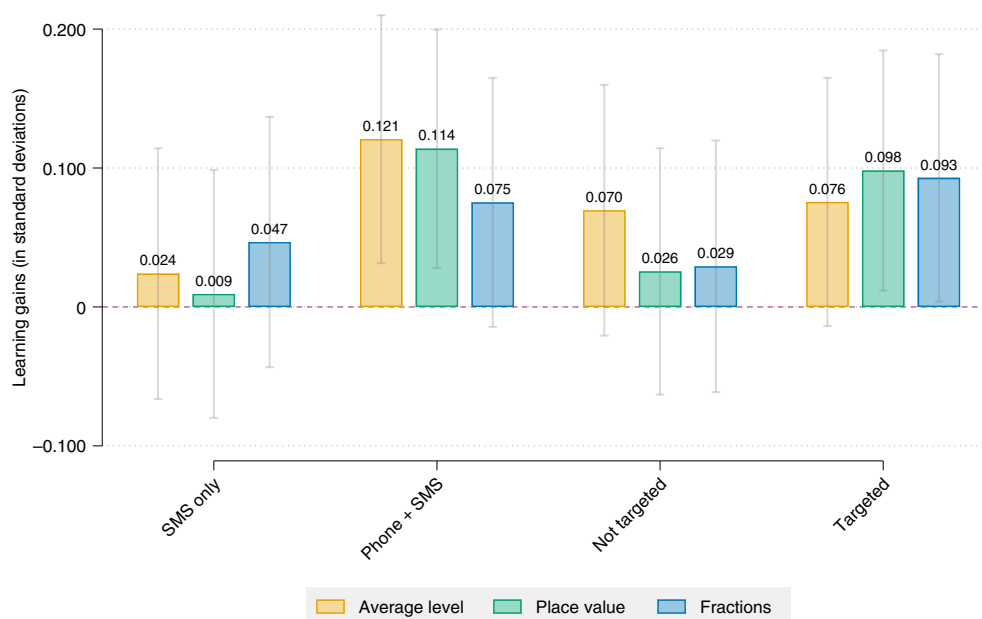
Fig. 1 | Treatment effects on learning outcomes. Data are presented as treatment effects relative to the control group ±95% confidence intervals. Full statistical results for the treatment effects are presented in Table 1. Effects are expressed in terms of standard deviations for comparable units. Each colour bar represents a distinct learning question. 'Average level' reports skill on the ASER 0 to 4 scale corresponding to no operations, addition, subtraction, multiplication and division ($n = 2,815$). 'Place value' refers to a distinct place value problem ($n = 2,881$) and 'Fractions' refers to a distinct question asking students to solve a fractions problem ($n = 2,751$). Each group—'SMS only', 'Phone + SMS', 'Not targeted' and 'Targeted'—refer to randomized treatment groups pooled across the designated category.

shows learning gains of 0.167 standard deviations (95% CI 0.046, 0.289; $P = 0.007$). As we show later, these results are robust to several validity checks. We find no statistically significant effects on average for the SMS-only intervention across all three learning proficiencies—average level, place value and fractions ($P = 0.602$, 0.837 and 0.309, respectively).

These results reveal that combined phone and SMS low-tech interventions can generate substantial learning gains, and that SMS messages alone are not as effective ($P = 0.033$). This suggests that SMS messages might not be as effective as direct instruction on their own; instead, they might be best placed as a complement to direct instruction through phone calls as in this study or as an accountability nudge for education systems, for example, as reminders for parents to monitor their child's academic progress[12].

To put the effect sizes of the joint phone and SMS treatment in context, ref. [48] provides benchmarks based on a review of 1,942 effect sizes from 747 randomized controlled trials (RCTs) evaluating education interventions with standardized test outcomes. In this review, 0.10 is the median effect size. A review in ref. [49] also finds 0.10 median effect sizes across 130 RCTs in low- and middle-income countries. Our findings show effect sizes that are around or above the median effect size, with a relatively cheap and scalable intervention. We further include non-standardized effect sizes in Extended Data Fig. 3. We find a 31% reduction in absolute innumeracy (students who cannot do any numerical operations) and an average level gain on the ASER assessment of 0.15 levels (95% CI 0.039, 0.262; $P = 0.008$). As a benchmark, a highly effective in-school educational programme, Teaching at the Right Level, achieved average improvement in math ASER levels of 0.09 to 0.13 in Bihar, India[15]. Furthermore, the learning gains observed were achieved in a total dosage of just 3 h of direct instruction spread over 8 weeks. If effects persist with a higher dosage, up to a 1–2 ASER level gain could potentially be achieved with 20–40 h of instruction, a typical educational programme dosage. Note that learning gains observed might be driven by either learning gains, minimizing learning loss or a combination of both.

In Extended Data Fig. 4, we explore heterogenous treatment effects along three dimensions: student gender, student grade and baseline school exam performance. These variables are typical predictors of learning and were available at baseline. We find limited evidence of heterogeneity along any of these margins, with interaction effects showing no significant effect (see figure for fully reported results). This suggests that the programme works equally well across these subpopulations. One possible explanation for the striking lack of heterogeneity in treatment effects is the focus of the intervention on foundational concepts, which applied to nearly all students. Moreover, since the phone calls were a one-on-one interaction, this ensured that no student was left behind.

Validity checks. We run a series of validity checks for our remote assessments and treatment effects. First, we randomize problems that test the same proficiency, a version of a reliability test used in the psychometric literature[17]. We randomize five problems for each proficiency including for addition, subtraction, multiplication, division and fractions (Table 2). We find that each random problem across all proficiencies is not statistically significantly different compared with a base random problem. Relatedly, we find no difference in treatment effects by the random question received for each proficiency. These tests reveal that the phone-based learning assessment has a high level of internal reliability. Details of statistical results, including P values, standard errors and F-tests are shown in Table 2.

We further disentangle cognitive skills gains from effort effects, which have been shown to affect test scores[18]. In our context, where learning outcomes are measured remotely in the household, effort might be particularly important. We test this hypothesis with a real-effort task requiring one to spend time to think about the question and exert effort or motivation to answer it beyond simple numerical proficiency (see Methods). As shown in column 1 of Extended Data Fig. 5, around 29% of students could answer this question in the control group, and we find no statistically significant changes in effort as a result of any of the interventions ($\beta = 0.016$,

95% CI −0.026, 0.058; $P = 0.448$ and $\beta = 0.021$, 95% CI −0.021, 0.0630; $P = 0.335$). Column 2 shows the effect on average level as a reference. These results indicate that learning gains due to the intervention are largely a function of cognitive skill rather than effort on the test.

It is also possible that learning gains are a matter of familiarity with the content in the intervention groups which received exposure to similar material as on the endline assessment. The familiarity hypothesis is partially tested by randomizing problems of the same proficiency, since this exogenously varies the question asked to minimize overlap with any particular question asked during the intervention itself; this does not change our results. We also test the familiarity hypothesis by including content not covered during the intervention, but which is related, such as place values; as noted earlier, we find that in the phone and SMS group, learning gains can translate to this skill.

We further explore a psychometric validity assessment known as the known-groups method. This approach quantifies whether test scores detect signal across groups that are known to differ[50]. We explore differences in learning level by student age and grade in the control group, two of the factors known to most affect differences in cognitive skills in the status quo. We find in Extended Data Fig. 6 that the assessment detects large and statistically significant differences across both dimensions. For each grade, students score around half an ASER level higher ($P < 0.001$), demonstrating the assessments' ability to differentiate among known groups ($\beta = 0.476$, 95% CI 0.377, 0.576).

We include a series of additional robustness checks in Supplementary Tables 1 and 2, including P values using randomization inference[51] and a joint test of significance for key foundational numeracy learning outcomes. We find small differences in P values overall, and that overall results hold, probably because of the large study sample size, which reduces the likelihood of these P values differing substantially (see Supplementary Tables 1 and 2 for full statistical results).

Lastly, we explore how effects vary on the basis of whether instruction is targeted to the learner's learning level. As seen in Table 1, we find an effect on average level for targeted content of $\beta = 0.076$ (95% CI −0.014, 0.165; $P = 0.097$) and $\beta = 0.070$ on average level for non-targeted content (95% CI −0.021, 0.160; $P = 0.130$). The direct comparison between targeted and non-targeted instruction has a P value of 0.896. Targeted instruction translated to increased learning when compared with the control and improves understanding of place values by 0.098 standard deviations (95% CI 0.012, 0.185; $P = 0.026$). Targeted instruction also benefits learning higher-order competencies such as understanding fractions, with 0.093 standard deviation gains against the control (95% CI 0.004, 0.182; $P = 0.041$). There were no significant effects on learning for non-targeted instruction against the control. The difference between targeted and non-targeted instruction is not statistically significant (see Table 1).

Mechanisms. We explore parental demand and engagement mechanisms. Parental engagement in both interventions is high, with column 1 of Extended Data Fig. 7 showing 92.1% of parents reporting their child attempted to solve any of the problems in the SMS only group (95% CI 0.903, 0.938; $P < 0.001$), and slightly higher engagement of 95.2% in the phone call plus SMS group (95% CI 0.939, 0.966; $P < 0.001$). Table 3 column 3 also shows significant increases in parents' self-efficacy and perceptions as a result of both interventions. Parents report 4.9 (95% CI 0.7, 9.1; $P = 0.023$) and 8.6 (95% CI 4.4, 12.9; $P < 0.001$) percentage points greater self-efficacy in supporting their child's learning in the SMS only, and phone and SMS group, respectively. We also find that parents' confidence that their child made progress on their learning increases from 6.6 (95% CI 2.4, 10.9; $P = 0.002$) to 10.5 (95% CI 6.2, 14.8; $P < 0.001$) percentage points. Moreover, parents of children in the phone call plus SMS

group update their beliefs about their child's learning level in tandem with their child's learning progress (see Table 3 columns 1 and 2). These results reveal that parents are engaged in the intervention and notice their child's progress.

Parents' engagement in their child's math learning might displace other educational activities and non-educational activities, such as returning to work when lockdowns were lifted. In column 6 of Table 3, we find no statistically significant educational crowd-out for both interventions, with no reduction in educational engagement overall ($\beta = -0.001$, 95% CI −0.019, 0.017; $P = 0.933$ and $\beta = -0.002$, 95% CI −0.020, 0.016; $P = 0.809$). In column 5, we find no evidence that parental engagement crowds out non-educational activities such as return to work, with no statistically significant increase in unemployment in the SMS plus phone intervention ($\beta = -2.9$, 95% CI −6.3, 0.5; $P = 0.092$).

Altogether, these results show that remote instruction can change parental beliefs and investments, which play an important role in their child's learning. The Supplementary Information contains details on each of the mechanisms mentioned here, as well details on other robustness checks performed.

Discussion

This paper provides experimental estimates on minimizing the fallout of the COVID-19 pandemic on learning. We show that remote instruction and remote assessment can promote learning. We find that low-tech phone calls plus SMS interventions have large and cost-effective effects on household engagement in education and learning, while SMS messages alone may not.

Both low-tech interventions are relatively low cost. The cost was US$5 per child in the SMS group and $19 dollars per child in the phone and SMS group. Given average treatment effects in the phone and SMS group of 0.12 standard deviations, this translates to 0.63 standard deviation gains for the phone and SMS group per US$100. For those who engaged in all sessions of the programme with a treatment effect of 0.17 standard deviations, this translates into 0.89 standard deviations gained per US$100.

These estimates are cost-effective relative to the literature. As a comparison, providing additional textbooks in Kenya had no effect on learning; halving class size in Kenya and India also had no effect on learning; and conditional cash transfers in Malawi yielded around 0.1 standard deviation per US$100[52]. Another relevant cost-effectiveness comparison is tutoring programmes. An evaluation of remote tutoring with college students in Italy during COVID-19 finds large learning gains of 0.26 standard deviations[44]. A recent review shows that tutoring programmes have been consistently effective across 96 randomized trials[53]. The phone call intervention in our trial compares closely with some of these tutoring programmes. A prominent example yielded 0.19 to 0.31 standard deviation learning gains and cost US$2,500 per child[42]. These comparisons show that the intervention in this study yields similar effects to some of the most effective interventions in the education literature; furthermore, the intervention is contextualized to low-resource contexts and, in some cases, can be more than an order of magnitude cheaper.

Since Botswana is a middle-income country, we note that cost conversions might be necessary when thinking through external validity of cost-effectiveness estimates to a low-income setting[54]. We consider purchasing power parity conversion rates to assess cost differences across contexts, although this is an imperfect conversion. In 2020, the purchasing power parity conversion to US dollars in Botswana was 4.5 according to World Bank data. In contrast, in Kenya, another example country context, the purchasing power parity conversion is 44. Thus, for the same total cost and assuming similar effectiveness, the cost-effectiveness in Kenya could be up to 10 times higher. Future research could collect cost as well as effectiveness data to directly compare cost-effectiveness of similar approaches across settings.

We also show that mobile phones provide a cheap and scalable way to collect information on student learning levels. We find learning gains are robust to a variety of phone-based robustness tests, including randomized problems across the same proficiency and differentiating effort from cognitive skills with real-effort tasks. We further find that gains persist in the phone and SMS treatment across multiple waves of assessment.

In terms of mechanisms, we find high parental engagement in educational activities with their children, high demand and greater self-efficacy to support their child's learning, as well as partial gains in accurate perceptions of their child's level. This finding reveals that parental investments in education can improve their child's learning outcomes even in a low-resource context.

Of note, this study is limited to one context and tests a subset of potential low-tech interventions. Future research might explore similar studies across contexts to assess how well the approach can be adapted across low- and middle-income countries, with growing evidence on similar approaches already emerging[44,45,55,56]. In addition, alternative high-access and low-cost technologies could be compared, such as WhatsApp, or phone calls only with no SMS. For example, some studies have compared WhatsApp for English language instruction to in-person instruction[57]. We present results after a few months in a high-needs school disruption setting. Future research might explore long-run effects. Research shows that short-term school disruption can cause lasting and accumulating damage[5]. It is possible that stemming learning loss during shocks could have far-reaching benefits, with students able to reintegrate into school instruction quickly; it is also possible that benefits fade if all students catch up as schools reopen.

The results in this study have immediate implications for global policy during the current school disruptions, revealing cost-effective and scalable approaches to stem learning loss during the pandemic. Moreover, school closures occur in settings beyond the COVID-19 pandemic, including teacher strikes, summer holidays, public health crises, during adverse weather events, natural disasters, and in refugee and conflict settings. In moments when schooling is disrupted, particularly for families with fewer resources at home, outside-school interventions are needed. Doing so at scale requires cheap, low-technology solutions that can reach as many families as possible. To this end, the results from this study have long-run implications for the role of technology and parents to serve as partial educational substitutes during school disruption, and to provide cost-effective remote instruction and assessment.

Methods

Our study complies with all relevant research protocols. It received Institutional Review Board (IRB) approval from Columbia University Teacher's College (IRB Protocol No. 20-299) and is registered on the AEA RCT registry (AEARCTR-0006044 on 25 June 2020; https://www.socialscienceregistry.org/trials/6044). IRB approval by Columbia University was deemed appropriate and sufficient; the intervention was not sensitive, informed consent was provided by adult caregivers with all participants providing assent, and protocols by the local research institutions that have jurisdiction over education interventions recently decentralized research approvals to regional units and prioritized direct respondent consent procedures rather than providing a centralized IRB protocol. Note that no datapoints were excluded from the analyses.

Population and sampling. A few days before the government announced that schools were closing as a result of the state of emergency, we collected 7,550 phone numbers from primary schools. This response built on an active presence in schools by Youth Impact (previously known as Young 1ove), one of the largest non-government organizations in Botswana, which was conducting educational programming in partnership with the Ministry of Basic Education. These numbers were collected for students in grades 3 to 5; however, only a subset of these numbers were reachable often due to arbitrary reasons, such as misrecorded digits. After phone collection and verification, facilitators called all numbers to confirm interest from parents in receiving remote learning support via phone. As described in Supplementary Information Section A, the final sample characteristics match those of a nationally representative sample on a host of indicators, such as learning levels and parent education.

For parents who opted into remote learning support, we provided two low-tech interventions: (1) one-way bulk SMS texts with multiple numeracy 'problems of the week' and (2) SMS bulk texts with live phone call walkthroughs of the problems on a 15–20 min phone call. Both low-tech interventions were intentionally designed to be simple so as to be digestible via phone by parents, teachers and students, and scalable by governments.

Treatments. The first intervention was a weekly SMS containing several simple math problems; for example, "Sunshine has 23 sweets. She goes to the shops to buy 2 more. How many does she have altogether?". The SMS was sent at the beginning of each week via a bulk texting platform. The SMS contained a message with 160 to 320 characters that could fit in one or two texts. Supplementary Fig. 4 shows an example weekly message of practice problems focused on place value.

The second intervention was a weekly phone call ranging in typical length from 5–20 min in addition to the weekly SMS, which was sent at the beginning of the week. On the call, the facilitator asked the parent to find the student and put the call on speaker. This arrangement allowed both the parents and the student to hear the facilitator at the same time and to engage in learning. The facilitator confirmed that the student had received the SMS message sent and answered any questions related to the task. Furthermore, the facilitator provided the student with a math question to go over and practice. The calls served to provide additional learning support as well as motivation and accountability. Supplementary Fig. 5 includes a subset of a sample phone call script.

A subset of phone numbers also received an additional intervention: targeted instruction to each child's learning level. We used data on learning levels from a midline phone-based learning assessment to send tailored text messages to each student in the fifth week. For example, students who knew addition received subtraction problems, whereas students who knew multiplication were sent division problems. This targeted instruction programme used data collected at week 4 to have near real-time data to target instruction. We collected additional endline survey data and conducted learning assessments which enabled evaluation of the targeted instruction component of the intervention.

The targeted instruction component of the intervention relates to a literature on targeted instruction. An educational approach called 'Teaching at the Right Level' (TaRL), a classroom-based intervention evaluated over 20 years targeting instruction by learning level rather than by age or grade, has been shown to produce cost-effective gains in learning across multiple studies. This approach has worked when delivered by teachers or volunteers[13–16,58] and when using adaptive computer software[13,59]. We tested a particularly low-cost and scalable approach to targeted instruction using phone-based assessments and instruction.

Data collection. We conducted two waves of data collection. The endline occurred after 4 months and a midline occurred shortly before the halfway point. The endline survey consisted of 17 questions including a learning assessment, parental engagement in educational activities, and parental perceptions of their own self-efficacy and their child's learning. A portion of the survey was conducted with the parent and learning outcomes were collected by directly assessing the child over the phone.

The learning assessment was adapted from the ASER test, which consists of multiple numeracy items, including 2-digit addition, subtraction, multiplication and division problems. Supplementary Fig. 6 shows a sample assessment. We focused on basic numeracy interventions and assessment. Learning gains might also translate to literacy; however, we did not measure literacy gains in this study. To maximize the reliability of the phone-based assessment, we introduced a series of quality-assurance measures: students had a time cap of 2 min per question to minimize the likelihood of family members in the household assisting the child, and we asked each child to explain their work and only marked a problem correct if the child could correctly explain how they solved the problem. We assigned facilitators to phone numbers using an arbitrary match sorted by phone number order. On average, each facilitator was assigned to about 30 phone numbers. Less than 1.5% of facilitators that provided weekly intervention calls surveyed the same household, providing for objective assessment. While imperfect, these measures provide a level of verification to maximize the likelihood that the test captures child learning. We previously discussed practical steps to implement learning measurement via phone[60]. We also conducted several checks to validate measures, as described below.

In addition to the ASER test, we evaluated the children's ability to answer a simple place value word problem such as "Katlego has 77 apples and organizes them by place value. How many tens does she have?" to capture learning outcomes beyond a core set of mathematical operations. We included a series of additional questions to identify mechanisms driving learning gains. This includes a measure of student effort using the following question: "The day before two days from now is Saturday. What day is today?". Note that this question is not necessarily a pure measure of student effort and could capture other related capabilities. There is no standard measurement of student effort. Real-effort tasks range from solving mazes[61] to adding series of 2-digit numbers[62], and 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[63]. Since we aimed to differentiate numerical ability from effort, we built on arithmetic real-effort tasks but chose a

problem that had easier arithmetic than our numeracy assessment (single digit addition and subtraction), yet required non-arithmetic effort. We also included a higher-order numeracy question to assess whether learning gains translate to material not covered directly in the intervention. In particular, we asked a question on fractions such as '$\frac{3}{8} + \frac{5}{8} = ?$'. We further conducted a reliability assessment by randomizing five different questions of each proficiency (addition, subtraction, multiplication, division and fractions) to formally assess the reliability of the learning assessment questions[17]. For example, for a division problem, we had one problem that asked students to divide 68 by 5 and another problem where 38 is divided by 3. Both are 2-digit division problems with remainder. If results for both problems have similar results, given that they measure the same latent ability, this increases our confidence in learning estimates.

We also included questions on parental engagement, perceptions and self-efficacy. We measured learning engagement by asking parents if they recalled their child attempting any of the problems sent over the last few weeks. We included a measure of a parent's perception of their child's numeracy level by directly matching their perception of their child's level to their child's actual learning level. If a parent estimates the highest level their child can do is subtraction, and their child indeed performs up to subtraction level, we code this as 'correct'. If the parent overestimates or underestimates their child's level, we code this as 'incorrect'. We also captured parents' confidence in supporting their child's learning at home and whether they felt their child made progress during the school closure period. We coded a dummy variable for whether parents are 'very confident' for both indicators. Additional questions included information on whether the caregiver has returned to work. Finally, demographic questions recorded the child's age, grade and gender. A sample survey is included in Supplementary Information (Additional Supplementary Information (1)).

We also conducted a similar midline assessment to cross-randomize targeted instruction (described above) and asked about demand for remote learning services if schools were to reopen.

Empirical strategy. We estimated treatment effects of the SMS only and phone and SMS intervention using the following specification:

$$Y_{ij} = \alpha_0 + \beta_1 SMS_j + \beta_2 PhoneSMS_j + \delta_s + \varepsilon_{ij}$$

where Y_{ij} is an outcome for child i in randomly assigned household j. SMS is an indicator variable coded to one for the SMS message only treatment group and zero otherwise, and PhoneSMS is an indicator variable coded to 1 if a household received both an SMS and a phone call and zero otherwise. δ_s is a strata indicator, which indicates whether a child participated in education programming immediately before the intervention. α_0 is a constant, ε_{ij} is an error term; β_1 and β_2 represent treatment effects of the SMS arm as well as the Phone and SMS arm relative to the control group, respectively. We used this specification to measure the impact of each intervention on students' learning level, engagement, and parents' perceptions of their child's level and self-efficacy. We included the primary child identified for instruction in each household level j, which is determined by the caregiver's phone number and is the unit of randomization. The vast majority of households had one child in the household at the time of the programme. Only 413 households had 2 children in the household, and only 52 households had 3 or more children in the household. Moreover, the number of children was unbiased across treatment arms (a regression which codes the number of kids assessed per household as a continuous variable finds a coefficient of −0.02 fewer kids in any treatment group relative to the control, with a P value 0.329 revealing no statistically significant differences in the number of students per household across arms (95% CI −0.052, 0.011)). As a robustness test, we ran a regression on the main learning results for households with only one participating child per household and found average learning gains of 0.136 standard deviations compared with 0.12 for the full sample (95% CI 0.041, 0.232; $P = 0.005$).

We also estimated the effect of targeted instruction with the following specification:

$$Y_{ij} = \alpha_0 + \beta_1 Targeted_j + \beta_2 NotTargeted_j + \delta_s + \varepsilon_{ij}.$$

Given randomization and equivalent treatment and control groups, each specification identifies causal effects of the intervention. In Supplementary Tables 3 and 4, we show no statistically significant survey response rate differences between treatment groups compared with the control group or each other, suggesting that endline outcomes are unbiased across study groups. We also show no statistically significant differences between groups on baseline characteristics, providing evidence that randomization was successfully implemented. Note that all statistical tests performed are two-tailed, and given the large sample size in this study, data distribution is assumed to be plausibly normal.

Statistical power. We conducted power calculations in Stata using estimates based on the literature. We estimated that our study sample is sufficiently large to detect 0.075 standard deviation gains in learning. Given that median effects in the education literature are approximately 0.10 standard deviations[48,49], these initial power calculations suggest that we could detect clinically meaningful results informed by the literature with high statistical certainty.

Data availability
This study used primary outcome data collected through surveys. The data will be anonymized and made available for replication on GitHub at https://github.com/nangrist/schoolsout.

Code availability
The code to replicate analysis and tables will be made publicly available on GitHub at https://github.com/nangrist/schoolsout.

References
1. *Education: From Disruption to Recovery* (UNESCO, 2020); https://en.unesco.org/covid19/educationresponse
2. Cooper, H., Nye, B., Charlton, K., Lindsay, J. & Greathouse, S. The effects of summer vacation on achievement test scores: a narrative and meta-analytic review. *Rev. Educ. Res.* **66**, 227–268 (1996).
3. Slade, T. S., Piper, B., Kaunda, Z., King, S. & Ibrahim, H. Is 'summer' reading loss universal? Using ongoing literacy assessment in Malawi to estimate the loss from grade-transition breaks. *Res. Comp. Int. Educ.* **12**, 461–485 (2017).
4. Jaume, D. & Willén, A. The long-run effects of teacher strikes: evidence from Argentina. *J. Labor Econ.* **37**, 1097–1139 (2019).
5. Andrabi, T., Daniels, B. & Das, J. Human capital accumulation and disasters: evidence from the Pakistan earthquake of 2005. *J. Hum. Resour.* https://doi.org/10.3368/jhr.59.2.0520-10887r1 (2021).
6. Bacher-Hicks, A., Goodman, J. & Mulhern, C. Inequality in household adaptation to schooling shocks: Covid-induced online learning engagement in real time. *J. Public Econ.* **193**, 104345 (2021).
7. Chetty, R., Friedman, J. N., Hendren, N. & Stepner, M. *The Economic Impacts of Covid-19: Evidence from a New Public Database Built Using Private Sector Data* Working Paper 27431 (National Bureau of Economic Research, 2020).
8. Engzell, P., Frey, A. & Verhagen, M. D. Learning loss due to school closures during the COVID-19 pandemic. *Proc. Natl Acad. Sci. USA* **118**, 17 (2021).
9. Parolin, Z. & Lee, E. K. Large socio-economic, geographic and demographic disparities exist in exposure to school closures. *Nat. Hum. Behav.* **5**, 522–528 (2021).
10. Carvalho, S. & Crawfurd, L. *School's Out: Now What?* (Center for Global Development, 2020); https://www.cgdev.org/blog/schools-out-now-what
11. Pansiri, N., Tsayang, G. & Young 1ove. *A Situational Analysis of Basic Literacy and Numeracy at Early Grade Levels in Botswana.* Unpublished Report (2017).
12. Bergman, P. Parent-child information frictions and human capital investment: evidence from a field experiment. *J. Polit. Econ.* **129**, 286–322 (2021).
13. Banerjee, A. V., Cole, S., Duflo, E. & Linden, L. Remedying education: evidence from two randomized experiments in India. *Q. J. Econ.* **122**, 1235–1264 (2007).
14. Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R. & Khemani, S. Pitfalls of participatory programs: evidence from a randomized evaluation in education in India. *Am. Economic J. Econ. Policy* **2**, 1–30 (2010).
15. Banerjee, A. et al. From proof of concept to scalable policies: challenges and solutions, with an application. *J. Econ. Perspect.* **31**, 73–102 (2017).
16. Duflo, A., Kiessel, J. & Lucas, L. *External Validity: Four Models of Improving Student Achievement.* Working Paper 27298 (National Bureau of Economic Research, 2020).
17. Crocker, L. M. & Algina, J. *Introduction to Classical and Modern Test Theory* (Holt, Rinehart and Winston, 1986).
18. Gneezy, U. et al. Measuring success in education: the role of effort on the test itself. *Am. Econ. Rev. Insights* **1**, 291–308 (2019).
19. Aker, J. C., Ksoll, C. & Lybbert, T. J. Can mobile phones improve learning? evidence from a field experiment in Niger. *Am. Econ. J. Appl. Econ.* **4**, 94–120 (2012).
20. Aker, J. C., Boumnijel, R., McClelland, A. & Tierney, N. Payment mechanisms and antipoverty programs: evidence from a mobile money cash transfer experiment in Niger. *Econ. Dev. Cult. Change* **65**, 1–37 (2016).
21. Aker, J. C. & Ksoll, C. Can ABC lead to sustained 123? The medium-term effects of a technology-enhanced adult education program. *Econ. Dev. Cult. Change* **68**, 1081–1102 (2020).
22. York, B. N., Loeb, S. & Doss, C. One step at a time: the effects of an early literacy text messaging program for parents of preschoolers. *J. Hum. Resour.* **54**, 537–566 (2018).
23. Doss, C., Fahle, E. M., Loeb, S. & York, B. N. More than just a nudge supporting kindergarten parents with differentiated and personalized text messages. *J. Hum. Resour.* **54**, 567–603 (2019).

24. Cortes, K. E., Fricke, H., Loeb, S., Song, D. S. & York, B. N. Too little or too much? Actionable advice in an early-childhood text messaging experiment. *Educ. Finance Policy* **16**, 209–232 (2021).

25. Bettinger, E., Cunha, N., Lichand, G. & Madeira, R. *Are the Effects of Informational Interventions Driven by Salience?* Working Paper No. 350 (University of Zurich, Department of Economics, 2021).

26. Lichand, G. & Wolf, S. *Arm-Wrestling in the Classroom: The Non-Monotonic Effects of Monitoring Teachers*. Working Paper No. 357 (University of Zurich, Department of Economics, 2021).

27. Kraft, M. A. & Rogers, T. The underutilized potential of teacher-to-parent communication: evidence from a field experiment. *Econ. Educ. Rev.* **47**, 49–63 (2015).

28. Berlinski, S., Busso, M., Dinkelman, T. & Martínez, C. A. *Reducing Parent-School Information Gaps and Improving Education Outcomes: Evidence from High-Frequency Text Messages*. Working Paper 28581 (National Bureau of Economic Research, 2021).

29. Cunha, N., Lichand, G., Madeira, R. & Bettinger, E. *What is It About Communicating with Parents?* Working Paper (Stanford CEPA, 2017); https://cepa.stanford.edu/sites/default/files/cunha_cover_paper1.pdf

30. Siebert, W. S., Wei, X., Wong, H. L. & Zhou, X. *Student Feedback, Parent-Teacher Communication, and Academic Performance: Experimental Evidence from Rural China*. Discussion Paper No. 11347 (IZA Institute of Labor Economics, 2018).

31. De Walque, D. & Valente, C. *Incentivizing School Attendance in the Presence of Parent-Child Information Frictions*. Policy Research Working Paper 8476 (The World Bank, 2018).

32. Rogers, T. & Feller, A. Reducing student absences at scale by targeting parents' misbeliefs. *Nat. Hum. Behav.* **2**, 335–342 (2018).

33. Bergman, P. & Chan, E. W. Leveraging parents through low-cost technology: the impact of high-frequency information on student achievement. *J. Hum. Resour.* **56**, 125–158 (2021).

34. Musaddiq, T., Prettyman, A. & Smith, J. School messaging platforms and student attendance. *Metro Atlanta Policy Lab for Education, Georgia Policy Labs* https://gpl.gsu.edu/download/school-messaging-platforms-and-student-attendance-policy-brief/?wpdmdl=1958&refresh=61a69b3152 2d11638308657 (2020).

35. Gallego, F. A., Malamud, O. & Pop-Eleches, C. Parental monitoring and children's internet use: the role of information, control, and cues. *J. Public Econ.* **188**, 104208 (2020).

36. Bergman, P. How behavioral science can empower parents to improve children's educational outcomes. *Behav. Sci. Policy* **5**, 52–67 (2019).

37. Godino, J. G. et al. Text messaging and brief phone calls for weight loss in overweight and obese English- and Spanish-speaking adults: a 1-year, parallel-group, randomized controlled trial. *PLoS Med.* **16**, e1002917 (2019).

38. Bettinger, E. et al. *Does EdTech Substitute for Traditional Learning? Experimental Estimates of the Educational Production Function*. Working Paper No. 26967 (National Bureau of Economic Research, 2020).

39. *World Development Report 2018: Learning to Realize Education's Promise* (The World Bank, 2018); https://www.worldbank.org/en/publication/wdr2018

40. Angrist, N. et al. *How to Improve Education Outcomes Most Efficiently? A Comparison of 150 Interventions Using the New Learning-Adjusted Years of Schooling Metric*. Policy Research Working Paper No. 9450 (The World Bank, 2020).

41. Angrist, N., Djankov, S., Goldberg, P. K. & Patrinos, H. A. Measuring human capital using global learning data. *Nature* **592**, 403–408 (2021).

42. Guryan, J. et al. *Not Too Late: Improving Academic Outcomes Among Adolescents*. Working Paper No. 28531 (National Bureau of Economic Research, 2021).

43. Josephson, A., Kilic, T. & Michler, J. D. Socioeconomic impacts of COVID-19 in low-income countries. *Nat. Hum. Behav.* **5**, 557–565 (2021).

44. Carlana, M. & La Ferrara, E. *Apart but Connected: Online Tutoring and Student Outcomes during the COVID-19 Pandemic*. Discussion Paper No. 14094 (IZA Institute of Labor Economics, 2021).

45. Hassan, H., Islam, A., Siddique, A. & Wang, L. C. *Telementoring and Homeschooling During School Closures: A Randomized Experiment in Rural Bangladesh* (IZA Institute of Labor Economics, 2021); https://conference.iza.org/conference_files/covid_2021/siddique_a25232.pdf

46. Azevedo, J. P., Hasan, A., Goldemberg, D., Geven, K. & Iqbal, S. A. Simulating the potential impacts of COVID-19 school closures on schooling and learning outcomes: a set of global estimates. *World Bank Res. Obs.* **36**, 1–40 (2021).

47. Evans, D. K. & Acosta, A. M. Education in Africa: what are we learning? *J. Afr. Econ.* **30**, 13–54 (2021).

48. Kraft, M. A. Interpreting effect sizes of education interventions. *Educ. Res.* **49**, 241–253 (2020).

49. Evans, D. K. & Yuan, F. *How Big are Effect Sizes in International Education Studies*. Working Paper No. 545 (Center for Global Development, 2020).

50. Hattie, J. & Cooksey, R. W. Procedures for assessing the validities of tests using the "known-groups" method. *Appl. Psychol. Meas.* **8**, 295–305 (1984).

51. Young, A. Channeling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results. *Q. J. Econ.* **134**, 557–598 (2019).

52. Kremer, M., Brannen, C. & Glennerster, R. The challenge of education and learning in the developing world. *Science* **340**, 297–300 (2013).

53. Nickow, A., Oreopoulos, P. & Quan, V. *The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence*. Working Paper No. 27476 (National Bureau of Economic Research, 2020).

54. Evans, D. K. & Popova, A. Cost-effectiveness analysis in development: accounting for local costs and noisy impacts. *World Dev.* **77**, 262–276 (2016).

55. Crawfurd, L., Evans, D. K., Hares, S. & Sandefur, J. *Teaching and Testing by Phone in a Pandemic*. Working Paper No. 591 (Center for Global Development, 2021).

56. Schueler, B. E. & Rodriguez-Segura, D. *A Cautionary Tale of Tutoring Hard-to-Reach Students in Kenya*. EdWorkingPaper No. 21-432 (Annenberg Institute at Brown University, 2021).

57. Dehghan, F., Rezvani, R. & Fazeli, S. A. Social networks and their effectiveness in learning foreign language vocabulary: a comparative study using WhatsApp. *CALL EJ* **18**, 1–13 (2017).

58. Duflo, E., Dupas, P. & Kremer, M. Peer effects, teacher incentives, and the impact of tracking: evidence from a randomized evaluation in Kenya. *Am. Econ. Rev.* **101**, 1739–1774 (2011).

59. Muralidharan, K., Singh, A. & Ganimian, A. J. Disrupting education? Experimental evidence on technology-aided instruction in India. *Am. Econ. Rev.* **109**, 1426–1460 (2019).

60. Angrist, N. et al. Practical lessons for phone-based assessments of learning. *BMJ Glob. Health* **5**, e003030 (2020).

61. Gneezy, U., Niederle, M. & Rustichini, A. Performance in competitive environments: gender differences. *Q. J. Econ.* **118**, 1049–1074 (2003).

62. Niederle, M. & Vesterlund, L. Do women shy away from competition? Do men compete too much? *Q. J. Econ.* **122**, 1067–1101 (2007).

63. Zamarro, G., Hitt, C. & Mendez, I. When students don't care: reexamining international differences in achievement and student effort. *J. Hum. Cap.* **13**, 519–552 (2019).

64. Angrist, N., Bergman, P., Brewster, C. & Matsheng, M. *Stemming Learning Loss During the Pandemic: A Rapid Randomized Trial of a Low-Tech Intervention in Botswana*. Working Paper WPS/2020-13 (Center for the Study of African Economics, 2020).

Acknowledgements

Learning (PAL) network. The infrastructure built by this coalition before COVID-19 enabled this rapid trial and response.

Author contributions

N.A. and M.M. conceived the study and oversaw data collection efforts. N.A. conducted the analysis, generated tables and figures, and wrote the drafts of the manuscripts. N.A. and P.B. oversaw survey and experimental design. P.B. edited drafts of the manuscripts. N.A., P.B. and M.M. provided substantial intellectual inputs throughout the study and manuscript drafting.

Competing interests

The authors declare no competing financial interests. Noam Angrist and Moitshepi Matsheng are co-founders of Youth Impact (previously called Young 1ove), one of the research and implementing partners of the primary research.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-022-01381-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-022-01381-z.

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

Peer review information *Nature Human Behaviour* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

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

	Study Sample	National Data
Regions Represented	8	9
Demographics		
Female	0.52	0.50
Percent Rural	0.29	0.29
Primary School Leaving Exam Scores		
Percent A	0.16	0.15
Percent B	0.21	0.18
Percent C	0.41	0.37
Caregiver Characteristics		
Beyond Secondary School Completion	0.29	0.26

Extended Data Fig. 1 | Sample characteristics and representativeness. This table reports and compares characteristics of the study sample with national-level indicators. Data for national level indicators is taken from a variety of sources: Percentage female and primary school leaving exam scores are taken from Botswana's Examinational Council Primary School Leaving Exam (PSLE) data; Rural population (% of total population) and national-level school enrolment, tertiary (% gross) are taken from the World Bank are taken from the World Bank. The study sample refers to school-level indicators matched to the original school of the study sample for the "demographics" and "primary school leaving exam scores" categories and to individuals characteristics for the caregiver characteristics.

	(1) Avg Level	(2) Avg Level	(3) Avg Level
Phone + SMS	0.121 (0.046) [0.008] {0.031,0.210}		
Phone + SMS - Per Session		0.028 (0.010) [0.007] {0.008,0.048}	
Phone + SMS - All Sessions			0.167 (0.062) [0.007] {0.046,0.289}
Observations	2815	1878	1878

Extended Data Fig. 2 | Treatment on the treated effects. This table shows treatment effects in terms of standard deviations. Column (1) reports intention-to-treat (ITT) effects at endline. Column (2) reports treatment-on-the-treated (TOT) using instrumental variables estimation with random assignment to the Phone and SMS group as an instrument for a continuous measure of participation per session in the Phone and SMS group. Column (3) reports extrapolated treatment-on-the-treated (TOT) estimates in the Phone and SMS group if households attended all sessions. We do not have similarly rich week-by-week implementation data in the SMS group to conduct a meaningful TOT analysis. The observation count is lower in Columns (2)-(3) than Column (1) since we exclude the SMS group in the regression. Standard errors are in parentheses, p-values are in square brackets, and 95% confidence intervals are in curly brackets. All statistical tests are two-tailed.

	(1) Avg Level	(2) Innumerate	(3) Division	(4) Place Value	(5) Fractions
SMS Only	0.030	-0.010	0.011	0.004	0.021
	(0.057)	(0.013)	(0.020)	(0.019)	(0.021)
	[0.602]	[0.460]	[0.594]	[0.837]	[0.309]
	{-0.083,0.142}	{-0.035,0.016}	{-0.028,0.049}	{-0.034,0.042}	{-0.019,0.061}
Phone + SMS	0.150	-0.029	0.050	0.049	0.034
	(0.057)	(0.012)	(0.020)	(0.019)	(0.021)
	[0.008]	[0.022]	[0.013]	[0.009]	[0.100]
	{0.039,0.262}	{-0.053,-0.004}	{0.010,0.090}	{0.012,0.085}	{-0.006,0.074}
Control Mean	2.459	0.093	0.235	0.759	0.721
Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes
Observations	2815	2815	2815	2881	2751
p-val: SMS = Phone	0.033	0.121	0.053	0.017	0.528

Extended Data Fig. 3 | Learning gains, non-standardized and by proficiency. This table reports results from a regression estimating treatment effects of the SMS only and phone and SMS intervention on learning gains which are not standardized. Average level 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). We disaggregate and report proficiencies on the low and high end: innumeracy (defined as whether a child can do no operations) and whether a child can do division. Place value refers to a distinct place value question. Fractions refers to a distinct question to solve a higher-order fractions problems. Standard errors are in parentheses, p-values are in square brackets, and 95% confidence intervals are in curly brackets. All statistical tests are two-tailed.

	Avg Level		
	(1)	(2)	(3)
	Female	Grade	School Performance
SMS Only	0.094	0.175	0.391
	(0.085)	(0.294)	(0.773)
	[0.267]	[0.551]	[0.613]
	{-0.072,0.260}	{-0.401,0.752}	{-1.125,1.907}
Phone + SMS	0.196	0.140	0.079
	(0.084)	(0.293)	(0.787)
	[0.020]	[0.633]	[0.920]
	{0.031,0.361}	{-0.434,0.714}	{-1.464,1.622}
Heteroeneity Effects Variable	0.204	0.476	-0.001
	(0.081)	(0.051)	(0.002)
	[0.012]	[0.000]	[0.597]
	{0.045,0.363}	{0.376,0.576}	{-0.005,0.003}
SMS Only # Heterogeneity Effects Variable	-0.130	-0.036	-0.001
	(0.115)	(0.071)	(0.003)
	[0.259]	[0.608]	[0.625]
	{-0.356,0.096}	{-0.175,0.102}	{-0.007,0.004}
Phone + SMS # Heterogeneity Effects Variable	-0.109	-0.000	0.000
	(0.114)	(0.070)	(0.003)
	[0.340]	[0.995]	[0.973]
	{-0.333,0.115}	{-0.138,0.137}	{-0.006,0.006}
Control Mean	2.459	2.459	2.459
Strata Fixed Effects	Yes	Yes	Yes
Observations	2794	2795	2170

Extended Data Fig. 4 | Learning gains, heterogeneous treatment effects. This table reports heterogeneous treatment effects by three dimensions: gender (if female), grade level (ranging from grades 3 to 5) and baseline school-level exam performance at the origin school of each participating child. Each column represents on model, which includes the dimension specified (for example, female) as an interaction term, referred to as "heterogeneity effects variable". The dependent variable is average level, which 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). Standard errors are in parentheses, p-values are in square brackets, and 95% confidence intervals are in curly brackets. All statistical tests are two-tailed.

	(1) Effort Task	(2) Avg Level
SMS Only	0.016	0.030
	(0.021)	(0.057)
	[0.448]	[0.602]
	{-0.026,0.058}	{-0.083,0.142}
Phone + SMS	0.021	0.150
	(0.021)	(0.057)
	[0.335]	[0.008]
	{-0.021,0.063}	{0.039,0.262}
Control Mean	0.290	2.459
Strata Fixed Effects	Yes	Yes
Observations	2732	2815
p-val: SMS = Phone	0.839	0.033

Extended Data Fig. 5 | Robustness check: effort on the test. This table reports results from a regression estimating effects of the SMS only and phone and SMS intervention on a real-effort task. Effort is contrasted with results on average learning level. Standard errors are in parentheses, p-values are in square brackets, and 95% confidence intervals are in curly brackets. All statistical tests are two-tailed.

	(1) Avg Level	(2) Avg Level
Age of Student	0.179	
	(0.027)	
	[0.000]	
	{0.127,0.231}	
Grade of Student		0.476
		(0.051)
		[0.000]
		{0.377,0.576}
Observations	935	938

Extended Data Fig. 6 | Known-groups validity test. This table reports the relationship between two variables known to correlate strongly with learning outcomes: age and the grade of students. We report the relationship in the control group, as a validity assessment of whether the test used can detect expected learning differences among "known groups". Standard errors are in parentheses, p-values are in square brackets, and 95% confidence intervals are in curly brackets. All statistical tests are two-tailed.

	Engagement	Demand		
	(1) Did Problems	(2) Phone and SMS	(3) SMS Only	(4) None
SMS Only	0.921	-0.027	0.077	-0.005
	(0.009)	(0.030)	(0.026)	(0.005)
	[0.000]	[0.363]	[0.003]	[0.322]
	{0.903,0.938}	{-0.086,0.031}	{0.026,0.129}	{-0.015,0.005}
Phone + SMS	0.952	0.177	-0.102	0.003
	(0.007)	(0.026)	(0.021)	(0.007)
	[0.000]	[0.000]	[0.000]	[0.639]
	{0.939,0.966}	{0.125,0.229}	{-0.144,-0.060}	{-0.010,0.016}
Control Mean	0.000	0.693	0.176	0.009
Observations	3405	1478	1478	1478
p-val: SMS = Phone	0.005	0.000	0.000	0.139

Extended Data Fig. 7 | Engagement and demand. This table reports results from a regression estimating effects of the SMS only and phone and SMS intervention on engagement questions at endline and demand questions at midline. We code engagement at zero for the control group since, by definition, they did not receive problems to respond to. For demand, we report demand at midline since this question was asked at the halfway point, with particular emphasis on demand for the interventions even if schools were to re-open. The observation count is lower for demand since a random subset of households received the midline. Standard errors are in parentheses, p-values are in square brackets, and 95% confidence intervals are in curly brackets. All statistical tests are two-tailed.

nature portfolio

Corresponding author(s):	Noam Angrist
Last updated by author(s):	Mar 1, 2022

Reporting Summary

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

Statistics

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

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

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	This study used primary outcome data collected through direct surveys. Some additional data was used to analyze how representative the sample is. This data was obtained from the Ministry of Basic Education and the Botswana Examinations Council, as well as the World Bank.
Data analysis	We conduct all analysis in Stata 17. The code to replicate analysis and tables will be made publicly available on a public Github link specified in the paper.

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

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

This study used primary outcome data collected through surveys. The data will be anonymized and made available for replication in a public Github link specified in the paper.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

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

Study description	We conduct a large-scale randomized trial with 4,500 households testing two low-technology interventions – SMS messages and phone calls – with caregivers to support educational instruction for their child in Botswana.
Research sample	The study took place in Botswana with 4,550 households. We compare our sample to national level indicators and find that those who enroll are broadly representative of national characteristics. Botswana has nine regions in total and our sample covers 8 out of all 9 regions. Our study sample includes over 103 schools which represents around 15 percent of all primary schools in the country. We compare our study sample to national-level indicators from the Ministry of Basic Education using data on enrollment and gender composition from 2017. We find a similar gender split between 50 to 52 percent in both samples. We also find a similar ratio of rural students in our sample to the national average of 29 percent. We compare study schools on the Primary School Leave Examinations (PSLE) from the Botswana Examinations Council in grades 7. We find similar distributions of learning: the percentage of students who score an A, B, and C is 16, 21, and 41 percent in study schools, respectively, and 14, 17, and 36 for all primary schools in the nation. In addition, we collect simple descriptive data on child age, grade, and gender in surveys. Around 50 percent of our sample is female; the average age of students is 9.7; 28.5 percent of students are in grade 3, 39.1 percent in grade 4 and 32.4 percent in grade 5. The average age of parents or caregivers participating in the randomized trial was 35, and 68 percent of parents were female. In a subsample of households where we collected data on parent education, 29 percent of parents had completed schooling beyond secondary school, relative to 26 percent on average nationally based on data from the World Bank. These measures suggest the sample of parents have similar education levels to the national average.
Sampling strategy	Phone numbers were collected in primary schools across the country prior to school closures for students and their caregivers in grades 3 to 5. The sampling strategy was designed to be national representative by collecting numbers in schools across the country. The sample size was determined based on power calculations using Stata software. To detect a .1 standard deviation effect, 800 to 1,000 students we needed per treatment arm. In order to account for response rates, the total sample was increased to around 1,500 per arm, resulting in just over 4,500 students in the total sample across three arms.
Data collection	We conducted data collection through direct surveys using a mobile phone, and enumerators were treatment blind. We required the primary caregiver and the respondent (the child) to be together during the survey. The endline occurred after 4 months and a mid-line occurred shortly before the halfway point. The endline survey consists of 17 questions including a learning assessment, parental engagement in educational activities, and parental perceptions of their own self-efficacy and their child's learning. A portion of the survey was conducted with the parent and learning outcomes were collected by directly assessing the child over the phone. The learning assessment was adapted from the ASER test, a test used frequently in the literature and which consists of multiple numeracy items, including 2-digit addition, subtraction, multiplication, and division problems. In order to maximize the reliability of the phone-based assessment, we introduced a series of quality-assurance measures: students had a time cap of two minutes per question to minimize the likelihood of family members in the household assisting the child, and we asked each child to explain their work and only marked a problem correct if the child could correctly explain how they solved the problem. We also introduced a series of validity tests described in the paper in depth.
Timing	Mid April to Late July 2021.
Data exclusions	No data were excluded from the analyses.
Non-participation	71 percent of working phone numbers screened were reachable, interested, and gave consent to participate in the trial. As mentioned in the paper, the set of final participants were broadly national representative along multiple indicators, including learning levels, the degree of rural vs urban regions, parent education, and gender.
Randomization	Allocation to treatment groups was done via random assignment using Stata software and stratified by prior enrollment in educational programming.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	See above.
Ethics oversight	This trial received IRB Approval from Columbia University, Teacher's College (IRB Protocol #: 20-299) and is registered on the AEA RCT registry: AEARCTR-0006044.

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