Teaching and Testing by Phone in a Pandemic

Lee Crawfurd, David K. Evans, Susannah Hares, and Justin Sandefur

Abstract

How did children learn while schools were closed during 2020 due to the COVID-19 pandemic? In low-income countries where internet access is scarce, distance learning is often passive, via TV or radio, with little opportunity for teacher-student interaction. In this paper we evaluate the effectiveness of live tutoring calls from teachers, using a randomized controlled trial with 4,399 primary school students in Sierra Leone. Tutoring calls increased engagement in educational activity but had no effect on mathematics or language test scores, for girls or boys. We also make a methodological contribution, testing the reliability of student assessments conducted by phone. Phone-based assessments have sensible properties, but we find suggestive evidence that scores are higher than with in-person assessments, and there is differential item functioning across survey modes for most individual questions.

Keywords: Education, COVID, distance learning, teachers

JEL: I10, I20, I25, O15

Center for Global Development
www.cgdev.org

Working Paper 591
September 2021

Teaching and Testing by Phone in a Pandemic

Lee Crawfurd
Center for Global Development

David K. Evans
Center for Global Development

Susannah Hares
Center for Global Development

Justin Sandefur
Center for Global Development

This trial is registered in the AEA RCT registry at: AEARCTR-0005914 and received approval from the Sierra Leone Ethics and Scientific Review Committee. We thank Yatta Kanu, Robert Sam-Kpakra, Sensi Tech Hub (including Benjamin Franklyn Bright, James Yorpoi, James Tarawalie, Mahmoud Yanka, and a team of enumerators), and Montrose International (especially Lindsay Ellis) for support with instrument design and data collection. We thank Mari Shojo, Marcela Gutierrez Bernal, Alaka Holla, Shelby Carvalho, Elisabetta Aurino, Francisco Carballo, Stephanie Dobrowolski, and Paul Skidmore for helpful comments. All errors are our own.

The Center for Global Development is grateful for contributions from the World Bank Strategic Impact Evaluation Fund and the Bill & Melinda Gates Foundation in support of this work.

Lee Crawfurd, David K. Evans, Susannah Hares, and Justin Sandefur, 2021. "Teaching and Testing by Phone in a Pandemic." CGD Working Paper 591. Washington, DC: Center for Global Development. https://www.cgdev.org/publication/teaching-and-testing-phone-pandemic

Center for Global Development
2055 L Street NW
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

The views expressed in CGD Working Papers are those of the authors and should not be attributed to the board of directors, funders of the Center for Global Development, or the authors' respective organizations.

1. Introduction

Schools closed worldwide in March 2020 in response to the COVID-19 pandemic. These closures present a number of policy challenges for governments. First, children miss out on direct learning in classrooms. In several countries, children lost ten percent or more of the total time they were expected to spend in-person at school over the course of their lives (Evans et al., 2021). Second, children may forget much of what they already learned in school. In high- and low-income countries, students–especially lower income students–regress in their academic skills during academic breaks (Alexander et al., 2007; Slade et al., 2017). Third, some may not return once schools re-open: following closures due to the 2014-2016 Ebola crisis, re-enrollment in Sierra Leone was high but imperfect (Kastelic et al., 2015), and gross enrollment fell slightly from 2013 to 2015 (World Bank, 2021).

To keep students engaged and learning during school closures, governments around the world abruptly shifted to distance learning. In low-income countries, the medium with the greatest reach is radio. The Government of Sierra Leone, like many others, announced a national radio teaching programme shortly after schools closed in March 2020. Many governments and partners complemented radio teaching with SMS-based reminders. In Sierra Leone, a number of non-state organizations, including our implementation partner Rising Academy Network, supported the Government to develop content for the radio teaching programme and provided SMS reminders. Missing from these and similar efforts, however, is any direct interaction between teachers and students. Teachers add significant value to student learning (Chetty et al., 2014), but little of that value is explained by teacher characteristics (Bau and Das, 2020), suggesting that effective interactions may drive teacher effects.

In this paper we evaluate the effectiveness of live phone tutorials in increasing engagement with radio lessons, improving learning, and ultimately ensuring that children re-enrol in school upon re-opening. We compare phone tutorials delivered by private-school teachers to those delivered by government-school teachers. We also make a methodological contribution, testing a phone-based learning assessment against the same assessment delivered in-person.

To do this we designed a randomized control trial, in which 4,399 students were randomly assigned to one of three treatment groups. The first group received SMS reminders to listen to educational radio. The second group received SMS reminders and weekly phone tutorials from private school teachers. The third group received SMS reminders and weekly phone tutorials from government school teachers. We also cross-randomise survey mode, with a sub-sample of 500 students assigned to be surveyed and tested in-person rather than by phone.

We find no effect of calls by either private or government teachers on mathematics or language test scores. This is robust to controls for student characteristics and school fixed effects, and differences in survey mode (in-person or phone). We do find some positive effects of tutoring calls on educational engagement by parents and children. Tutoring calls increase an index of student activity by 0.31 σ and an index of parent activity by 0.34 σ. Overall re-enrolment is 99 percent, so we see no effect of calls on re-enrolment.

Random assignment to survey mode was imperfectly implemented. We see a large negative correlation between being surveyed in-person and test scores, but this is not apparent in the randomised allocation or when using randomised assignment as an instrument for actual survey mode. The phone survey has good internal consistency, but we do see differences in item functioning between phone and in-person surveys.

Our results join three other studies that evaluate the effect of phone tutorials during the COVID-19 pandemic (as well as others in process). The first found mathematics learning gains in Botswana of between 0.16 and 0.29 σ for an SMS and live phone call intervention. Students were sent math problems by SMS and then called by NGO staff to work through the problems (Angrist et al., 2020a). The second found no benefits on mathematics performance in Kenya of either short 5-minute ”accountability checks” or 15-minute tutoring calls (Schueler and Rodriguez-Segura, 2021). The third found large effects of 0.56 σ in mathematics and 0.66 σ in literacy in Bangladesh. The intervention was a 30-minute telephone mentoring session with a student volunteer from a local university (Hassan et al., 2021). The intervention that we study is focused on encouraging engagement with radio instruction, following guides to review material covered in radio broadcasts. We measure literacy as well as numeracy, test for differences between government and private school teacher delivery of tutoring calls, and achieve a much lower attrition rate. One potentially important difference between the two studies that found positive effects (Angrist et al. in Botswana and Hassan et al. in Bangladesh) and the two that did not (Schueler and Rodriguez-Segura in Kenya, and this paper on Sierra Leone), is that the former worked with the sub-set of families who opted in to their studies, whereas the latter attempted to work with all the children enrolled in the relevant schools and grades. Treatment effects could be larger for more motivated students (and families) who chose to opt-in to distance learning, than for those who would not have chosen to opt-in.

Other literature considers distance-learning and digital communication from before the current crisis began. For example a radio-based math instruction programme in Nicaragua in the late 1970s increased test scores (Jamison et al., 1981). A qualitative evaluation provided evidence that a radio learning programme in Sierra Leone during the Ebola epidemic helped to keep children connected to education (Barnett et al., 2018). The lack of adult support to children was cited as a key weakness of this program.

SMS reminder messages have proven effective at improving educational outcomes in some contexts. Nudges in Brazil reduced dropout in 2020 (Lichand and Christen, 2021). Weekly SMS messages and monthly quizzes in rural China improved student academic outcomes (Mo et al., 2014). SMS messages and phone calls can also useful for engaging parents during normal times (Barrera et al., 2020; Berlinski et al., 2021; Kraft and Rogers, 2015; Doss et al., 2018). But the size of text message effects tends to be modest relative to the impact of in-person interactions with teachers (Araujo et al., 2016; Bau and Das, 2020).

Our study also adds to the little that has been written about assessing learning by phone. There is experimental evidence that phone surveys on other topics can be reliable (Garlick et al., 2020), and other research from developing countries showing that survey mode (e.g., paper versus computer-assisted) does make a difference for measured outcomes in both education (Singh, 2020) and other sectors (Caeyers et al., 2012). Yet phone-based assessment offers potential for significant cost-savings over in-person learning assessments (Angrist et al., 2020b). For example phone-based assessments trialled in India during the COVID-19 crisis cost USD 3.5 per student, compared with typical in-person costs of around USD 5-13 per student (Joshi and Sudhanshu, 2021).

The rest of this paper is structured as follows; section 2 provides more background about the program context, section 3 discusses the interventions, section 4 outlines the experimental design, section 5 the data, section 6 the results, and section 7 concludes.

2. Background

The COVID-19 crisis affected Sierra Leone much as it did many of the country's neighbors. Sierra Leone recorded a total of 2,611 confirmed cases and 76 deaths in 2020 (Dong et al., 2020). Awareness of COVID-19 was high (Fitzpatrick et al., 2021). The economic impact was severe – small business profits fell by 50 percent between March and June 2020, and average wage earnings fell 20 percent, with increases in household debt and reduced food consumption (Meriggi et al., 2020).

Turning to education, students globally lost an average of two-thirds of an academic year of schooling in 2020 (UNESCO, 2021). In some low-income countries, losing this much education can substantive proportion of children's total lifetime expected schooling (Evans et al., 2021). In Sierra Leone, schools closed on 31st March 2020 until further notice. Primary schools reopened for exam grades (Grade 6) only on 1st July 2020. Schools re-opened for all children for the next academic year on 5th October 2020. A nationally representative survey conducted in early October 2020 found that 91 percent of parents intended to send their children back to in-person school (Cuccaro et al., 2021). A low dropout rate (3 percent) was also found in another third-party survey of a representative sample of students in Rising Academy schools across Sierra Leone, Liberia, and Ghana, between January-March 2021 (Caballero Montoya et al., 2021).

Only around half of children in Sierra Leone were engaged in any educational activity while schools were closed, according to a nationally representative survey conducted in July 2020. Less than half listened to educational radio, spending on average four hours per week listening to radio lessons (from a maximum possible of 7 hours for grades 1-3 or 5 hours for grade 4-6). 99 percent of parents expected their children to return to school, but only around half expected their children to be promoted to the next grade (Foster, 2020). This re-enrolment was later confirmed in a second survey in November/December 2020, which found that 97-99 percent of previously enrolled students had returned. Actual grade promotion rates were higher than expected by parents, at around 75 percent (Foster, 2021).

Learning outcomes were dire even before the crisis. A 2014 national Early Grade Reading Assessment (EGRA) found that 97 percent of children in class 2 could not read (DSTI Media, 2019). Only 83 percent of children complete primary school (World Bank, 2021).

The programme we study was implemented by the non-government education provider Rising Academies in partnership with the Government of Sierra Leone. Rising Academies launched in Sierra Leone in 2014 and provided emergency education to children who were out of school due to school closures during the Ebola epidemic. Rising Academies manages 157 private and government schools in Sierra Leone, Liberia, and Ghana, and works closely with government in Sierra Leone and Liberia. Public schools managed by Rising Academies in Liberia have been shown to be effective (Romero and Sandefur, 2019; Romero et al., 2020). Prior to the pandemic, Rising Academies had been supporting 25 government primary schools since January 2020 as part of the government's Education Innovation Challenge program. Education Innovation Challenge schools are government schools staffed by government teachers, in which one of five non-state operators have been invited to test pedagogical and other innovations with the potential to improve the quality of teaching and learning at scale. The intervention we study took place in these Education Innovation Challenge schools, government schools supported by Rising Academies.

The Sierra Leone Ministry of Education broadcast educational content by radio for all grade levels, from lower primary to secondary. The Ministry has had a radio education unit since the end of the civil war in 2002 as a complement to schools (Alghali et al., 2005; Mangesi, 2007), and it broadcast educational radio during Ebola-related school closures. New radio learning content was developed by government and partners (including Rising Academies) specifically for the COVID-19 school closures to replace in-person instruction. Rising Academies produced Mathematics and English lessons for lower and upper primary. The Radio Teaching Program broadcast Rising's lower primary lessons three days per week on national radio. Rising Academies also broadcast two hours of Mathematics and English lessons for lower and upper primary every day on six local radio stations (Lamba and Reimers, 2020).

Before the pandemic, 73 percent of all households owned a mobile phone, 55 percent owned a radio, 20 percent owned a television, and 5 percent owned a computer (Statistics Sierra Leone (Stats SL) and ICF, 2020). Government therefore did not offer any distance-learning provision by television or online.

Despite these available resources, children may miss out on radio instruction simply through limited attention to the time schedule while at home. To encourage participation, Rising Academies sent SMS reminders to students, including (in principle) all households in all treatment arms of our experiment described below. The phone number listed for each student or guardian received a total of 48 SMS messages over 18 weeks, or an average of 2.7 messages per week. Messages were either simple reminders about the time of the education radio broadcast, an exercise to be completed, or a piece of advice for parents. The SMS number was free to respond to, and students were encouraged to reply with their answers. Messages were addressed from Rising Academies.

3. Intervention

In addition to their other work contributing to the government's national radio programme and sending SMS reminders about these programmes, Rising Academies designed and implemented a tutorial phone call intervention, designed to be complementary to the radio programming, for students from the 25 government schools that they were supporting as part of the Education Innovation Challenge. Rising collected around 5,600 phone numbers of students from these schools in the two days prior to school closures in March 2020. Students were then called so that teachers could recap lessons delivered by radio and answer questions. Interaction is critical to learning, such that there are limits to the overall effectiveness of entirely one-way instruction delivered through mass media such as radio. Delivering actual instruction by phone allows for two-way communication, so teachers can check for the understanding of children and adjust instruction in real-time as necessary.

Several studies have shown that the same intervention can have bigger effects when delivered by an NGO than when delivered by government (Bold et al., 2018; Vivalt, 2020; Kerwin and Thornton, 2021). We therefore test the same intervention delivered by private school teachers employed by the implementer (Rising Academies) and by public school teachers employed by the government. Students from the public schools in our sample were randomised to be either called by private school teachers employed by Rising Academies, or by government school teachers. As the implementer has more direct influence over its own employees, we expect this version to test the potential of the intervention at high fidelity, and the version with government teachers to give greater insight into the potential for scalability.

The interventions began on 25th May 2020 and continued through the end of August 2020. Interventions were initially planned to last for 12 weeks from May to July, and were extended into August for a total of 16 weeks of programming. Educational radio programs were broadcast on national radio and on six local radio stations.

The intervention was delivered by 80 private school teachers and 80 government school teachers. Each teacher was assigned an average of 35 students, and that teacher stayed with the same group of students throughout. Each teacher taught one subject (reading or math) and grade level (upper or lower primary) in the phone tutorials. Teachers did not teach their own usual class.

The private school teachers involved in delivering the intervention had been working for Rising Academies for an average of three years. Government school teachers had been introduced to Rising Academies through the ”Education Innovation Challenge” government partnership programme that started in January 2020. Both the private and public school teachers continued to receive their normal salary whilst schools were closed. They received phone calling credit to cover the cost of calls. In May 2020 government teachers received a pre-agreed 30 percent pay rise, the largest rise in a decade (Murray, 2020). Government school teachers volunteered to participate in the tutoring intervention. The government and private school teachers in our sample earn similar amounts. The private school teachers have an average of 7 years of experience, half as much as the government school teachers with an average of 14 years of experience. The private school teachers are more likely to have university education than the government school teachers (Table A1).

The intervention aimed to deliver a weekly call to each student from each of their two assigned teachers (one focused on math and one on reading). All households from whom phone numbers were gathered were included in the randomization. Each call was expected to last for around fifteen minutes. Teachers identified themselves as teachers, and carried out telephone-based tutorials. These tutorials were consistent with the curriculum of the radio programming. The calls reviewed and recapped the material covered in the radio broadcast, following a detailed guide. Programme monitoring data suggests that private school teachers placed more calls than government school teachers. The average respondent in the private teacher treatment arm received ten out of a maximum possible 16 calls focused on mathematics and nine out of 16 on language. Respondents in the government teacher arm received seven out of 16 planned calls on mathematics and six out of 16 on language. Students may not have received all 16 of the planned calls in part due to difficulties with phone signal, timing of calls, or getting access to a shared family phone. However, the difference in the number of calls received from private school and government school teachers is most likely due to differing incentives facing those employed directly by the implementer.

4. Experimental Design and Econometric Specification

We randomized pupils into a control group or one of the two treatment groups.[1] Randomization was stratified according to baseline test scores and grade of students (where baseline data were available).[2] In the follow-up survey, students were also cross-randomized to either in-person or phone survey tracking. The number of students at each stage is shown in Figure 1 below.

We estimate the following specification: we regress each outcome Y_{is} on an indicator variable for whether the student received any calls (from either public or private), and an indicator for whether the student received calls from a public school teacher in particular.

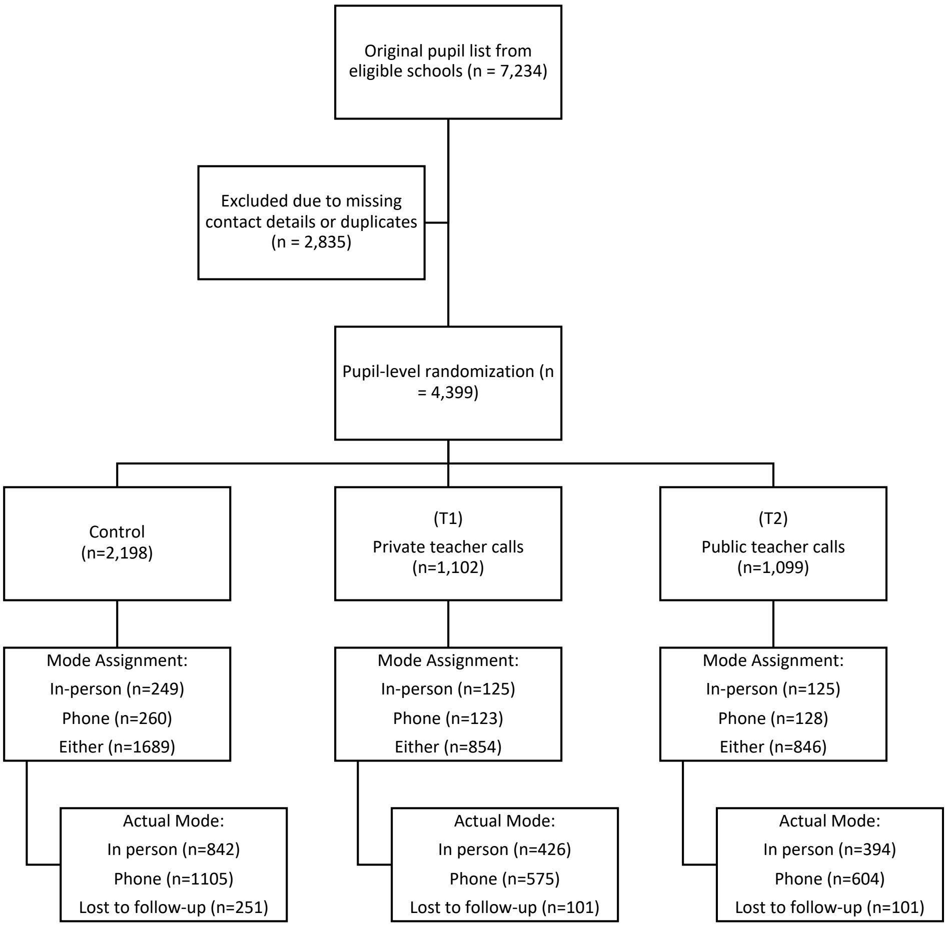
$$Y_{is} = \beta_0 + \beta_1 Calls_i + \beta_2 Public_i + \gamma X_i + Z_s + \epsilon_{is}$$

Our coefficients of interest are β_1 and β_2. We also include student controls X_i and school fixed effects Z_s, and calculate robust standard errors.

[1] We discuss potential ethical issues with the design in Appendix B.

[2] We had intended to assign a fourth group as a pure control group to receive neither SMS reminders nor tutoring calls, however in practice this group also received SMS reminders and so we include them in the SMS only group.

Figure 1: Consort Diagram


Original pupil list from eligible schools (n = 7,234)
Excluded due to missing contact details or duplicates (n = 2,835)
Pupil-level randomization (n = 4,399)
Control (n=2,198)
(T1) Private teacher calls (n=1,102)
(T2) Public teacher calls (n=1,099)
Mode Assignment: In-person (n=249) Phone (n=260) Either (n=1689)
Mode Assignment: In-person (n=125) Phone (n=123) Either (n=854)
Mode Assignment: In-person (n=125) Phone (n=128) Either (n=846)
Actual Mode: In person (n=842) Phone (n=1105) Lost to follow-up (n=251)
Actual Mode: In person (n=426) Phone (n=575) Lost to follow-up (n=101)
Actual Mode: In person (n=394) Phone (n=604) Lost to follow-up (n=101)

5. Data

5.1. Baseline

Prior to school closures, the implementing agency (Rising Academies) collected contact information for 5,566 students from 4,407 households, along with their grade, school, date of birth, and father's name. For a sub-set of 3,034 of these students (in grade 3-6), the implementing agency conducted basic literacy and numeracy assessments adapted from the ASER Centre tools, between 25th February and 20th March 2020.

5.2. Interim Survey

We conducted a short interim independent survey between 10th and 19th September 2020, shortly after the end of treatment at the end of August. This survey targeted a sub-sample of 815 children. Of these, 413 children (51 percent) were able to be tracked. The focus of this interim survey was on time spent engaging with distance learning. We calculate indices of parent and child educational activity. For parents, this index comprises binary indicators for whether they talk about school with their child, read to their child, pay for tutoring, call their teacher, and know the FM frequency for educational radio. For children, this index comprises binary indicators for whether they watched an educational TV, listened to any educational radio, read, were taught by the parent, and spent as much time overall as their parents wanted on educational activity.

5.3. Endline Survey

Schools reopened on 5 October 2020. A key outcome we are interested in is the effect of treatment on re-enrolment. Full enrollment in Sierra Leone typically takes several weeks from the day that school starts, so we began our main survey five weeks after the reopening of school, on Monday (9th November 2020), and ten weeks after the end of the intervention. We collect data on test scores, re-enrolment in school, and time spent on distance learning.

Students were asked to estimate roughly how many minutes they spent per day on all educational activities in a typical week between April and July 2020 while schools were closed. Parents were asked the same question in phone-based surveys.

5.4. Learning Assessment

This paper makes the methodological contribution of assessing the validity of a phone-based learning assessment. There is very little existing literature on best practices for conducting learning assessments by phone (Angrist et al., 2020b). We designed an assessment that could be administered verbally either by phone or in person. We randomise a sub-sample of 500 children to be interviewed in-person, allowing us to directly compare the effect of survey mode on student outcomes.[3] In-person surveys took place at schools. We select a combination of items from Early Grade Reading and Mathematics Assessments (EGRA and EGMA), ASER assessments, and items used orally in in-person assessments in urban India (Banerjee et al., 2017). Parents were told that the questions were not a test of their child's knowledge, with all responses remaining anonymous. Here we discuss sources of validity evidence for our learning assessment across five areas: content, cognition, coherence, correlation, and consequence (Ho, 2020; American Educational Research Association et al., 2018).

1. **Content:** All of the question items from our assessment are relevant for the content of the tuition that students received. Specifically, we selected items that are similar to questions to be asked by teachers in the scripts for the tutorials. In mathematics this includes counting and simple arithmetic, and in English this includes a test of vocabulary, spelling, and aural comprehension.

2. **Cognition:** We piloted our assessment with a small sample of 32 households to confirm that children responded to the questions in the way that we anticipated. Based on the pilot we updated the assessment to include a definition of words that students were asked to spell.

[3] 500 is a less than ideal sample size, but we were constrained by budget and what logistics seemed feasible during the uncertainty of the pandemic. Some evidence suggests that the item-response theory (IRT) models we use can be feasibly estimated with samples as small as 150 (Şahin and Anıl, 2017).

3. **Coherence:** Items in the mathematics and language assessments have a high level of internal reliability in both in-person and phone samples, and higher inter-item correlation in the phone samples (Table A2). This suggests that the questions are measuring the same underlying construct (mathematics and language ability). We construct test score outcomes using item-response theory (IRT) (Das and Zajonc, 2010). This allows us to estimate the underlying unobserved traits of mathematical and language ability, while allowing the difficulty and discrimination of individual question items to vary. This is a more conceptually accurate approach than the more common approach of simply giving the percentage of correct answers, which gives the same weight to questions of different difficulty. The method of aggregating test questions can have large implications for estimated effect sizes (Singh, 2015). IRT also allows us to test whether questions have different difficulty and discrimination across the two survey modes (i.e., Differential Item Functioning or DIF). We first estimate a two parameter logistic model with the 12 mathematics items, and a hybrid partial credit and two parameter model for the 11 language items. We then estimate differential item functioning across the two survey modes with logistic regression (Tables A3 and A4), following Swaminathan and Rogers (1990).[4]. In order to compare test scores between individuals who were surveyed in person or by phone, we then re-estimate the IRT models, only using the subset of items which appear to perform similarly across mode to link scores across the two assessments. There is differential item functioning (either uniform or non-uniform) for 16 of the 23 individual survey questions between the actual in-person and phone-based survey modes, at the 5 percent level (Table A3). Comparing responses by randomized mode assignment, there is differential item functioning for only 8 of the 23 questions (Table A4).

[4] We also show the DIF graphically in Appendix Figures A1 and A2, and the full distribution of test scores by survey mode and assignment in Figure A3.

4. **Correlation:** Our assessments are highly correlated with the baseline in-person ASER assessments administered by the programme implementer. This correlation is not statistically significantly different for those assigned to in-person or phone assessment (Table A5).

5. **Consequence:** Similar assessments to ours have been used in a range of contexts for monitoring school performance. Conducting these assessments by phone holds the potential to substantially reduce the costs of this monitoring, if phone assessments can be shown to be reliable.

5.5. Sample Characteristics and Representativeness

Our study takes place with students from 25 government primary schools in four districts; Western Area Urban (Freetown), Bo, Kailahun, and Kenema. These schools were selected by government to receive pedagogical support from Rising Academies beginning in the 2019/2020 school year as part of the government's "Education Innovation Challenge". Several private providers were competitively selected to support different schools. Rising Academies started supporting these schools in January 2020.

Compared to other schools in the country, those in our sample schools are larger and more likely to be in Freetown, but are by no means elite schools. They have similar levels of basic amenities as other schools nationwide, such as electricity (8 percent), drinking water (64 percent), handwashing facilities (68 percent), and toilets (84 percent) (Table 1). Most students were aged between 7 and 17 (with 1 percent of outliers aged between 3 and 20).

Table 1: All schools' vs. sample schools' characteristics

	All Schools	Our Sample
	Percent	Percent
Feeding Programme	41	76
Recreational Facilities	56	20
Electricity	12	8
Drinking Water	67	64
Handwashing Facilities	66	68
Toilets	74	84
In Freetown	11	36
	Mean	Mean
Years in Operation	27.5	39.4
Total Enrolment	195.3	285.0
N	6,895	25

Note: This table shows descriptive statistics for the schools from our sample and how they compare to all schools nationwide. Data is drawn from the Ministry of Basic and Senior Secondary Education (MBSSE) Education Management Information System (EMIS) 2019 Annual School Census. A map of school locations is shown in the Appendix Figure A4.

5.6. Balance and Attrition

Randomization was stratified by student sex, grade, and baseline test scores. A balance test is shown at Table A6 showing no statistically significant difference in mean values for these variables across treatment groups.

Overall we were able to track 90 percent of students. Just over half of these were surveyed by phone. Data collection was conducted sequentially, first calling all numbers (except the 500 student sub-sample randomly reserved for in-person surveying), before moving to in-person tracking. This allows us to show how the characteristics of those able to be tracked by phone differs to those we could track in person, as well as the characteristics of those we were not able to track at all. None of the treatment arms are statistically significantly correlated with tracking by phone, but students were less likely to be reached by phone if they lived outside of Freetown, and if their parents had not completed any school. This suggests that surveys conducted entirely by phone are likely to under-represent the most marginalised. With regards to overall attrition, students who received tutoring calls were marginally more likely to be tracked. Students in Grade Six were 7 percentage points less likely to be found overall, and in Freetown 7 percentage points less likely (Table A7).

Our response rate compares favourably to purely phone-based assessments (Angrist et al., 2020a; Schueler and Rodriguez-Segura, 2021; Etang and Himelein, 2020), highlighting the importance of in-person tracking to minimise attrition.

6. Results

6.1. Implementation

Administrative data on SMS messaging shows that over 92 percent of phone numbers received all of the planned SMS messages (three per week). Parents in the tutoring call groups were 64 percentage points more likely to report receiving a call from a teacher, and 25 percentage points more likely to report receiving SMS messages. Out of a maximum of 16 potential calls in each subject, students in the private school teacher group received an average of 10 calls in mathematics and 9 calls in language, compared to students in the government school teacher group who received an average of 7 calls in mathematics and 6 in language. Parents reported that calls lasted an average of 22 minutes, and that children spent on average just over one hour per day listening to educational radio. We see a 0.34 σ (standard deviation) effect of receiving calls on the index of parent educational activity, and 0.31 σ on the index of child educational activity (Table 2). We see no statistically significant differences in time spent on learning, retrospectively reported by either parent or child. Of parents who reported that their children spent less time on education than they would have liked (31 percent of parents), the most common reason given for this was "no motivation or interest." The coefficient on calls is equivalent to a 10 percentage point increase in the probability of listening to educational radio at all, though this estimate is only marginally statistically significant at the 10 percent level.

6.2. Outcomes

Almost all (99.7 percent) of respondents in our sample report that their child has re-enrolled in school and attended in the last week, so we do not see any difference in this outcome by treatment status.

There is no effect of tutoring calls on mathematics or language test scores, by either private or government teachers (Table 3). With the upper bound of the 95 percent confidence interval around the estimate of the marginal effect of calls, we can rule out effects larger than 0.07 σ in mathematics and 0.06 σ in language - or 0.12 σ and 0.12 σ using Lee Bounds to bound possible remaining bias due to attrition. Coefficients on other covariates have the expected sign effect and significance - with a 0.1σ effect of baseline test scores.

Looking at individual test question items, we see a small statistically significant effect for just two of the 12 mathematics items (0.03-0.05 σ) and for one of the 11 language items (0.03 σ, see Tables A8 and A9). Results are little changed when aggregating test items using item-response theory estimates or a simple total of correct questions (Table A10). Across sub-groups, we see similarly insignificant results for literacy and mathematics for girls and for boys (Table A11). We also see no statistically significant interactions between treatment and student sex, grade, parent education, or baseline test scores (Table A12). We observe no difference in effects by the intensity of treatment (number of calls actually successfully placed) (Table A13).

Table 2: Implementation and Effects on Time Use

	Effect of Calls (T1/T2)	Marg Effect of Pub. Teach (T2)	Control Mean	Obs.
Outcomes:				
Received any SMS	.25***	-.0798	.608	529
	(.0569)	(.0581)		
Received any calls	.638***	.0445	.215	529
	(.0522)	(.0526)		
Parent Activity:				
Parent Activity Index	.34**	-.207	-.135	529
	(.142)	(.139)		
Talks about school	.093**	-.0371	.785	529
	(.0472)	(.046)		
Reads to child	.0918	-.117	.377	529
	(.0715)	(.0753)		
Pays for tutoring	.00923	-.0467	.54	529
	(.0691)	(.0729)		
Calls teacher	.112**	-.039	.0792	529
	(.0502)	(.0571)		
Knows FM frequency	.038*	.013	.331	1495
	(.0222)	(.0238)		
Child Activity:				
Child Activity Index	.307**	-.188	-.0978	529
	(.139)	(.15)		
Educational TV	.047	-.127**	.192	529
	(.0581)	(.0562)		
Educational Radio	.0969*	.0584	.596	529
	(.0578)	(.0648)		
Reading	.0422	.0189	.691	529
	(.0643)	(.0696)		
Parent teaching	.0231	-.1	.4	529
	(.0709)	(.0739)		
As much as parent would like	.111*	-.103	.672	529
	(.0567)	(.0678)		
Time Spent on Learning:				
Mins/day (Sep Report)	1.69	-1	85.3	529
	(5.23)	(6.01)		
Radio mins/day (Sep Rpt)	5.62	6.04	40.8	529
	(5.14)	(5.69)		
Mins/day (Dec Report)	-2.63	-.228	81.3	2289
	(1.79)	(2.01)		
Mins/day (Dec Rpt, Child)	-2.79*	-.195	83.8	3953
	(1.46)	(1.67)		

Note: Outcome variables are binary unless otherwise indicated. Variables with 529 observations are from the small September 2020 interim survey, others are from the full December 2020/January 2021 endline survey. All regressions include school fixed effects and robust standard errors.

Table 3: Effect of Treatment on Test Scores

	Maths		Language	
	(1)	(2)	(3)	(4)
Effect of calls (T1 or T2)	0.006	-0.003	-0.019	-0.017
	(0.038)	(0.035)	(0.038)	(0.035)
Marginal effect of gov. teachers (T2)	-0.010	-0.006	0.020	0.015
	(0.044)	(0.040)	(0.044)	(0.039)
School fixed effects		Yes		Yes
Controls		Yes		Yes
Lee Bounds				
(Lower)	-0.057	-0.057	-0.124	-0.124
(Upper)	0.118	0.118	0.103	0.103
Observations	3,946	3,946	3,946	3,946

Notes: Robust standard errors in parentheses.

Controls include student age, sex, grade, and baseline test scores.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

6.3. Mode effects

We see substantial mode effects in the OLS specification. Students surveyed in-person scored -0.45σ worse in mathematics than those surveyed by phone. We do not see these mode effects with the randomly assigned intent to be surveyed in-person, but this may be due to low compliance with the randomisation. Of 499 children randomly assigned to be surveyed in-person, 233 were found and surveyed in-person, and 186 were surveyed by phone. An additional 1,429 students originally planned to be surveyed by phone (but who were unreachable) were then found and surveyed in-person at their school. Random assignment to in-person interview increased the probability that an individual was actually interviewed in-person by 11 percentage points. Including control variables in the instrumental variable specification changes the sign of the estimated effect of in-person assessment, but these estimates have large standard errors and are mostly not statistically significant (Table 4).[5] Qualitative evidence suggests that differences between

[5] In Appendix Tables A14, A15, and A16, we show reduced form effects in the in-person and phone sub-samples, and interactions with parent education.

in-person and phone learning assessment scores may be in part explained by students interviewed by phone being helped by their parents, despite the request of interviewers not to do so (Sam-Kpakra, 2021).

Table 4: Effects of Survey Mode on Test Scores (IV)

	Maths (OLS)		IV 1st stage		Maths (IV)	
	(1)	(2)	(3)	(4)	(5)	(6)
Assigned to in-person test			0.151***	0.111***		
			(0.026)	(0.028)		
In-Person (Actual)	-0.506***	-0.447***			0.389	-0.303
	(0.032)	(0.030)			(0.347)	(0.409)
School fixed effects		Yes		Yes		Yes
Controls		Yes		Yes		Yes
F-Stat			34.6	10.2		
Observations	3,946	3,946	3,946	3,946	3,946	3,946
	Language (OLS)		IV 1st stage		Language (IV)	
	(1)	(2)	(3)	(4)	(5)	(6)
Assigned to in-person test			0.151***	0.111***		
			(0.026)	(0.028)		
In-Person (Actual)	-0.323***	-0.219***			0.710**	-0.065
	(0.033)	(0.031)			(0.343)	(0.418)
School fixed effects		Yes		Yes		Yes
Controls		Yes		Yes		Yes
F-Stat			34.6	10.2		
Observations	3,946	3,946	3,946	3,946	3,946	3,946

Notes: Robust standard errors in parentheses.
The outcome in columns (3) and (4) is whether the individual was actually surveyed in person.
Controls include student age, sex, grade, and baseline test scores.

6.4. Cost effectiveness

Our analysis of program cost data follows a format outlined by the World Bank, designed to allow comparability of program costs across countries (Holla, 2019). The average cost of the SMS treatment is $2 per participant, and the average cost of the tutoring call treatment is $40 per participant. This average cost includes phone charges, teacher salaries, and management staff time in the design and oversight of the program, with all cost components disaggregated as far as possible.

6.5. Discussion

To better understand how the programme was perceived by participants, we commissioned a parallel qualitative study with pupils, parents, and teachers. This included a total of 23 focus group discussions with both treatment and control group pupils and parents, at 11 of the 25 programme schools, spread across all four districts (Sam-Kpakra, 2021). It also included 5 interviews with public and private school teachers. Interviews with children found a number of reasons that could explain poor overall performance. Some found the timing of calls challenging. For example, if parents had to work during the day then a child may have either had to try and take the tuition call at a noisy and distracting location such as a marketplace, or take the call in the evening, when they were tired. Some more rural locations had challenges with mobile phone network and electricity supply. In some areas, some pupils reported that teachers only spoke Krio and English, and not the local language (Mende). Some pupils mentioned that they struggled without being able to see their teacher writing on the blackboard as they were used to.

> *"Sometimes when the teacher called my father will not be at home at the moment and he will ask the teacher to call at night and when the teacher calls at night, I won't be able to have total understanding because at that time I had started becoming sleepy, I will just pretend that I understood the*

> *lesson but in actual sense I do not*". (Primary School Pupils, Western Urban District)

The interviews reported by Sam-Kpakra (2021) also offer an explanation for the substantially higher scores for tests conducted by phone rather than in-person. Some pupils and parents openly admitted that they had helped with answers that the child was stuck on.

> *"As for me what I mostly did was when they ask my child and he doesn't know the answer, I push the phone far away and tell him the answer".* (Parent, Kenema District)

Finally, interviews did raise some concerns about possible spill-overs across groups. Several pupils and parents from treated households noted that they invited friends and neighbours to listen to the tuition call together.

> *"My child was not fortunate to be part of the mobile phone teaching program. But fortunately, one of his friends invited him as he was part of the mobile phone teaching program organised by Rising Academy".* (Parent, Kailahun District)

7. Conclusion

In this paper we tested the effect of live tutoring calls from teachers designed to complement distance learning delivered by radio. We find a small positive effect on engagement with education, but no effect on mathematics and language test scores. We also find substantial differences in average student learning assessment results when conducted by phone compared to in-person. We don't see any effect on school re-enrolment, as this is over 99 percent of respondents.

One limitation to this study is the focus on learning outcomes. Another component of the radio programming and SMS reminders was around improving parenting practices designed to improve child well-being, which we did not measure as an outcome.

While most countries around the world have re-opened their schools, surges of COVID-19 cases may lead to further closures, and future adverse events will lead to school closures in individual countries. This study suggests a need for further experimentation in terms of how to help students stay engaged and learning when schools close. Furthermore, our substantial differences across modes of assessments (phone versus in-person) suggest the need for more research if phone-based assessments are to be a viable tool for measuring student learning.

References

Alexander, Karl L, Doris R Entwisle, and Linda Steffel Olson, "Summer learning and its implications: Insights from the Beginning School Study," *New Directions for Youth Development*, 2007, *2007* (114), 11–32.

Alghali, A.M., Edward D.A. Turay, Ekundayo J.D. Thompson, and Joseph B.A. Kandeh, "Environmental Scan on Education in Sierra Leone with Particular Reference to Open and Distance Learning and Information and Communication Technologies," 2005.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, and Psychological Testing (U.S.), *Standards for educational and psychological testing*, American Educational Research Association, Washington, DC, 2018.

Angrist, Noam, Peter Bergman, Caton Brewster, and Moitshepi Matsheng, "Stemming learning loss during the pandemic: A rapid randomized trial of a low-tech intervention in Botswana," *Available at SSRN 3663098*, 2020.

_ , _ , David K. Evans, Susannah Hares, Matthew C. H. Jukes, and Thato Letsomo, "Practical lessons for phone-based assessments of learning," *BMJ Global Health*, July 2020, *5* (7), e003030. Publisher: BMJ Specialist Journals Section: Practice.

Araujo, M. Caridad, Pedro Carneiro, Yyannú Crúz-Aguayo, and Norbert Schady, "Teacher quality and learning outcomes in kindergarten," *Quarterly Journal of Economics*, 2016.

Asiedu, Edward, Dean Karlan, Monica P. Lambon-Quayefio, and Christopher R. Udry, "A Call for Structured Ethics Appendices in Social Science Papers," *Proceedings of the National Academy of Sciences*, July 2021.

Banerjee, Abhijit V., Swati Bhattacharjee, Raghabendra Chattopadhyay, and Alejandro J. Ganimian, "The Untapped Math Skills of Working Children in India: Evidence, Possible Explanations, and Implications," 2017.

Barnett, Sarah, Jetske van Dijk, Abdulai Swaray, Tamba Amara, and Patricia Young, "Making Multisectoral Collaboration Work: Redesigning an education project for child friendly radio: a multisectoral collaboration to promote children's health, education, and human rights after a humanitarian crisis in Sierra Leone," *The BMJ*, 2018, *363.*

Barrera, Oscar, Karen Macours, Patrick Premand, and Renos Vakis, "Texting Parents about Early Child Development: Behavioral Changes and Unintended Social Effects," Technical Report, The World Bank December 2020.

Bau, Natalie and Jishnu Das, "Teacher Value-Added in a Low-Income Country," *AEJ: Economic Policy*, 2020.

Berlinski, Samuel, Matias Busso, Taryn Dinkelman, and Claudia Martínez A., "Reducing Parent-School Information Gaps and Improving Education Outcomes: Evidence from High-Frequency Text Messages," Working Paper 28581, National Bureau of Economic Research March 2021. ZSCC: NoCitationData[s0] Series: Working Paper Series.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur, "Experimental evidence on scaling up education reforms in Kenya," *Journal of Public Economics*, December 2018, *168*, 1–20. 00000.

Caeyers, Bet, Neil Chalmers, and Joachim De Weerdt, "Improving consumption measurement and other survey data through CAPI: Evidence from a randomized experiment," *Journal of Development Economics*, May 2012, *98* (1), 19–33.

Chetty, Raj, John N. Friedman, and Jonah Rockoff, "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, 2014.

Cuccaro, Filippo, Luciana Debenedetti, and Andreas Holzinger, "New RECOVR Survey Findings in Sierra Leone Highlight Socio-Economic Fallout from COVID-19 as New Restrictions Are Put in Place," 2021.

Das, Jishnu and Tristan Zajonc, "India shining and Bharat drowning: Comparing two Indian states to the worldwide distribution in mathematics achievement," *Journal of Development Economics*, July 2010, *92* (2), 175–187.

Dong, Ensheng, Hongru Du, and Lauren Gardner, "An interactive web-based dashboard to track COVID-19 in real time," *The Lancet Infectious Diseases*, May 2020, *20* (5), 533–534. Publisher: Elsevier.

Doss, Christopher J., Erin M. Fahle, Susanna Loeb, and Benjamin N. York, "More than a Nudge: Supporting Kindergarten Parents with Differentiated and Personalized Text Messages," Technical Report NBER Working Paper No. 24450 2018.

DSTI Media, "Sierra Leone invests $1.5 million to bring education innovation to schools for better learning outcomes," 2019.

Etang, Alvin and Kristen Himelein, "Monitoring the Ebola Crisis Using Mobile Phone Surveys," in Johannes Hoogeveen and Utz Pape, eds., *Data Collection in Fragile States: Innovations from Africa and Beyond*, Cham: Springer International Publishing, 2020, pp. 15–31.

Evans, David K., Susannah Hares, Amina Mendez Acosta, and Christelle Saintis, "It's Been a Year Since Schools Started to Close Due to COVID-19," 2021.

Fitzpatrick, Anne E, Sabrin A Beg, Laura C Derksen, Anne Karing, Jason T Kerwin, Adrienne Lucas, Natalia Ordaz Reynoso, and Munir Squires, "Health Knowledge and Non-Pharmaceutical Interventions During the COVID-19 Pandemic in Africa," Working Paper 28316, National Bureau of Economic Research January 2021.

Foster, Elizabeth, "Sierra Leone Covid-19 Impact Monitoring Survey (CIMS): Round 1 Results," Technical Report, World Bank, UNICEF, Statistics SL 2020.

_ , "Sierra Leone Covid-19 Impact Monitoring Survey (CIMS): Round 2 Results," Technical Report, World Bank, UNICEF, Statistics SL 2021.

Garlick, Robert, Kate Orkin, and Simon Quinn, "Call Me Maybe: Experimental Evidence on Frequency and Medium Effects in Microenterprise Surveys," *The World Bank Economic Review*, June 2020, *34* (2), 418–443.

Hassan, Hashibul, Asad Islam, Abu Siddique, and Liang Choon Wang, "Telementoring and homeschooling during school closures: A randomized experiment in rural Bangladesh," 2021, p. 48.

Ho, Andrew, "five Cs of validation: How can we bring current validity theory to practice?," Working Paper, Harvard University, Cambridge, MA 2020.

Holla, Alaka, "Capturing cost data: a first-mile problem," 2019.

Jamison, Dean T, Barbara Searle, Klaus Galda, and Stephen P Heyneman, "Improving elementary mathematics education in Nicaragua: An experimental study of the impact of textbooks and radio on achievement.," *Journal of Educational psychology*, 1981, *73* (4), 556.

Joshi, Pratibha and Sharma Sudhanshu, “Role for Phone Based Assessments in Education Systems,” RISE Programme Blog 2021.

Kastelic, Kristen Himelein, Mauro Testaverde, Abubakarr Turay, and Samuel Turay, “The socio-economic impacts of Ebola in Sierra Leone: results from a high frequency cell phone survey (round three),” Technical Report, The World Bank 2015.

Kerwin, Jason T. and Rebecca L. Thornton, “Making the Grade: The Sensitivity of Education Program Effectiveness to Input Choices and Outcome Measures,” *The Review of Economics and Statistics*, May 2021, *103* (2), 251–264.

Kraft, Matthew A and Todd Rogers, “The underutilized potential of teacher-to-parent communication: Evidence from a field experiment,” *Economics of Education Review*, 2015, *47*, 49–63.

Lamba, Keya and Fernando Reimers, “Sierra Leone and Liberia: Rising Academy Network on air,” Education continuity stories series, OECD Publishing, Paris 2020.

Lichand, Guilherme and Julien Christen, “Behavioral nudges prevent student dropouts in the pandemic,” Working Paper 363, Working Paper 2021.

Mangesi, Kofi, “Survey of ICT and Education in Africa : Sierra Leone Country Report,” 2007.

Meriggi, Niccolo, Macartan Humphreys, Abou Bakarr Kamara, Matthew Krupoff, Madison Levine, Herbert, Mcleod, Mushfiq Mobarak, Wilson Prichard, Ashwini Shridhar, Peter van der Windt, and Maarten Voors, “Tracking the economic consequences and responses to COVID-19 in Sierra Leone,” Working Paper, International Growth Center May 2020.

Mo, Di, Renfu Luo, Chengfang Liu, Huiping Zhang, Linxiu Zhang, Alexis Medina, and Scott Rozelle, “Text messaging and its impacts on the health and education of the poor: evidence from a field experiment in rural China," *World development*, 2014, *64*, 766–780.

Montoya, Erika Caballero, Sokhna Mously Fall, Jeff McManus, and Frida Njogu-Ndongwe, "Challenges and Opportunities as Students Return to School: Evidence from Caregiver and Staff Surveys across Rising Academy Network Schools," 2021.

Murray, Francis H., "Sierra Leone teachers laud salary increment," 2020.

Romero, Mauricio and Justin Sandefur, "Beyond Short-term Learning Gains:," Technical Report CGD Working Paper 521 2019.

_ , _ , and Wayne Aaron Sandholtz, "Outsourcing Education: Experimental Evidence from Liberia," *American Economic Review*, February 2020, *110* (2), 364–400.

Şahin, Alper and Duygu Anıl, "The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory," *Educational Sciences: Theory & Practice*, 2017.

Sam-Kpakra, Robert, "Sierra Leone - Distance Learning During COVID-19: Qualitative Research Report," 2021.

Schueler, Beth E. and Daniel Rodriguez-Segura, "A Cautionary Tale of Tutoring Hard-to-Reach Students in Kenya," July 2021.

Singh, Abhijeet, "How standard is a standard deviation? A cautionary note on using SDs to compare across impact evaluations in education," 2015.

_ , "Myths of Official Measurement:Auditing and Improving Administrative Data in Developing Countries," Technical Report, Research on Improving Systems of Education (RISE) July 2020.

Slade, Timothy S, Benjamin Piper, Zikani Kaunda, Simon King, and Hibatalla Ibrahim, "Is 'summer'reading loss universal? Using ongoing literacy assessment in Malawi to estimate the loss from grade-transition breaks," *Research in Comparative and International Education*, 2017, *12* (4), 461–485.

Statistics Sierra Leone (Stats SL) and ICF, "Sierra Leone Demographic and Health Survey 2019," Technical Report, Stats SL and ICF, Freetown, Sierra Leone, and Rockville, Maryland, USA 2020. ZSCC: NoCitationData[s0].

Swaminathan, Hariharan and H. Jane Rogers, "Detecting Differential Item Functioning Using Logistic Regression Procedures," *Journal of Educational Measurement*, 1990, *27* (4), 361–370. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-3984.1990.tb00754.x.

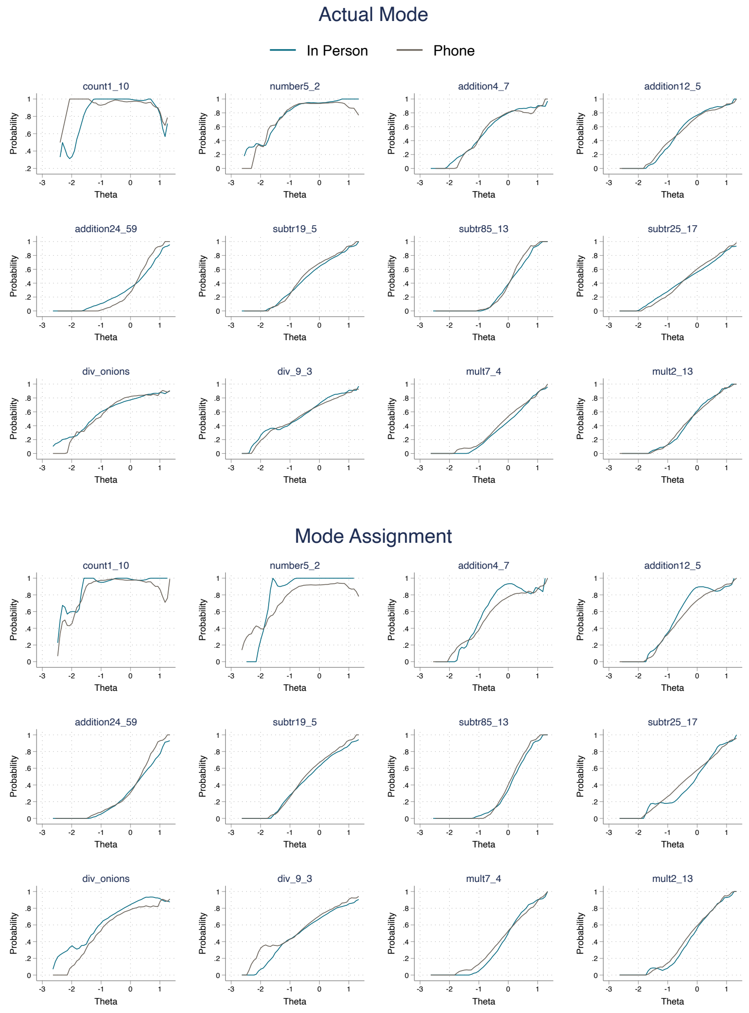
UNESCO, "UNESCO figures show two thirds of an academic year lost on average worldwide due to Covid-19 school closures," 2021.

Vivalt, Eva, "How Much Can We Generalize From Impact Evaluations?," *Journal of the European Economic Association*, December 2020, *18* (6), 3045–3089.

World Bank, "World Development Indicators (World Bank)," 2021.

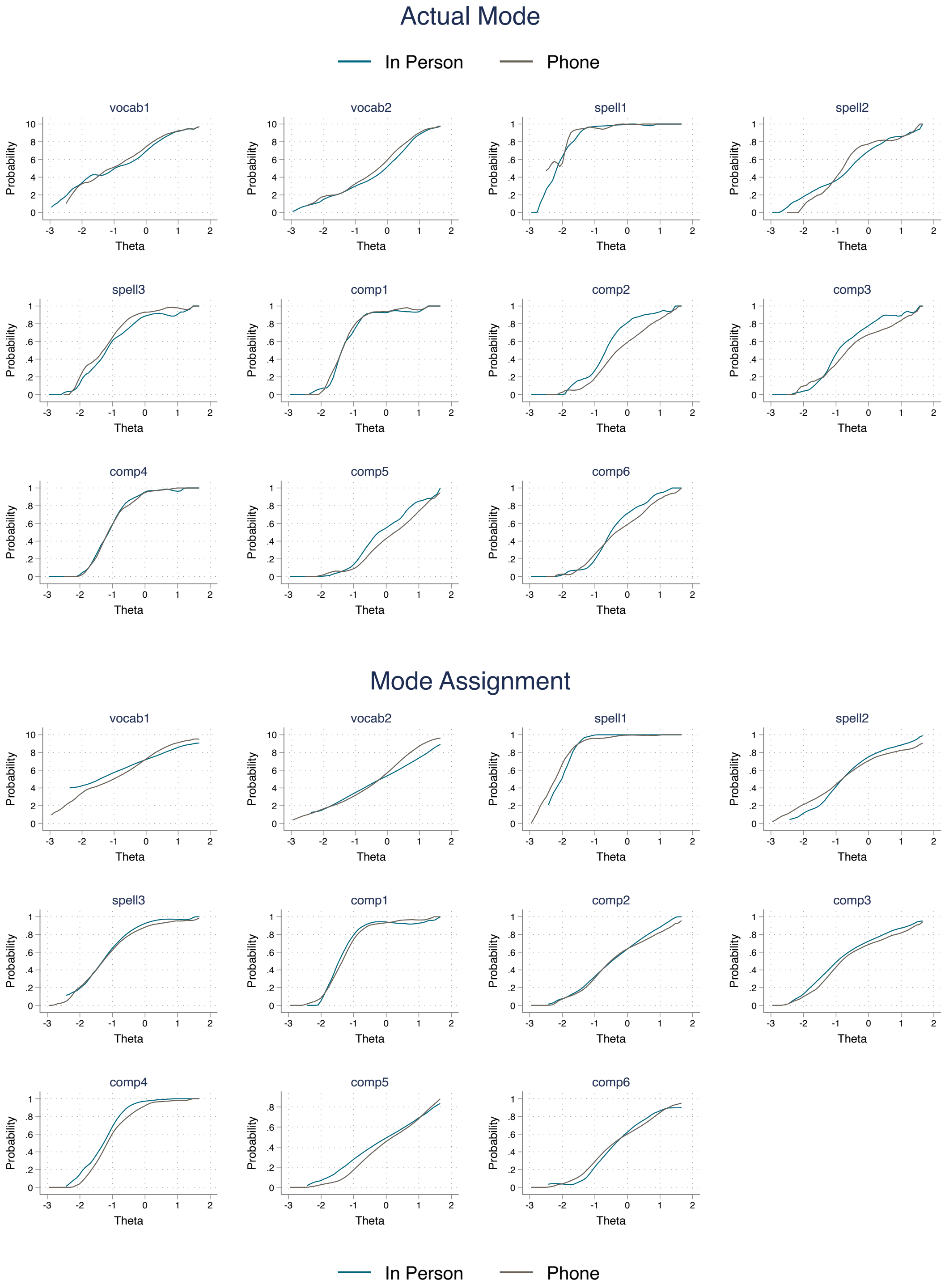
Appendix A Additional Figures and Tables

Figure A1: Differential Item Functioning (DIF) - Mathematics



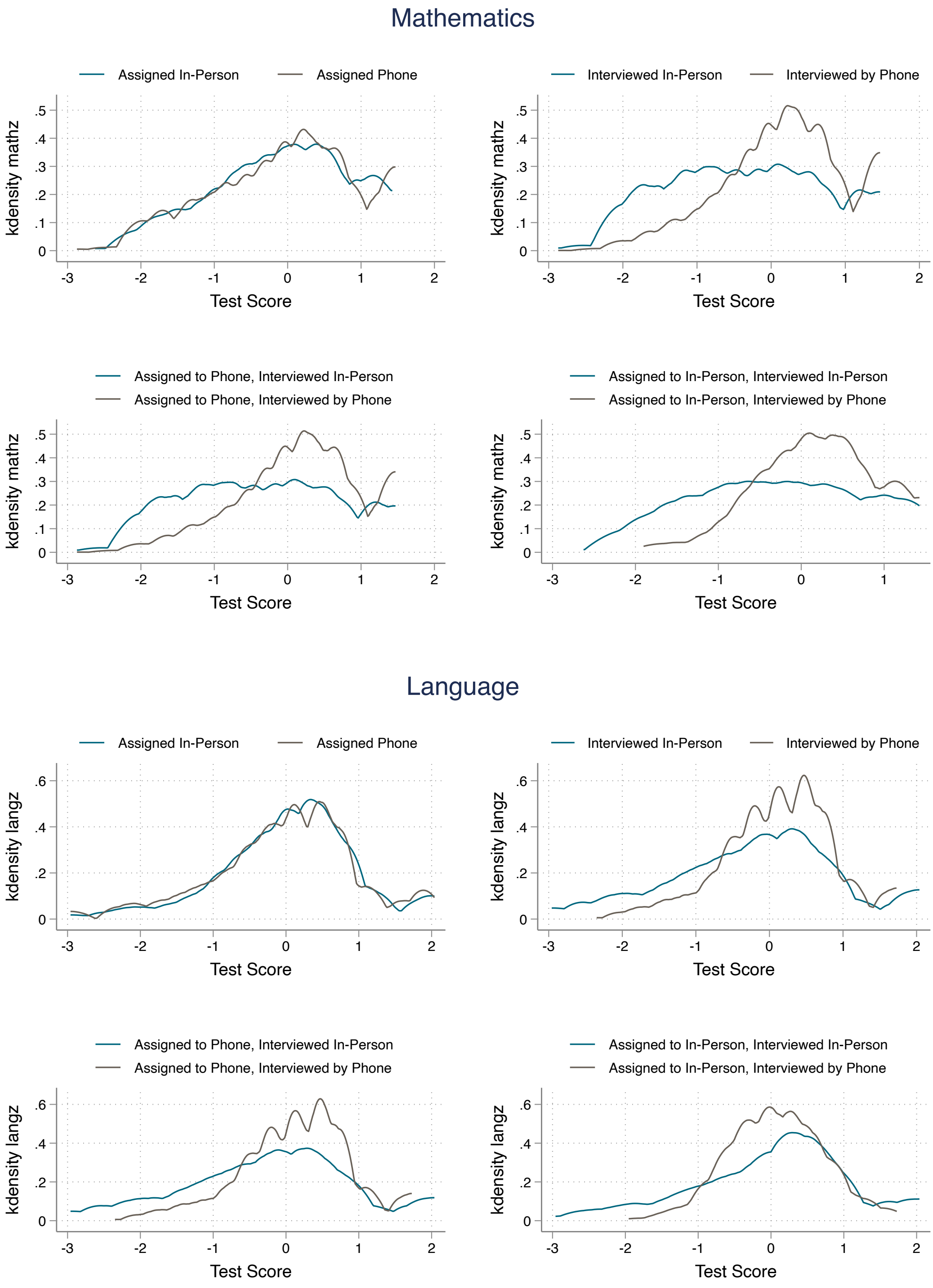
Note: This figures graphs the probability of answering each question correctly, by estimated ability (Theta), by survey mode.

Figure A2: Differential Item Functioning (DIF) - Language



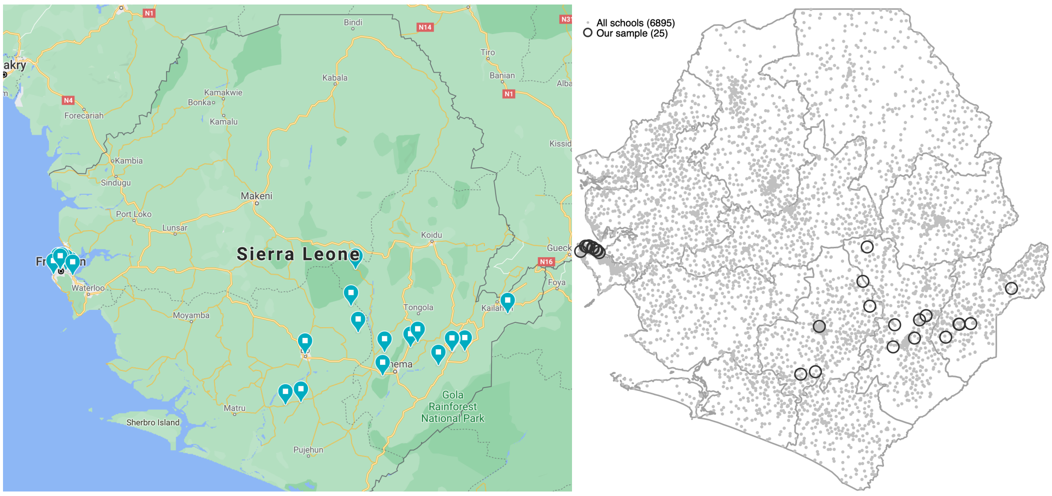
Note: This figures graphs the probability of answering each question correctly, by estimated ability (Theta), by survey mode.

Figure A3: Test score distribution by Survey Mode



Note: This figure shows the distribution of test scores by survey mode.

Figure A4: Maps of Experimental Schools



Note: This figure shows the location of the 25 schools included in our sample. They are located four districts: Western Area Urban (Freetown), Bo, Kailahun, and Kenema.

Table A1: Intervention Teacher Characteristics

	Private	Government
Salary (median, Leones)	1,100,000	1,000,000
Experience (mean years, total)	6.5	14.2
Number of Teachers, by Education Level		
- Secondary School	5	5
- Teaching Certificate	4	33
- Higher Teaching Certificate	21	28
- BSc (Education)	4	1
- BSc (Other)	21	5
- Other	24	7
- Total	79	79

Note: This table shows descriptive statistics for private and government school teachers who delivered the tutoring call intervention, based on a survey conducted by the implementing organisation.

Table A2: Phone assessment reliability

	Alpha	Omega	Correlation	N(Items)	N(Students)
In-person (Maths)	0.86	0.87	0.35	12	1,666
Phone(Maths)	0.93	0.92	0.52	12	2,734
In-person (Language)	0.85	0.86	0.35	11	1,666
Phone(Language)	0.89	0.90	0.43	11	2,734

Note: This table shows Cronbach's Alpha, McDonald's omega, and the inter-item correlation for the sub-samples of the data tested in-person and by phone.

Table A3: Differential Item Functioning: Actual Mode

	Non-Uniform		Uniform	
	Chi2	Prob.	Chi2	Prob.
count1_10	26.73	0.00	0.90	0.34
number5_2	0.01	0.92	0.39	0.53
addition4_7	5.05	0.02	1.78	0.18
addition12_5	0.16	0.69	5.46	0.02
addition24_59	44.08	0.00	49.04	0.00
subtr19_5	8.55	0.00	0.19	0.66
subtr85_13	9.77	0.00	9.49	0.00
subtr25_17	2.01	0.16	1.07	0.30
div_onions	1.79	0.18	6.52	0.01
div_9_3	0.00	0.96	8.90	0.00
mult7_4	11.11	0.00	11.43	0.00
mult2_13	0.63	0.43	3.82	0.05
vocab1	20.95	0.00	27.79	0.00
vocab2	2.61	0.11	0.15	0.70
spell1	2.28	0.13	0.22	0.64
spell2	2.67	0.10	23.73	0.00
spell3	2.83	0.09	29.08	0.00
comp1	0.09	0.76	2.01	0.16
comp2	6.96	0.01	172.57	0.00
comp3	13.05	0.00	40.98	0.00
comp4	0.29	0.59	3.40	0.07
comp5	10.36	0.00	138.78	0.00
comp6	29.98	0.00	45.04	0.00

Table A4: Differential Item Functioning: Assigned Mode

	Non-Uniform		Uniform	
	Chi2	Prob.	Chi2	Prob.
count1_10	0.98	0.32	0.00	0.98
number5_2	6.70	0.01	6.86	0.01
addition4_7	1.49	0.22	3.31	0.07
addition12_5	0.47	0.49	10.71	0.00
addition24_59	0.06	0.81	1.57	0.21
subtr19_5	0.52	0.47	3.62	0.06
subtr85_13	0.47	0.49	6.29	0.01
subtr25_17	0.17	0.68	1.43	0.23
div_onions	1.89	0.17	4.10	0.04
div_9_3	0.74	0.39	6.60	0.01
mult7_4	1.16	0.28	0.96	0.33
mult2_13	0.08	0.77	0.07	0.79
vocab1	1.95	0.16	0.25	0.62
vocab2	0.93	0.34	0.70	0.40
spell1	6.02	0.01	0.01	0.92
spell2	1.07	0.30	3.94	0.05
spell3	0.12	0.72	0.28	0.60
comp1	8.03	0.00	1.28	0.26
comp2	0.61	0.44	0.56	0.45
comp3	0.56	0.45	0.19	0.66
comp4	0.19	0.66	3.04	0.08
comp5	0.17	0.68	17.06	0.00
comp6	2.43	0.12	0.07	0.80

Table A5: Correlation between baseline and endline tests

	(1) Maths	(2) Language	(3) Maths	(4) Language
Assigned to in-person test	-0.077	0.097		
	(0.063)	(0.137)		
Baseline maths	0.235***		0.171***	
	(0.020)		(0.022)	
Baseline maths x In-Person (Assigned)	0.062			
	(0.052)			
Baseline literacy		0.144***		0.079***
		(0.018)		(0.021)
Baseline literacy x In-Person (Assigned)		-0.023		
		(0.049)		
In-Person (Actual)			-0.538***	-0.499***
			(0.034)	(0.083)
Baseline maths x In-Person (Actual)			0.155***	
			(0.036)	
Baseline literacy x In-Person (Actual)				0.157***
				(0.033)
School fixed effects	Yes	Yes	Yes	Yes
Observations	3946	2393	3946	2393
R^2	0.076	0.117	0.132	0.131

Notes: Robust standard errors in parentheses.

* p < 0.10, ** p < 0.05, *** p < 0.01

Table A6: Baseline Balance

	(1)	(2)	(3)	F-test
	Control	T1 Pr Tchr	T2 Gov Tchr	for joint
Variable	Mean/SE	Mean/SE	Mean/SE	orthogonality
Age	11.43	11.37	11.42	0.66
	(0.03)	(0.05)	(0.05)	
Male	0.48	0.48	0.50	0.63
	(0.01)	(0.02)	(0.02)	
Baseline grade	3.54	3.54	3.53	0.97
	(0.04)	(0.05)	(0.05)	
Baseline test score	0.02	0.01	0.00	0.85
	(0.02)	(0.02)	(0.02)	
N	2198	1102	1099	

Notes: P-values reported for F-test, which is estimated with school fixed effects. Standard errors reported in parentheses. Total observations is 3,399.

Table A7: Predictors of Attrition

	Found By Phone		Found At All	
	(1)	(2)	(3)	(4)
Effect of calls (T1 or T2)	0.008	0.012	0.023**	0.022**
	(0.019)	(0.019)	(0.011)	(0.011)
Marginal effect of gov. teachers (T2)	0.031	0.024	-0.000	0.001
	(0.022)	(0.022)	(0.012)	(0.012)
Age		-0.001		0.004
		(0.005)		(0.003)
Assigned to in-person test		-0.227***		-0.034*
		(0.026)		(0.017)
In Grade 6		0.125***		-0.073***
		(0.023)		(0.016)
Freetown		0.140***		-0.064***
		(0.017)		(0.009)
Baseline test score		0.009		-0.008
		(0.011)		(0.007)
Parent: Primary		0.074***		
		(0.021)		
Parent: Secondary		0.065***		
		(0.020)		
Parent: Tertiary		0.081***		
		(0.025)		
Outcome mean	0.58	0.59	0.90	0.90
Observations	3,953	3,888	4,399	4,399

Notes: Robust standard errors in parentheses.

* p < 0.10, ** p < 0.05, *** p < 0.01

Table A8: Effects on individual Maths items

	Counting		Addition			Subtraction			Division		Multiplication	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Effect of calls (T1 or T2)	0.018	0.013	0.046***	0.019	0.007	0.016	0.030*	-0.014	0.015	0.002	-0.004	0.000
	(0.011)	(0.012)	(0.016)	(0.017)	(0.016)	(0.018)	(0.017)	(0.018)	(0.016)	(0.017)	(0.018)	(0.018)
Mrgnl effect gov tchrs (T2)	0.003	0.006	-0.023	-0.024	-0.006	0.014	-0.014	0.018	-0.003	0.012	0.015	0.003
	(0.013)	(0.014)	(0.018)	(0.019)	(0.019)	(0.020)	(0.019)	(0.021)	(0.018)	(0.020)	(0.020)	(0.020)
Assigned to in-person test	0.003	0.004	-0.010	0.008	-0.021	-0.045*	0.004	-0.024	-0.034	-0.014	0.001	0.011
	(0.016)	(0.017)	(0.023)	(0.024)	(0.024)	(0.025)	(0.024)	(0.026)	(0.022)	(0.025)	(0.025)	(0.025)
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,399	4,399	4,399	4,399	4,399	4,399	4,399	4,399	4,399	4,399	4,399	4,399
R^2	0.050	0.040	0.081	0.087	0.094	0.089	0.077	0.073	0.061	0.066	0.091	0.103

Table A9: Effects on individual Language items

	Vocabulary		Spelling			Comprehension					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Effect of calls (T1 or T2)	-0.078	0.030	0.015	0.008	0.031**	0.007	0.023	-0.000	0.023	-0.016	0.009
	(0.087)	(0.096)	(0.012)	(0.017)	(0.015)	(0.015)	(0.018)	(0.017)	(0.015)	(0.017)	(0.018)
Mrgnl effect gov tchrs (T2)	-0.058	-0.163	0.003	0.014	-0.001	0.009	0.012	0.023	0.012	0.045**	0.007
	(0.099)	(0.111)	(0.014)	(0.019)	(0.018)	(0.017)	(0.020)	(0.020)	(0.017)	(0.020)	(0.020)
Assigned to in-person test	-0.150	-0.210	0.008	-0.051**	-0.012	-0.014	0.002	-0.001	-0.004	0.038	0.005
	(0.129)	(0.143)	(0.017)	(0.024)	(0.022)	(0.021)	(0.025)	(0.025)	(0.022)	(0.025)	(0.025)
School FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,953	3,953	4,399	4,399	4,399	4,399	4,399	4,399	4,399	4,399	4,399
R^2	0.159	0.155	0.050	0.101	0.073	0.070	0.082	0.077	0.084	0.086	0.097

Table A10: Effects of treatments on IRT vs. Simple Total Test Scores

	Maths		Language	
	(1) IRT	(2) Total	(3) IRT	(4) Total
Effect of calls (T1 or T2)	-0.003	-0.006	-0.017	-0.014
	(0.035)	(0.035)	(0.035)	(0.034)
Mrgnl effect gov tchrs (T2)	-0.006	-0.009	0.015	0.029
	(0.040)	(0.040)	(0.039)	(0.039)
School FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	3,946	3,946	3,946	3,946
R^2	0.198	0.201	0.201	0.219

Notes: Robust standard errors in parentheses.

Controls include student age, sex, grade, and baseline test scores.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A11: Effect of Treatment on Test Scores, by Sex

	Maths		Language	
	(1) Boys	(2) Girls	(3) Boys	(4) Girls
Effect of calls (T1 or T2)	-0.028	0.019	-0.016	-0.029
	(0.049)	(0.050)	(0.051)	(0.048)
Marginal effect of gov. teachers (T2)	-0.036	0.028	0.021	0.015
	(0.056)	(0.056)	(0.057)	(0.053)
School fixed effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	1916	2029	1916	2029

Notes: Robust standard errors in parentheses.

Controls include student age, grade, and baseline test scores.

* p < 0.10, ** p < 0.05, *** p < 0.01

Table A12: Heterogenous effects on Maths Scores

	(1)	(2)	(3)	(4)
Effect of calls (T1 or T2)	0.035	-0.012	-0.069	-0.006
	(0.040)	(0.044)	(0.072)	(0.029)
Male X Treat	-0.084			
	(0.057)			
Parent Education X Treat		0.010		
		(0.026)		
BL Grade X Treat			0.018	
			(0.017)	
BL Test Scores X Treat				-0.034
				(0.035)
School FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	3,946	3,881	3,946	3,946
R^2	0.198	0.215	0.198	0.198

Notes: Robust standard errors in parentheses.

Controls include student age, sex, grade, and baseline test scores.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A13: Effect of Treatment Intensity (Number of Calls)

	Maths			Language		
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	OLS	IV	OLS	OLS	IV
Number of Maths calls	0.012***	0.004	-0.001			
	(0.003)	(0.003)	(0.005)			
Number of Language calls				0.001	0.002	0.001
				(0.004)	(0.003)	(0.006)
School FE	Yes	Yes		Yes	Yes	
Controls		Yes	Yes		Yes	Yes
Observations	3,946	3,946	3,946	3,946	3,946	3,946
R^2	0.036	0.198	0.171	0.090	0.201	0.126

Notes: In the IV estimates the number of calls is instrumented for by treatment status.

Robust standard errors in parentheses.

Controls include student age, sex, grade, and baseline test scores.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A14: Effects of Survey Mode (Actual) on Maths Scores (ITT)

	Assigned			Actual		
	(1) All	(2) Phone	(3) In-Person	(4) All	(5) Phone	(6) In-Person
Effect of calls (T1 or T2)	-0.013	-0.013	0.055	-0.010	-0.019	0.008
	(0.037)	(0.037)	(0.100)	(0.039)	(0.038)	(0.058)
Mrgnl effect gov tchrs (T2)	0.000	0.001	-0.050	-0.014	-0.003	-0.021
	(0.042)	(0.042)	(0.114)	(0.044)	(0.043)	(0.068)
Assigned to in-person test	-0.073					
	(0.067)					
Calls X In-Person	0.093					
	(0.106)					
Calls (Gov) X In-Person	-0.060					
	(0.120)					
In-Person (Actual)				-0.451***		
				(0.042)		
Calls X In-Person				0.012		
				(0.070)		
Calls (Gov) X In-Person				-0.009		
				(0.082)		
School FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3,946	3,527	419	3,946	2,284	1,662
R^2	0.198	0.194	0.265	0.242	0.226	0.222

Notes: Robust standard errors in parentheses.

Controls include student age, sex, grade, and baseline test scores.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A15: Effects of Survey Mode (Assignment) on Maths Scores, by Parent Education

	(1) None	(2) Primary	(3) Secondary	(4) Tertiary	(5) All
Effect of calls (T1 or T2)	-0.025	-0.001	0.060	-0.056	0.001
	(0.049)	(0.063)	(0.050)	(0.071)	(0.028)
In-Person (Assigned)	-0.072	0.100	-0.242***	0.101	0.132
	(0.093)	(0.119)	(0.080)	(0.099)	(0.085)
Parent Education:					
Primary					0.224***
					(0.041)
Secondary					0.324***
					(0.039)
Tertiary					0.478***
					(0.049)
In-Person (Assigned) × Primary					-0.003
					(0.134)
In-Person (Assigned) × Secondary					-0.481***
					(0.110)
In-Person (Assigned) × Tertiary					-0.117
					(0.120)
School FE	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
Observations	1,307	863	1,162	546	3,881
R^2	0.226	0.218	0.228	0.256	0.221

Notes: Robust standard errors in parentheses.

Controls include student age, sex, grade, and baseline test scores.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A16: Effects of Survey Mode (Actual) on Maths Scores, by Parent Education

	(1) None	(2) Primary	(3) Secondary	(4) Tertiary	(5) All
Effect of calls (T1 or T2)	-0.044	0.008	0.059	-0.057	-0.008
	(0.047)	(0.061)	(0.050)	(0.070)	(0.028)
In-Person (Actual)	-0.474***	-0.398***	-0.302***	-0.196**	-0.512***
	(0.052)	(0.068)	(0.058)	(0.083)	(0.050)
Parent Education:					
Primary					0.154***
					(0.045)
Secondary					0.187***
					(0.042)
Tertiary					0.326***
					(0.050)
In-Person (Actual) × Primary					0.106
					(0.081)
In-Person (Actual) × Secondary					0.143*
					(0.074)
In-Person (Actual) × Tertiary					0.296***
					(0.093)
School FE	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
Observations	1,307	863	1,162	546	3,881
R^2	0.273	0.252	0.241	0.263	0.255

Notes: Robust standard errors in parentheses.

Controls include student age, sex, grade, and baseline test scores.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix B Ethics Discussion

In this section we discuss the research ethics of this project, following the structure proposed in Asiedu et al. (2021).

1. Policy Equipoise and Scarcity: Clinical equipoise means there is genuine and meaningful uncertainty or disagreement amongst stakeholders on the outcome of the research (e.g., the cost-effectiveness of an intervention relative to alternatives). At the time of the design of this intervention there was very little evidence on the effectiveness of distance learning, particularly in low-income countries like Sierra Leone. SMS reminders have been shown to be effective at increasing engagement with a service, but little evidence existed for how reminders to listen to radio school would increase learning. Had the research not been conducted, the counterfactual situation that would have happened instead (i.e., no text messages or calls) would not have been predictably better for participants. Budget considerations limited the numbers of SMS messages that could be sent and calls made.

2. Role of Researchers with Respect to Implementation: The researchers in this study were fully independent of the program implementers. Researchers had no direct decision-making power over the implementation of the program and did not directly provide any component of the program. The role of researchers was limited to the random assignment of participants to treatment and control groups and the supervision of outcome surveys.

3. Potential Harms to Research Participants and Non-Participants from the Interventions or Policies: We consider the risk of potential harm to be very low from both the intervention and the research. The intervention consisted of two SMS messages per week and two 15 minute phone calls per week, all of which were designed to encourage students to engage with radio learning content. Thus the additional time cost to children and parents was low. We see no potential harms to non-participants.

4. Potential Harms from Data Collection (e.g., Surveying, Privacy, Data Management) or Research Protocols (e.g., Random Assignment): The research protocol for this study was approved by the Sierra Leone Research Ethics Board. All participants in the research provided verbal informed consent. Participants were compensated for participating in the research with 10,000 SLL of phone credit (approximately $1 USD). We did not observe any risks or negative outcomes from the data collection process. We see no potential harms to research staff in the conduct of the survey. The number of coronavirus cases and deaths in Sierra Leone was low at the time of data collection beginning (2,366 confirmed cases and 74 deaths). Most data collection was carried out over the phone. In-person data collection, which only applied to a subsample in the endline survey, was conducted in accordance with government guidelines on strict social distancing and hygiene measures.

5. Financial and reputational conflicts of interest: None of the researchers have financial conflicts of interest with regard to the results of the research. None of the researchers have potential reputational conflicts of interest with regards to the results of the research. We should note, however, that the implementing organization (Rising Academies Network, RAN) is a for-profit company with a possible financial interest in the results.

6. Intellectual freedom: There were no contractual limitations on the ability of the researchers to report the results of the study. The research team remained fully independent of RAN, and had no financial relationship with RAN; research funds were raised independently by the research team, and RAN staff were not involved in the analysis or interpretation of data.

7. Feedback to participants or communities: We did not budget for providing post-study feedback on results to participants. However, the results have been shared with the implementing partner (RAN) to inform the design of their programs and with the Government of Sierra Leone.

8. Foreseeable misuse of research results: We see no foreseeable and plausible risk that the results of the research will be misused and/or deliberately misinterpreted by interested parties to the detriment of other interested parties.