Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 26-1661

Measuring Foundational Learning by Phone: Evidence from a Randomized Crossover Experiment

Saloni Gupta
Brown University

Governments need reliable measures of foundational learning, but often rely on reports from the teachers and officials responsible for students' learning, who may face pressure to present favorable outcomes. Independent assessments address this problem, but conducting them in person is costly to repeat across many locations. Phone assessments could supply independent measures if they produce scores comparable to in-person testing. I test this condition using a randomized crossover experiment in Uttar Pradesh, India, in which Grade 2 and 3 children took the same literacy and numeracy tests by phone and in person at home, in randomized order, a median of five days apart. Across 428 children, the pooled difference between phone and in-person scores across the two subjects is a precisely estimated null. The findings support independent phone assessments as a means of collecting aggregate learning data to supplement teacher-reported progress.

VERSION: November 2026

Suggested citation: Gupta, Saloni. (2026). Measuring Foundational Learning by Phone: Evidence from a Randomized Crossover Experiment. (EdWorkingPaper: 26-1661). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/7yma-zm98

Measuring Foundational Learning by Phone: Evidence from a Randomized Crossover Experiment

Saloni Gupta*

November 2026

Abstract

Governments need reliable measures of foundational learning, but often rely on reports from the teachers and officials responsible for students' learning, who may face pressure to present favorable outcomes. Independent assessments address this problem, but conducting them in person is costly to repeat across many locations. Phone assessments could supply independent measures if they produce scores comparable to in-person testing. I test this condition using a randomized crossover experiment in Uttar Pradesh, India, in which Grade 2 and 3 children took the same literacy and numeracy tests by phone and in person at home, in randomized order, a median of five days apart. Across 428 children, the pooled difference between phone and in-person scores across the two subjects is a precisely estimated null. The findings support independent phone assessments as a means of collecting aggregate learning data to supplement teacher-reported progress.

JEL Codes: I21, I25, O15, C93
Keywords: phone assessments, independent monitoring, foundational learning, principal–agent problem, India

*Brown University. Email: saloni_gupta@brown.edu. I thank Alejandro Ganimian and Alex Eble for invaluable advising and feedback throughout this project. Rahul Ahluwalia, Niharika Gupta, Kumar Satyam, Prakhar Singh, Vaani Chopra and Azeem Panjwani at Central Square Foundation were instrumental in the study design and implementation. I gratefully acknowledge Central Square Foundation for institutional support. Central Square Foundation implemented the experiment and collected the data. The analysis was approved by the Teachers College, Columbia University Institutional Review Board under an exemption for secondary research (Protocol 21-421). The experiment was not pre-registered, and the analyses were not pre-specified. A retrospective registration has been submitted to the AEA RCT Registry (https://www.socialscienceregistry.org/trials/7414).

1 Introduction

Primary enrollment in low-income countries has nearly caught up with that in high-income countries, but schooling has not translated into learning. Many children complete primary school without basic literacy and numeracy; in India, fewer than three in ten Grade 3 children in government schools can read a Grade 2 text or solve a subtraction problem (Pritchett, 2013; World Bank, 2018; ASER, 2024). Independent household surveys document these learning gaps, but the data reported by education systems themselves can present a much more favorable picture. When independent proctors re-administered the same questions in two Indian states, the scores that teachers had reported for their own students proved roughly twice the retest scores in Madhya Pradesh and 16 to 20 percentage points higher in Andhra Pradesh (Singh and Berg, 2024; Singh, 2024).[1]

Governments typically obtain learning data from the teachers and officials responsible for students' learning. This creates a principal–agent problem: those responsible for learning also report the outcomes and may face pressure to present favorable results. Such distortions make independent measurement essential as governments adopt explicit learning targets. India's NIPUN Bharat mission sets the goal that every child attains foundational literacy and numeracy by the end of Grade 3. It requires that progress toward this goal be tracked independently for each state, district and block (GOI, 2021). Such tracking can rely on external assessors visiting schools, as in many high-income countries, or on assessments of children at home, as in India's household learning surveys.

Assessments conducted in person, whether in schools or at home, are expensive to repeat and to spread across a district, so they are often confined to a few locations or to samples drawn from a limited set of schools. Phone assessments remove the need to travel to each school or household. They can therefore reach children in almost every school and administrative unit of a district, including remote areas, so that a sample can be drawn to represent each of these units. Repeated in-person visits spread this widely would be prohibitively costly. Household access to phones is wide enough to support such reach: a survey of 12,500 households from low-income settings across ten Indian states reports that 90% own a smartphone (Central Square Foundation, 2026). The gains in reach and cost make phone assessments attractive for independent monitoring, but only if they yield results comparable to those from in-person assessments.

To assess whether phone and in-person assessments yield comparable scores, I conducted a randomized crossover experiment among children in Grades 2 and 3 in Uttar Pradesh,

[1] Score inflation was also documented on a diagnostic assessment with no formal consequences for students or teachers (Singh and Berg, 2024).

India. I compare phone and in-person literacy and numeracy scores for 428 children who took the same test at home in both modes, in randomized order. The assessments typically took place within a week of each other. Randomizing assessment order separates the mode difference from changes between assessments, such as learning or familiarity with the test.

I report four main findings. First, average phone and in-person scores are similar: the pooled difference across literacy and numeracy is 0.007 standard deviations (SD), estimated precisely enough to rule out average differences larger than approximately 0.08 SD in either direction. Subject-specific estimates are small and have opposite signs. The estimated mode difference does not vary detectably by the child's gender, grade or age, by the mother's education or by household income.

Second, similar average scores coexist with substantial variation in individual children's scores across assessments: phone scores correlate with the in-person benchmark at 0.57 in literacy and 0.50 in numeracy. Agreement in average scores therefore does not establish that phone assessments are suitable for diagnosing individual children's learning needs or planning individualized instruction.

Third, prompting by household members is reported more frequently by phone than in person (32 versus 20 percent), but is not clearly associated with higher scores. Excluding children with reported prompting leaves the estimated average mode difference small.

Fourth, the estimated mode difference in literacy is small whether children read from a school textbook at home or from material sent through WhatsApp or SMS. The findings support both approaches to supplying reading material for phone assessments.

These findings contribute to an emerging literature on the validity of phone-based assessment of learning. In prior comparisons, the phone and in-person assessments differ not only in mode but also in assessment order, test content, timing or setting, which makes it difficult to isolate the effect of phone administration. Angrist et al. (2026) validated the outcome measure of a tutoring trial by assessing the same Kenyan children on identical numeracy items, first by phone and then in person, and found a difference of 0.065 SD. Because every child took the phone test first, the comparison also reflects any change between assessments, such as increased familiarity with the test. In Sierra Leone, Crawfurd et al. (2023) compared mathematics and verbally administered vocabulary, spelling and listening-comprehension scores across modes. Children actually assessed by phone scored 0.43 SD higher in mathematics than those assessed in person, whereas the comparison by randomly assigned mode showed no statistically detectable difference. Assignment did not consistently determine administration: children unreachable by phone were assessed at school, while some assigned to in-person assessment were reached by phone. The authors therefore regard the mode comparison as inconclusive.

Rodriguez-Segura and Schueler (2022) examined both literacy and numeracy in Kenya. In both subjects, phone scores predicted individual school examination scores much less closely than a repeated in-person examination did. The numeracy benchmarks differed in content and were administered 10–14 months earlier; the literacy benchmarks were administered 1–2 months earlier. Sobers et al. (2023) obtained phone–in-person correlations of 0.57–0.80 across reading tasks and 0.76–0.83 for passage comprehension among the same children in Côte d'Ivoire. Their phone assessments took place at school, where a research assistant supplied the handset, after a home pilot encountered prompting by family members. Phone testing took place about a year after the first in-person assessment and five to six months before the second. The study examined individual agreement under supervised administration rather than average differences between modes. Phone scores also vary with the assessor: Rodriguez-Segura and Schueler (2023) randomly assigned assessors to more than 2,500 Kenyan children and found enumerator effects large enough to alter impact estimates.

This study extends the evidence on phone-based assessment to a randomized crossover comparison of literacy and numeracy, using identical tests for the same children at home, with assessments close in time. For literacy, it estimates average mode differences when reading material is supplied through school textbooks or digital messages. For both subjects, it examines agreement between individual scores under household conditions. Enumerators recorded prompting in both modes, so its frequency can be compared across modes and the estimates recomputed for children with no reported prompting.

Governments have already begun to assess foundational skills through phones. Several Indian states, including Telangana, Karnataka and Uttar Pradesh, now administer such assessments through WhatsApp chatbots, assessment calls embedded in phone tutoring, or teacher-administered applications that score oral reading with speech recognition.[2] Phone assessment also serves research: its use expanded during pandemic school closures, and it now provides outcome measures for trials of remote instruction and parenting support (Angrist et al., 2022, 2026; Hernández-Agramonte et al., 2024). The study provides a proof of concept for using independent phone assessments to supplement teacher-reported progress

[2]Telangana's School Education Department administers weekly curriculum-aligned assessments for Classes 1 to 10 through a WhatsApp chatbot (Samagra Shiksha Telangana, 2023). Karnataka is scaling Ganitha Ganaka, in which government teachers assess and tutor children by phone in six-week cycles (Angrist et al., 2026). Uttar Pradesh's NIPUN Lakshya application, used by teachers and mentors for spot assessments in Grades 1 to 3, scores oral reading with a speech-recognition model (Government of Uttar Pradesh, Department of Basic Education, 2026). Outside India, a government-partnered initiative in Botswana assessed primary-school children by phone during school closures (Angrist et al., 2020), and an SMS platform in Kenya delivers lessons and quizzes to primary-school children and reports their proficiency to teachers (Jordan et al., 2025).

with aggregate measures of foundational learning across districts or other administrative units.

2 Study Design and Data

2.1 Sample and experimental design

The experiment was conducted in the Lucknow and Varanasi districts of Uttar Pradesh among children enrolled in Grades 2 and 3 of government primary schools. Families were sampled from a database provided by the state government and recruited in four successive cohorts between September 2021 and January 2022; each cohort was screened, assigned and assessed as a group.[3] Screening calls established eligibility and obtained parental consent, after which 1,505 children were randomly assigned an assessment sequence. Figure 1 traces recruitment and participation through to the analysis samples.

Each child was scheduled for two assessments at home, one by phone and one in person, with the same instrument administered on both occasions, and I randomly assigned the order in which a child received them. Within each recruitment cohort, I grouped children by school cluster and grade and randomly assigned half of each group to begin by phone and the other half to begin in person. Siblings in the same household were assigned independently of one another. Separate teams of assessors, both trained and supervised by Central Square Foundation, conducted the phone and the in-person assessments. Parents were informed that participation would carry no benefits or penalties, and they were requested not to assist their children in answering. Among the children who completed both assessments, the median interval between the two was five days, with an interquartile range of three to eight days, so the two scores were obtained close in time.

Recruitment by phone also determined the study population, which comprises children in households that could be reached by phone and agreed to participate. Forty percent of the listed numbers were not reached even after two calls, and 19 percent of the households reached declined. Screening calls failed to connect or went unanswered for two main reasons: many phone numbers in the state database were no longer in service or had been recorded incorrectly, and many households did not answer repeated calls from an unknown number, a common response in India, where unsolicited calls are among the most frequent in the world (Truecaller, 2021). These losses at screening narrow the randomized sample to reachable,

[3]Primary schools in Uttar Pradesh reopened for in-person classes on 1 September 2021, after the pandemic closures. Assessments in early January 2022 coincided with a renewed school closure. All assessments took place at children's homes.

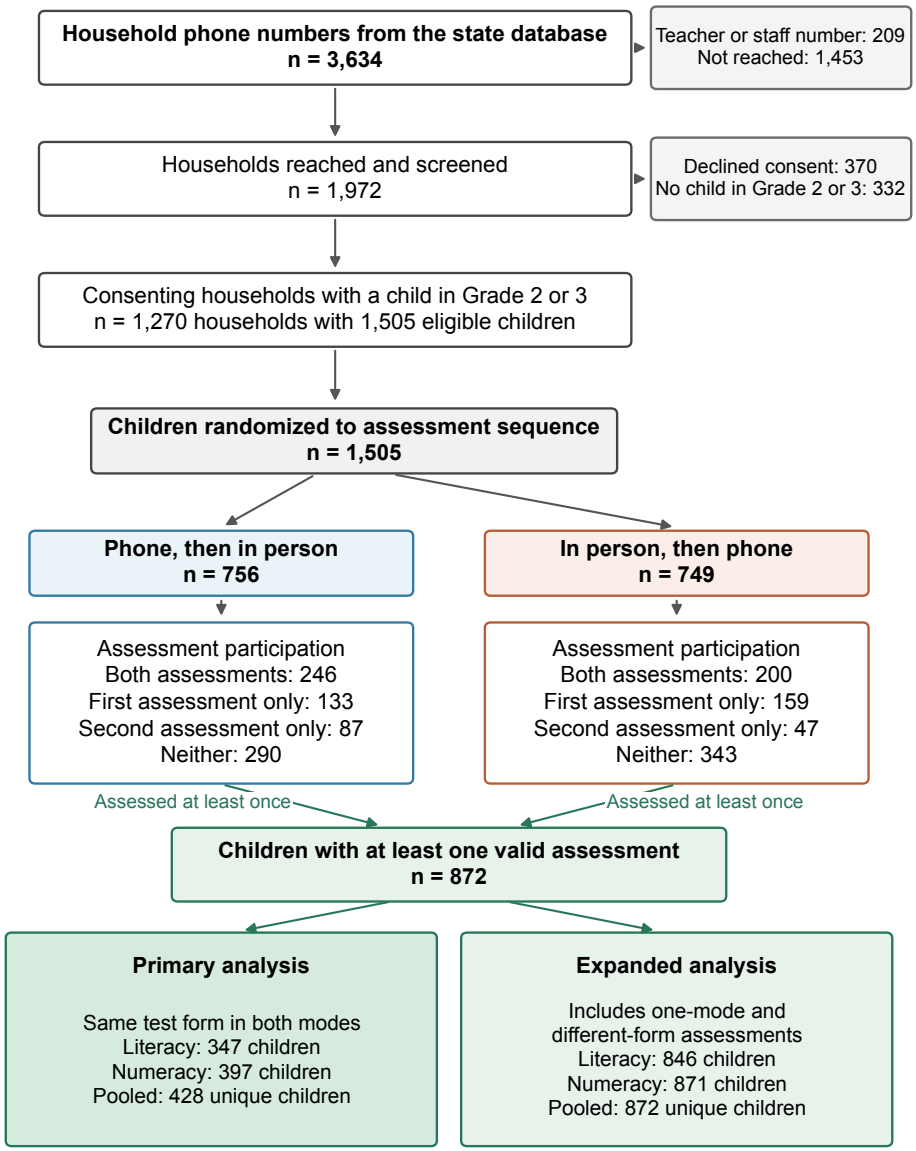


Figure 1: Screening, Assessment Participation and the Analysis Samples

Notes: Counts through the screening stages are household phone numbers from the state database. Counts from randomization onward are children, and a household with more than one eligible child contributes each child. A number was classified as not reached when two calls placed a few days apart did not connect, were not answered, or found the phone switched off or the number out of service. The first and second assessments are those scheduled in the assigned sequence, whether or not the child completed the first. A child is counted as assessed when a valid score was recorded in at least one subject at that assessment. The primary sample requires a valid score on the same test form in both modes. A child is missing from a subject sample for one of two reasons: a valid score is absent at one of the two assessments, which occurs almost only in literacy, where it reflects the lack of usable reading material at one assessment; or the two assessments used different forms, following a change in the recorded textbook or grade between them. The expanded sample comprises every valid assessment, including those of children who received different forms and of children assessed in one mode only. Assessments are subject assessments: one contact can supply both a literacy and a numeracy score. Subject samples overlap. The pooled count includes each child once, whether contributing literacy, numeracy or both.

consenting households; subsequent nonparticipation among randomized children is examined in Section 2.3.

2.2 Assessment instruments and protocols

The assessments measured foundational literacy and numeracy. Foundational literacy is the ability to read and understand a simple text, and foundational numeracy the ability to read and write numbers and to perform basic operations (UNESCO Institute for Statistics et al., 2020). Each subject had two test forms. The literacy tasks comprised listening and picture comprehension, letter and word recognition, and the reading of sentences and a passage. The two literacy forms were based on Grade 1 or Grade 2 textbook material, generally one grade below the children's enrolled grades. The numeracy tasks comprised number comparison, identifying the numbers before and after a given number, ordering, arithmetic and word problems. The two numeracy forms were designed for Grades 2 and 3, and each child received the form for the grade recorded at the assessment; numeracy was administered orally in both modes and required no material. The literacy forms share six items, and the numeracy forms share nine.[4] Appendix Table A4 reports the number of items and the sample size for each form.

In the phone assessments, children completed the reading tasks using either a school textbook at home or material sent to the household's phone. The textbook protocol was used for the first two recruitment cohorts and the messaging protocol for the last two. Under the textbook protocol, the assessor asked the child to open the school textbook to specified pages, and under the messaging protocol, the assessor sent the reading material and questions as WhatsApp images or SMS text. In both protocols the assessor asked the questions and recorded the child's answers during a voice call, and the corresponding in-person assessments used the same textbook pages or printed sheets.[5] The messaging protocol addressed the constraint that the required textbook was not always available at home.

[4]The literacy forms contain 26 and 29 scored items. Each letter and word the child is asked to read is scored as one item, and each sentence and the passage are scored as one item each, credited when the child reads more than half of the words correctly. The shared items are four listening-comprehension questions and two letters. Each numeracy form contains 14 questions, each scored as one item. Appendix Table A1 lists every item.

[5]Under the textbook protocol, the form followed the textbook available at home, irrespective of the child's enrolled grade; households more often had the previous year's textbook than the current one. Under the messaging protocol, the form followed the enrolled grade: Grade 2 children generally received Grade 1 material and Grade 3 children Grade 2 material. A change in the recorded textbook or grade between the two assessments therefore produced different forms at the two assessments. Among the 719 children contacted under the textbook protocol, 142 lacked a usable textbook at one or both assessments, because no textbook was available or the available book did not contain the required reading material. Enumerator notes describe, for example, children with a workbook instead of the textbook.

2.3 Participation and analysis samples

Of the 1,505 children assigned a sequence, 872 completed at least one valid assessment, 446 in both modes and 426 in one mode only (Figure 1). Assessments without usable reading material or recorded answers are treated as not administered (Appendix Table A3).

After assignment, most nonparticipation resulted from unsuccessful contact with households; other households declined, and some assessments were interrupted before completion. Screening characteristics are available for 1,192 of the 1,505 children.[6] For assessed children, the participation comparisons also use household characteristics reported during the assessments. Children assessed in both modes and children assessed in one mode are similar in gender, age, grade and location and differ slightly in household size and mother's schooling (Appendix Table A3). Section 5 examines sensitivity to selective participation by reweighting assessments using characteristics recorded at screening.

The primary analysis compares children assessed on the same test form by phone and in person, keeping test content fixed. It includes 347 children in literacy and 397 in numeracy. The pooled analysis includes 428 children with a same-form comparison in at least one subject; children assessed on the same form in both subjects contribute to both subject comparisons. Appendix Table A2 describes their characteristics.

To check whether the main finding depends on the restriction to same-form comparisons, the expanded analysis also includes children assessed in only one mode or on different test forms across modes. It includes 846 children in literacy and 871 in numeracy, with 872 children contributing to the pooled analysis (Appendix Table A6). Section 3 describes how the forms are linked and how each sample is analyzed.

3 Empirical Strategy

In the primary sample, I compare phone and in-person scores using child-by-test fixed effects and test-specific second-assessment indicators. For child i, test g and scheduled assessment t, I estimate

$$Y_{igt} = \alpha_{ig} + \beta\,\mathrm{Phone}_{it} + \gamma_g\,\mathrm{Second}_{it} + \varepsilon_{igt}. \tag{1}$$

Y_{igt} is the percentage-correct score, standardized using the in-person mean and SD for that test in the primary sample. Phone_{it} indicates phone administration, and Second_{it} indicates the second scheduled assessment. Child-by-test fixed effects absorb stable differences between

[6] For the first 313 children recruited, screening recorded eligibility and consent only. Subsequent screening calls also recorded children's age, gender, grade and textbook availability and the number of eligible children in the household, allowing observed characteristics to be compared between children who completed assessments and those who did not.

children, while the second-assessment indicators allow scores to change between sittings. The coefficient β estimates the average phone–in-person difference. I estimate separate literacy and numeracy effects and a pooled effect across all four test forms.

Because the expanded sample includes children assessed on different forms or in only one mode, I place the different test forms on a common scale using a two-parameter logistic (2PL) item response model, linking forms through their shared items and allowing items to differ in difficulty and discrimination. The resulting scores are standardized using the in-person mean and SD within each subject. For subject s, I estimate

$$\begin{aligned} Z_{ist} = \alpha_s + \beta\,\mathrm{Phone}_{it} + \gamma_s\,\mathrm{Second}_{it} + \delta_s\,\mathrm{PhoneFirst}_i \\ + X_i'\theta_s + \lambda_{s,c(i)} + u_{ist}. \end{aligned} \tag{2}$$

Z_{ist} is the standardized linked score, and $\mathrm{PhoneFirst}_i$ indicates assignment to phone assessment first. Assigned sequence is absorbed by the child-by-test fixed effects in equation (1); it enters equation (2) explicitly because the expanded specification omits these fixed effects. X_i contains the child's gender, age, grade and the number of eligible children in the household, as recorded at screening, together with missing-value indicators. The fixed effects $\lambda_{s,c(i)}$ identify groups defined by both recruitment cohort and block, with a separate category for the initial cohort without screening records. In the pooled regression, all coefficients other than β vary by subject.

Unlike the primary analysis, the expanded analysis also compares different children. It therefore adjusts for recorded differences between children assessed by phone and those assessed in person. Both analyses assume that the second-assessment indicators capture any effect of having taken the test before and that this effect does not depend on which mode came first. Section 5 examines sensitivity to selective participation.

Standard errors are clustered by child in both specifications; the expanded analysis uses a child-level bootstrap that also accounts for the estimation of the linked scores. I first examine average score differences, then compare skill profiles and individual scores, before examining household prompting and differences across children.

4 Results

Average score differences. Average scores are similar across assessment modes in the same-form paired sample. Table 1, Panel A reports a pooled phone effect of 0.007 SD, with a 95% confidence interval of $[-0.07, 0.08]$. The subject estimates are 0.069 SD in literacy (95%

Table 1: Effect of Phone Assessment on Test Scores

Panel A. Main specification: same form in both modes

	In-person score		Phone effect			
	Mean	SD	Estimate	(SE)	95% CI	Children
Literacy	75.8	26.8	+0.069	(0.047)	[−0.02, +0.16]	347
Numeracy	67.5	28.0	−0.046	(0.047)	[−0.14, +0.05]	397
Pooled	71.4	27.7	+0.007	(0.039)	[−0.07, +0.08]	428

Panel B. Expanded specification: all valid assessments, linked 2PL scores

	Estimate	(SE)	95% CI	Children
Literacy	+0.056	(0.044)	[−0.03, +0.14]	846
Numeracy	−0.063	(0.044)	[−0.15, +0.02]	871
Pooled	−0.005	(0.038)	[−0.08, +0.07]	872

Notes: Panel A uses the primary sample of 428 children with a same-form comparison in at least one subject. Scores are percentages correct, standardized using the in-person mean and SD within each form; descriptive means and SDs are in percentage points. Regressions include child-by-test fixed effects and test-specific second-assessment indicators. Standard errors are clustered by child.

Panel B uses all valid subject assessments, including different-form pairs and children assessed in one mode only. Shared items link forms through a two-parameter logistic item response model; scores are standardized using all in-person assessments within subject. Regressions control for assessment period, assigned sequence, screening characteristics and recruitment-cohort-by-block indicators, with subject-specific adjustments when pooled and no child fixed effects. Standard errors and percentile confidence intervals use a child-level bootstrap with 999 replications, repeating item calibration, scoring, standardization and regression. All confidence intervals are 95%. One contact can contribute both subject assessments; child counts across rows are not additive.

CI $[-0.02, 0.16]$) and −0.046 SD in numeracy (95% CI $[-0.14, 0.05]$). The subject-specific estimates are less precise than the pooled estimate and have opposite signs.

Results are similar in the expanded analysis using linked 2PL scores and all valid assessments (Table 1, Panel B). The estimated phone effects are 0.056 SD in literacy (95% CI $[-0.03, 0.14]$), −0.063 SD in numeracy (95% CI $[-0.15, 0.02]$) and −0.005 SD pooled (95% CI $[-0.08, 0.07]$). Both specifications yield pooled estimates close to zero, with confidence intervals ruling out average differences larger than approximately 0.08 SD in either direction.

Both ways of supplying reading material produce small estimated literacy differences. The phone effect is 0.065 SD when material is sent through WhatsApp or SMS (95% CI $[-0.03, 0.16]$), compared with 0.098 SD when children use school textbooks (95% CI $[-0.04, 0.24]$). The estimated phone effect under messaging is 0.032 SD lower than under textbook delivery (95% CI $[-0.20, 0.14]$; Table 2, Panel A). Children were not randomly assigned to the delivery methods, so the 0.032 SD difference between the protocol estimates may also reflect differences between the children assessed, whereas each protocol's own mode comparison uses randomized assessment order.

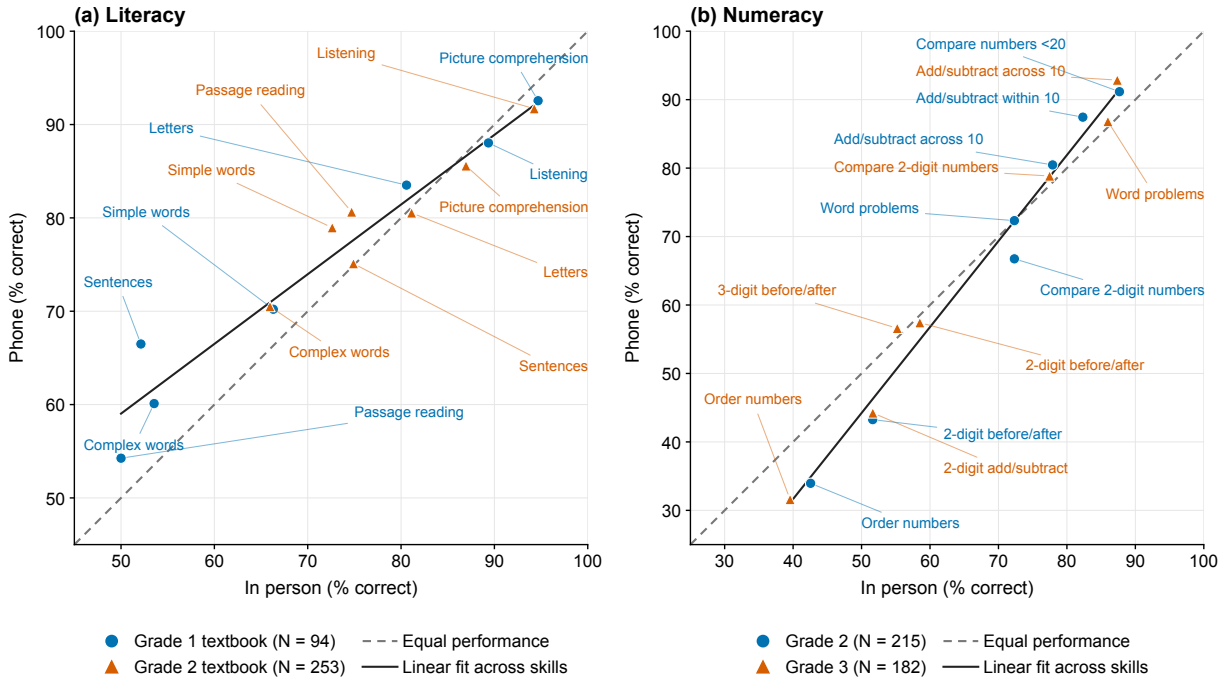


Figure 2: Percent Correct by Skill, Phone and In-Person Assessment

Notes: Each point is one skill on one test: the mean percent correct by phone (vertical axis) and in person (horizontal axis), using the primary sample for that test. Means are not adjusted for the order of assessment.

Skill profiles and internal consistency. Skill profiles broadly align across modes: skills with higher average in-person scores generally also have higher average phone scores (Figure 2). Among children with the same score on the remaining items of the form, several reading items and two addition and subtraction items are answered correctly more often by phone, while items on number comparison, numbers before and after, and ordering are answered correctly more often in person (Appendix Figure A1).

Score dispersion and internal consistency are broadly similar across modes (Appendix Table A4). Cronbach's alpha ranges from 0.85 to 0.95. Tests detect differences in dispersion and internal consistency only for Grade 3 numeracy, where both are lower by phone.

Individual score agreement. Phone and in-person scores correlate at 0.57 in literacy and 0.50 in numeracy (Appendix Table A4). Individual agreement is therefore moderate even when the same children take identical test forms at home, usually within a week of each other.

Household prompting. At the end of each assessment in both modes, the assessor recorded whether anyone had helped the child answer; during phone assessments, asses-

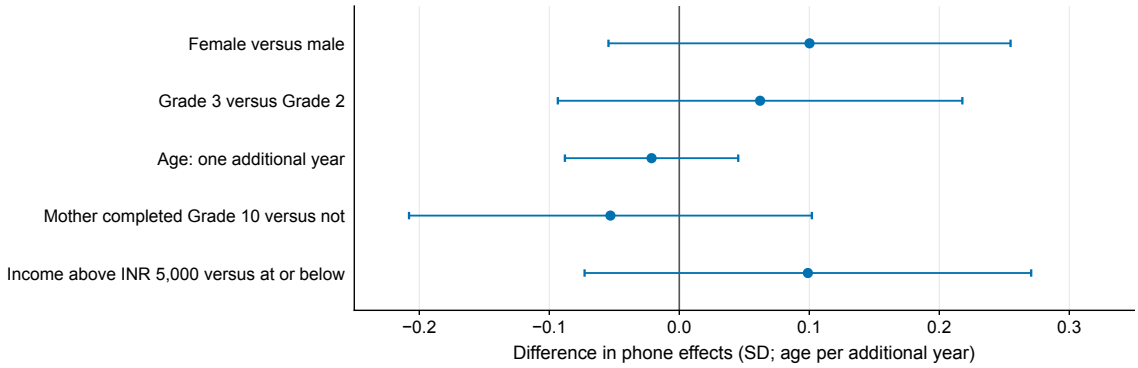


Figure 3: Heterogeneity of the Phone Effect

Notes: Sample: the primary same-form paired sample. Points show differences in estimated phone effects, with 95% confidence intervals based on standard errors clustered by child. Female and Grade 3 compare with male and Grade 2 children. Mother's education compares mothers who completed Grade 10 with those who did not. Income compares households reporting more than INR 5,000 per month with those reporting INR 5,000 or less; households that did not disclose income are excluded from that comparison. Age is continuous; its coefficient is the change per additional year. Each regression allows both the phone effect and test-specific second-assessment effects to vary with the characteristic. A joint test of the five interactions gives $p = 0.84$. The gender, grade and age comparisons use 428 children; maternal education uses 427 and income uses 352. The joint test uses 351 children with all five characteristics observed.

sors inferred prompting from cues such as another person's voice or prolonged pauses before the child answered. Among the 364 primary-sample children with reports at both assessments, suspected prompting was recorded more frequently by phone than in person: 32.4% versus 20.3% (Table 2, Panel B). However, the data do not show that reported prompting systematically raised scores. Within-child comparisons controlling for assessment mode and period yield an association between reported prompting and pooled scores of −0.028 SD (95% CI: $[-0.19, 0.13]$).

Restricting the comparison to children without prompting reported at either assessment leaves a small estimated average mode difference: −0.061 SD (95% CI: $[-0.15, 0.03]$; Panel C). These comparisons rely on assessor reports and cannot capture undetected prompting, such as silent visual cues on reading tasks.

Differences by child and household characteristics. Finally, the phone effect does not differ detectably by gender, grade, age, mother's education or household income (Figure 3). No estimated difference exceeds 0.10 SD in absolute value, every 95% confidence interval includes zero, and a joint test does not reject the null that all five interaction coefficients are zero ($p = 0.84$).

Table 2: Assessment Conditions and Phone–In-Person Score Differences

Panel A. Reading-material delivery

		Percent correct		Phone effect (SD)		
	Children	In person	Phone	Estimate	(SE)	95% CI
School textbook	208	73.2	75.6	+0.098	(0.070)	[−0.04, +0.24]
WhatsApp image or SMS text	139	79.7	81.8	+0.065	(0.050)	[−0.03, +0.16]
Difference (messaging minus textbook)				−0.032	(0.086)	[−0.20, +0.14]

Panel B. Reported prompting and its association with scores

	Children	In person (%)	Phone (%)	Adjusted gap (pp) [95% CI]
Any prompting	364	20.3	32.4	+11.2 [+5.1, +17.3]
	Coefficient	95% CI	Children	Child–test pairs
Literacy	+0.037	[−0.16, +0.23]	299	299
Numeracy	−0.091	[−0.27, +0.09]	333	333
Pooled	−0.028	[−0.19, +0.13]	364	632

Panel C. Phone effects without reported prompting

	Full paired sample		No reported prompting	
	[95% CI]	Children	[95% CI]	Children
Literacy	+0.069 [−0.02, +0.16]	347	−0.041 [−0.15, +0.07]	163
Numeracy	−0.046 [−0.14, +0.05]	397	−0.078 [−0.20, +0.04]	190
Pooled	+0.007 [−0.07, +0.08]	428	−0.061 [−0.15, +0.03]	200

Notes: Panel A uses valid same-form literacy pairs under each reading-material protocol. Textbook assessments used the same pages in both modes; messaging supplied WhatsApp images or SMS text by phone and printed sheets in person. Percent-correct means are unadjusted. The final row reports the messaging-minus-textbook difference in estimated phone effects. Protocols were used in different recruitment cohorts and were not randomly assigned.

Panel B uses children with prompting reports at both assessments. Prompting refers to household members helping the child answer during the assessment, as recorded by the assessor once for the whole assessment. The adjusted rate difference controls for child fixed effects and assessment period. Score associations additionally control for assessment mode. Coefficients are within-child associations between reported prompting and the standardized score, in SD. Panel C compares the full primary sample with children with no prompting reported at either assessment.

All score regressions include child-by-test fixed effects and test-specific second-assessment indicators, retain full-primary-sample in-person standardization within test, and use standard errors clustered by child. All confidence intervals are 95%. Prompting was not randomized; the associations and restrictions do not identify its causal effect.

5 Robustness

Of the 1,505 children assigned a sequence, 633 were never assessed and 426 were assessed in one mode only (Figure 1). Selective completion could affect the estimated mode difference if participation is related to how children perform under each mode. I assess this concern by weighting observed assessments by the inverse of their estimated completion probabilities, using characteristics recorded at screening and assigned assessment sequence. The analysis is restricted to children whose characteristics were recorded at screening, within recruitment cohort-by-block groups with at least one completed pair; weighted and unweighted estimates use the same observations (Appendix Figure A2).[7] On this restricted sample, reweighting changes the pooled estimate from −0.012 to 0.017 SD in the primary specification (95% bootstrap CI $[-0.07, 0.10]$), and from −0.020 to 0.022 SD in the expanded specification (95% bootstrap CI $[-0.06, 0.11]$). The small average differences therefore persist after adjustment for observed predictors of participation. These estimates do not address selection on unobserved characteristics.

Because completion rates differed by assigned sequence across recruitment cohorts, the effect of the second assessment could also differ by cohort. Allowing a separate second-assessment effect for each test in each recruitment cohort yields a pooled estimate of 0.006 SD (95% CI $[-0.07, 0.08]$), with estimates of 0.081 SD for literacy and −0.057 SD for numeracy.

The expanded analysis additionally depends on linking scores across test forms. Applying its linked 2PL scores to the primary same-form sample, while retaining child-by-test fixed effects and test-specific second-assessment indicators, yields a pooled estimate of −0.014 SD (95% bootstrap CI $[-0.09, 0.06]$). The pooled result is therefore little changed by this alternative score construction when the sample and regression specification are held fixed. Appendix Figure A2 summarizes the scoring and reweighting comparisons.

6 Conclusion

Governments need independent measures of learning to assess progress toward foundational literacy and numeracy goals. This study provides experimental evidence on whether phone assessments can serve this purpose. Comparing the same children on identical test forms at home, with assessment order randomized, yields a precisely estimated pooled difference

[7] The reweighting population comprises 1,044 randomized children; the initial cohort without screening records is excluded. The primary and expanded reweighted regressions use 333 and 574 children, respectively. Logistic completion models use assigned sequence, screening gender, age and grade, the number of eligible children in the household, an indicator for an incomplete screening record, and recruitment-cohort-by-block indicators. Primary-sample weights predict same-form pair completion separately by subject; expanded-sample weights predict a valid subject assessment separately for each scheduled period.

close to zero. For literacy, estimated mode differences are small whether children use a school textbook at home or material sent through WhatsApp or SMS.

The findings support phone assessments for aggregate monitoring; they do not establish their suitability for individual diagnosis or instructional planning.

Using phone assessments for population monitoring requires a sampling and contact strategy that accounts for households that cannot be reached or do not participate, together with procedures to limit prompting during assessment. The evidence comes from low-stakes assessments and does not establish comparability when results carry consequences. Independent phone assessments offer a way to supplement teacher-reported learning data across administrative levels.

Data availability. The de-identified assessment and screening data and the code that produce every exhibit will be made available at publication in the journal's replication repository. Fieldwork was conducted by Central Square Foundation; names and phone numbers are removed and children are identified by anonymous codes.

Funding. Data collection was funded by Central Square Foundation, which conducted the fieldwork. The foundation had no role in the analysis, interpretation or writing.

Competing interests. The author declares no competing interests. The author was not employed by Central Square Foundation during the research and received no payment from it.

Declaration of generative AI and AI-assisted technologies in the writing process. During the preparation of this work the author used generative AI tools (Claude, Codex and Gemini) to assist with drafting and editing the text. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

Angrist, Noam, Micheal Ainomugisha, Sai Pramod Bathena, Peter Bergman, Colin Crossley, Claire Cullen, Thato Letsomo, Moitshepi Matsheng, Rene Marlon Panti, Shwetlena Sabarwal, and Tim Sullivan, "Mobile education builds resilience during shocks in five countries," *Nature*, 2026, *657* (8132), 689–696.

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

_ , _ , **David K Evans, Susannah Hares, Matthew CH Jukes, and Thato Letsomo**, "Practical lessons for phone-based assessments of learning," *BMJ Global Health*, 2020, *5* (7), e003030.

ASER, "Annual Status of Education Report 2024," *ASER Centre*, 2024.

Central Square Foundation, "Bharat Survey for EdTech 2025: Survey Summary," Technical Report, Central Square Foundation 2026. Four-page summary, p. 2. Accessed October 24, 2026.

Crawfurd, Lee, David K Evans, Susannah Hares, and Justin Sandefur, "Live tutoring calls did not improve learning during the COVID-19 pandemic in Sierra Leone," *Journal of Development Economics*, 2023, *164*, 103114.

GOI, "National Initiative for Proficiency in Reading with Understanding and Numeracy," *Department of School Education & Literacy Ministry of Education*, 2021.

Government of Uttar Pradesh, Department of Basic Education, "Nipun Lakshya App, version 3.0.1," Android application, Mission Prerna 2026. Accessed October 26, 2026.

Hernández-Agramonte, Juan Manuel, Olga Namen, Emma Näslund-Hadley, and Maria Loreto Biehl, "Supporting early childhood development remotely: Experimental evidence from SMS messages," *Journal of Development Economics*, 2024, *166*, 103201.

Jordan, Katy, Christina Myers, Kalifa Damani, Phoebe Khagame, Albina Mumbi, and Lydia Njuguna, "Supporting equitable access to learning via SMS in Kenya: Impact on engagement and learning outcomes," *British Journal of Educational Technology*, 2025, *56* (4), 1530–1552.

Pritchett, Lant, *The Rebirth of Education: Schooling Ain't Learning*, Washington, DC: Center for Global Development, 2013.

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

_ **and** _ , "Assessors influence results: Evidence on enumerator effects and educational impact evaluations," *Journal of Development Economics*, 2023, *163*, 103057.

Samagra Shiksha Telangana, "Intinta Chaduvula Panta: WhatsApp chatbot-based weekly assessment for students of Classes 1 to 10," Proceedings dated 8 January 2023, Office of the State Project Director, Samagra Shiksha, School Education Department, Government of Telangana, Hyderabad 2023.

Singh, Abhijeet, "Improving administrative data at scale: Experimental evidence on digital testing in Indian schools," *The Economic Journal*, 2024, *134* (661), 2207–2223.

_ **and Petter Berg**, "Myths of official measurement: Limits to test-based education reforms with weak governance," *Journal of Public Economics*, 2024, *239*, 105246.

Sobers, Shauna-Marie, Hannah L. Whitehead, Konan Nana Anicet N'Goh, Mary-Claire Ball, Fabrice Tanoh, Hermann Akpé, and Kaja K. Jasińska, "Is a phone-based language and literacy assessment a reliable and valid measure of children's reading skills in low-resource settings?," *Reading Research Quarterly*, 2023, *58* (4), 733–754.

Truecaller, "Truecaller Insights: Top 20 Countries Affected by Spam Calls in 2021," Global Spam Report, Truecaller AB, Stockholm, 17 December 2021. Archived at https://web.archive.org/web/20211217074018/https://truecaller.blog/2021/12/17/top-20-countries-affected-by-spam-calls-in-2021/.

UNESCO Institute for Statistics, USAID, World Bank, FCDO, ACER, and Bill & Melinda Gates Foundation, "Global Proficiency Frameworks for Reading and for Mathematics: Grades 1 to 9," Technical Report, UNESCO Institute for Statistics, Montreal December 2020. Two volumes. Reading: https://www.uis.unesco.org/sites/default/files/medias/fichiers/2025/02/Global-Proficiency-Framework-Reading%20(1).pdf; Mathematics: https://www.uis.unesco.org/sites/default/files/medias/fichiers/2025/02/Global-Proficiency-Framework-Math.pdf.

World Bank, *World Development Report 2018: Learning to Realize Education's Promise*, Washington, DC: World Bank, 2018.

Online Appendix

Measuring Foundational Learning by Phone: Evidence from a Randomized Crossover Experiment

Saloni Gupta

Table A1: Assessment Items

Panel A. Literacy

Task (items: G1 / G2)	Grade 1 form (*Kalrav* Class 1)	Grade 2 form (*Kalrav* Class 2)	Scoring
Listening comprehension (4 / 4, shared)	The assessor reads a short story aloud, beginning आँगन में एक नटखट बिल्ली रहती है (*A naughty cat lives in the courtyard*). Four questions: which animals appear in the story; where the cat lives; what the cat saw while drinking milk; what the cat did on seeing the mouse.		1 if correct
Picture comprehension (2 / 2)	Picture on p. 36: names the objects; describes the picture	Picture on p. 12: names the objects; describes the picture	G1: one object named; picture described in sentences. G2: two or more objects; three or more sentences
Letters (8 / 8; two shared)	न ड क म प स फ र (p. 34)	हों ठ क ग म आं जी भ (p. 24)	1 per letter
Simple words (3 / 4)	जग, बस, श्रम (p. 34)	मगन, बड़ा, गाना, पानी (p. 40)	1 per word
Complex words (6 / 8)	रात, गिना, गीत, मुख, दूध, कृषि (p. 43)	गुड़, नीबू, मिर्च, हलवा, गाजर, इमली, संतरा, पकौड़ी (p. 70)	1 per word
Sentences (2 / 2)	आम खाते मिट्ठू लाल (*Mitthu, the red parrot, eats mangoes*); इमली खट्टी मीठी बेर (*Tamarind is sour, berries are sweet*) (p. 22)	सूरज रात में चमकता है (*The sun shines at night*); चाँद सूरज से बड़ा है (*The moon is bigger than the sun*) (p. 54)	1 per sentence if more than half of its words are read correctly
Passage, read for one minute (1 / 1)	35 words, beginning गेंद, छाता और डिब्बा तीन दोस्त थे (*A ball, an umbrella and a box were three friends*) (p. 49)	42-word poem, beginning बया हमारी चिड़िया रानी (*The baya, our queen of birds*) (p. 16)	1 if more than half of the words attempted are read correctly
Reading comprehension (administered, not scored)	None	First paragraph on p. 21, beginning बरसात का दिन था (*It was a rainy day*). Three questions: whose house was getting wet; where the water started dripping from; why grandmother was worried	Not scored

Panel B. Numeracy (14 items per form, administered orally; † marks the nine items shared by the two forms)

Task	Grade 2 form	Grade 3 form
Number comparison	Which is bigger, 5 or 8? Which is smaller, 19 or 14? Which is bigger, 45 or 38?† Which is smaller, 97 or 78?†	Which is bigger, 45 or 38?† Which is smaller, 97 or 78?†
Before, after and between	What comes before 85?† Which numbers come between 65 and 68?†	What comes before 85?† Which numbers come between 65 and 68?† What comes after 794? What comes between 321 and 323?
Ordering	Arrange 6, 3, 1 in increasing order; arrange 16, 19, 12 in decreasing order†	Arrange 16, 19, 12 in decreasing order†; arrange 89, 68, 70 in increasing order
Addition and subtraction	$5+3$; $9-4$; $6+4$†; $11-5$†	$6+4$†; $11-5$†; $34+28$; $42-23$
Word problems	Rahul has 8 rupees and his mother gives him 5 more; how many does he have?† Lata has 12 chocolates and gives 4 to a friend; how many are left?†	Same two problems†

Notes: Items are transcribed from the survey instrument; page numbers refer to the Uttar Pradesh Hindi textbook *Kalrav* from which the child read, and English glosses are given in italics. Each child received one literacy form and one numeracy form. The reading-comprehension questions on the Grade 2 literacy form are not scored because, in practice, they were administered differently by mode: in person the assessor read the questions to the child, whereas by phone the questions were sent as a text message and the child was asked to read them. Reading the questions aloud made a correct answer more likely in person, so the questions were excluded from the score.

Table A2: Child and Household Characteristics of the Primary Sample

Characteristic	Mean or percentage	SD	Children
Child characteristics			
Female (%)	51.3	–	427
Age (years)	7.85	1.21	428
Grade 3 (%)	45.8	–	428
Household characteristics			
Mother: below Grade 10 (%)	58.6	–	374
Father: below Grade 10 (%)	49.1	–	373
Mother: homemaker (%)	80.9	–	376
Father: wage labor (%)	61.7	–	376
Household members	6.50	2.17	375
Children in household	3.09	1.03	376
Monthly income $\leq$ INR 10,000 (%)	94.2	–	309
Total children			428

Notes: The sample comprises the 428 children with a same-form comparison in at least one subject. Each child appears once. Age and gender come from the screening record where one is linked and otherwise from the assessment; other characteristics use reports from the in-person assessment. Means and percentages use nonmissing responses; the final column gives the number contributing to each row. Parental education below Grade 10 includes no schooling and schooling through Grade 8. Income non-disclosure is treated as missing. Household members and children are top-coded at 10 and 5, respectively. SDs are reported for continuous characteristics.

Table A3: Participation and Attrition

Panel A. Screening characteristics by participation

	Both modes	One mode	Neither	p, both versus one	p, difference by sequence	p, neither versus assessed
Female (%)	52.0	50.4	49.5	0.683	0.889	0.531
Age (years)	7.73	7.77	7.68	0.660	0.101	0.317
Grade 3 (%)	42.5	48.7	44.3	0.121	0.228	0.745
Urban block (%)	18.2	19.6	15.4	0.653	0.123	0.116
Children	346	275	571			

Panel B. Household characteristics among assessed children

	Both modes Mean (N)	One mode Mean (N)	p, both versus one	p, difference by sequence
Mother below Grade 10 (%)	53.1 (343)	61.0 (272)	0.047	0.093
Father below Grade 10 (%)	44.7 (340)	50.0 (270)	0.194	0.706
Mother homemaker (%)	83.7 (344)	82.1 (273)	0.586	0.662
Father wage labor (%)	42.0 (345)	49.6 (274)	0.059	0.004
Household members	6.67 (344)	6.18 (273)	0.007	0.409
Children in household	2.88 (345)	2.95 (273)	0.461	0.217
Monthly income $\leq$ INR 10,000 (%)	84.1 (276)	88.8 (224)	0.117	0.108

Panel C. Participation by assigned sequence

	Phone first	In person first	p, difference
Both modes (%)	31.7	26.4	0.043
One mode (%)	22.3	23.8	0.543
Neither (%)	46.0	49.8	0.186
Children	600	592	
Both modes, among assessed children (%)	58.6	52.5	0.125
Children assessed at least once	324	297	

Notes: The sample comprises 1,192 randomized children with linked screening records. Assessment requires a valid score in at least one subject; participation in both modes need not yield a same-form pair. Panel A compares screening characteristics across participation groups. Urban blocks are Nagar Kshetra and Ramnagar, as classified by the school directory. Panel B compares household characteristics reported at the first completed assessment; available-response counts are in parentheses. Comparisons are unadjusted and reports may differ by assessment mode. Household members and children are capped at 10 and 5; undisclosed income is excluded. In Panels A and B, the first p-value compares both-mode with one-mode children; the second tests whether this difference varies by assigned sequence. The third p-value in Panel A compares never-assessed with assessed children. Panel C reports participation rates by assigned sequence and completion in both modes conditional on assessment at least once. Tests use heteroskedasticity-robust standard errors. In-person assessments in Malihabad are treated as not administered following the local team's dismissal during fieldwork; phone assessments are retained. Appendix Table A5 documents the administration problems and sensitivity estimates.

Table A4: Phone Effects and Score Properties by Test

Panel A. Phone effect by test

	Phone effect			Second-assessment effect (SE)	Children
Test	Estimate	(SE)	95% CI		
Literacy, G1 textbook	+0.078	(0.076)	[−0.07, +0.23]	−0.224 (0.076)	94
Literacy, G2 textbook	+0.067	(0.058)	[−0.05, +0.18]	+0.024 (0.058)	253
Numeracy, Grade 2	−0.056	(0.062)	[−0.18, +0.07]	+0.021 (0.062)	215
Numeracy, Grade 3	−0.035	(0.071)	[−0.17, +0.10]	+0.172 (0.071)	182

Panel B. Score properties by assessment mode

Test	Items	Children	Mode	SD	Ceiling (%)	α	Correlation
Literacy, G1 textbook	26	94	In person	27.0	24.5	0.94	0.62
			Phone	26.6	16.0	0.95	
Literacy, G2 textbook	29	253	In person	26.6	31.2	0.95	0.55
			Phone	26.2	29.2	0.95	
Numeracy, Grade 2	14	215	In person	27.5	17.7	0.89	0.53
			Phone	25.3	14.4	0.87	
Numeracy, Grade 3	14	182	In person	28.5	17.0	0.89	0.47
			Phone	24.9	8.2	0.85	

Notes: Both panels use children with a valid score on the same test in both modes. Panel A reports separate regressions by test, with child fixed effects and a second-assessment indicator, on percent-correct scores standardized using the in-person mean and SD for that test. The phone coefficient compares phone with in-person scores; the second-assessment coefficient captures a common change at the later assessment. Heteroskedasticity-robust standard errors are in parentheses; all confidence intervals are 95%. A test for differential carryover compares the sum of each child's phone and in-person scores between the two assigned sequences; the difference (in person first minus phone first) is −0.09 SD (95% CI [−0.47, +0.29]) in literacy, −0.12 ([−0.45, +0.21]) in numeracy and −0.11 ([−0.43, +0.22]) pooled.

Panel B describes the percent-correct scores of the same children by mode. Ceiling: share of children scoring 100%. α: Cronbach's alpha, a measure of internal consistency within one assessment (equal to KR-20, as all items are scored correct or incorrect). Equality of alpha is tested using a child-level bootstrap with 2,000 replications; equality of variance uses the Pitman–Morgan test for paired scores. Both tests reject equality only for Grade 3 numeracy ($p = 0.02$ for alpha and $p = 0.04$ for variance). Correlation: Pearson correlation between a child's phone and in-person scores on the test. Pooled across the two forms of each subject, the correlation is 0.57 in literacy and 0.50 in numeracy.

Table A5: Assessment Administration and Sensitivity to the Dismissed Team's Records

Panel A. Assessment administration and scores

	Dismissed team	Other teams, same cohort
In-person visits	70	288
Median recorded duration (minutes)	17.0	22.5
Visits with field notes (%)	10.0	45.5

			Mean score (%)		Phone minus in person (SD)
Team	Subject	Children	In person	Phone	[95% CI]
Dismissed team	Literacy	32	59.4	77.7	+0.684 [+0.318, +1.049]
	Numeracy	39	53.5	58.2	+0.163 [−0.129, +0.455]
Other teams,	Literacy	197	77.7	78.4	+0.031 [−0.105, +0.167]
same cohort	Numeracy	224	69.9	66.5	−0.128 [−0.249, −0.007]

Panel B. Sensitivity of the main estimates

		Phone minus in person (SD)		
Sample	Children	Literacy	Numeracy	Pooled
Primary specification: same-form pairs				
Current sample	428	+0.069	−0.046	+0.007
Restore otherwise valid visits	470	+0.126	−0.025	+0.045
Expanded specification: all valid assessments				
Current sample	872	+0.056	−0.063	−0.005
Restore otherwise valid visits	900	+0.099	−0.036	+0.030
Remove the whole affected block	822	+0.056	−0.060	−0.003

Notes: The local in-person team in Malihabad was dismissed during the October 2021 fieldwork, before analysis, after Central Square Foundation's field supervisors found it difficult to coordinate with the team and observed problems establishing rapport and confirming that required reading material was available. Its 70 in-person visits are excluded from the current sample; assessments by the regular phone team are retained.

Panel A compares all recorded visits for administration measures and otherwise valid same-form pairs for scores. Recorded duration can include pauses and data entry. Score means are unadjusted; standardized differences adjust for assessment order, with standard errors clustered by child. These comparisons do not identify the cause of score differences. Child counts are not additive across subjects.

Panel B retains all other validity rules: restoration adds 67 literacy and 69 numeracy assessments. Primary estimates retain the current sample's in-person standardization within form and child-by-test fixed effects with test-specific second-assessment terms. Expanded estimates recalibrate linked 2PL scores and repeat the screening-characteristic, sequence, period and recruitment-cohort-by-block adjustments for each sample. Expanded sensitivity estimates are reported without bootstrap confidence intervals. Removing the whole affected block leaves the primary sample unchanged because it contains no completed pairs from that block.

Table A6: Composition of the Expanded Sample

Subject	Children				Subject assessments
	Same form in both modes	Different forms across modes	One mode only	Total	
Literacy	347	67	432	846	1,260
Numeracy	397	48	426	871	1,316

Notes: Categories are defined separately for each subject. The first two categories contribute two valid subject assessments per child; the third contributes one. Thus literacy contributes $2(347+67)+432 = 1,260$ assessments, and numeracy contributes $2(397+48)+426 = 1,316$. A child may have valid scores in both modes for one subject but in only one mode for the other. Subject samples overlap: the expanded pooled analysis includes 872 unique children and 2,576 subject assessments. The primary pooled analysis includes 428 unique children with a same-form comparison in at least one subject. Children are counted once in each pooled total.

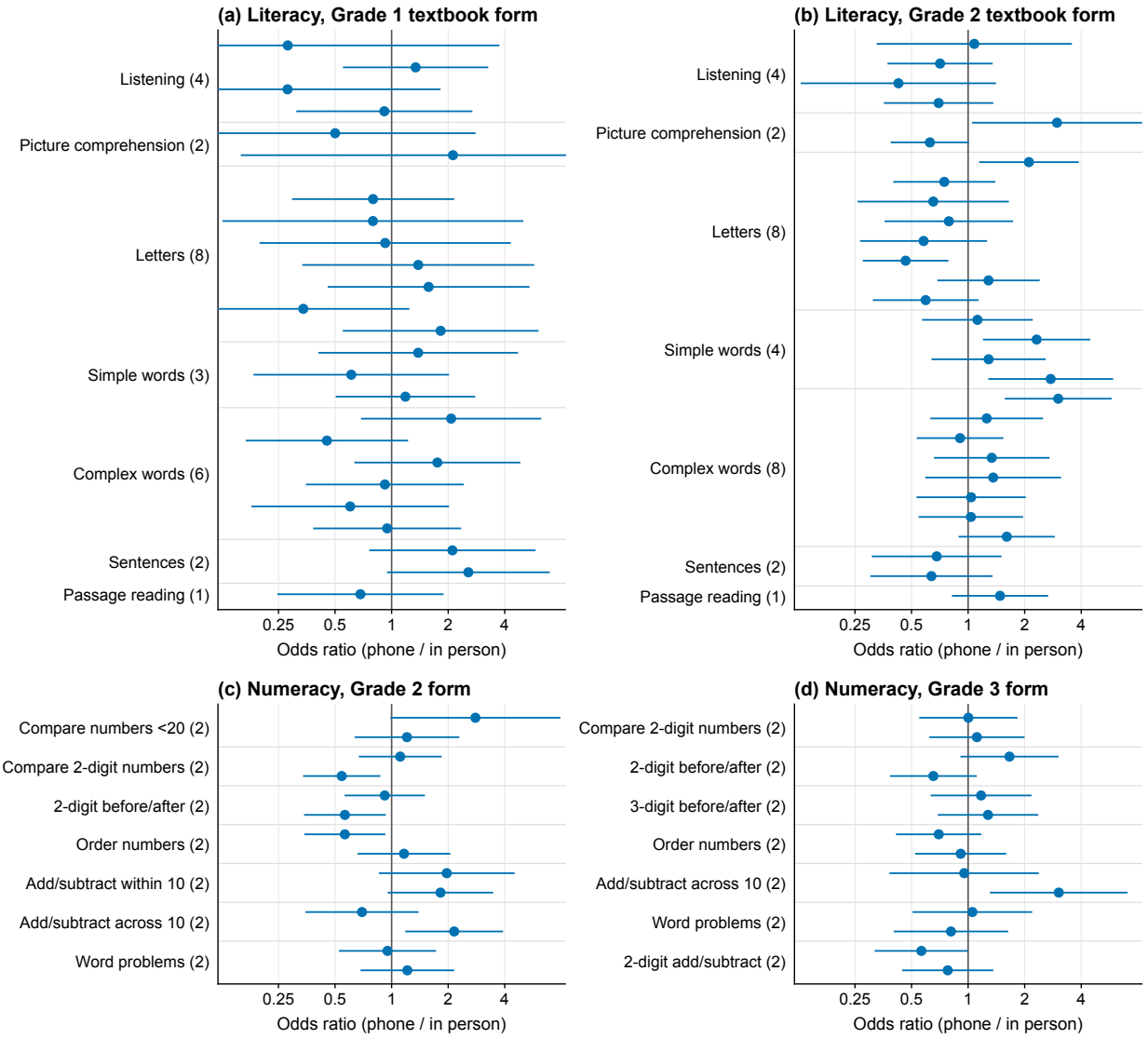


Figure A1: Item-Level Differences Between Assessment Modes

Notes: Sample: the primary same-form paired sample for each test (94, 253, 215 and 182 children, each assessed once by phone and once in person). Each point is one item, grouped by task (number of items in parentheses). For each item, children are matched on the number of the form's other items they answered correctly. Points show Mantel–Haenszel odds ratios of a correct answer by phone relative to in person, with 95% confidence intervals based on the Robins–Breslow–Greenland variance. Values above 1 indicate higher odds of a correct answer by phone; values below 1 indicate higher odds in person. The horizontal axis is logarithmic; intervals are truncated at the axis limits. A child-level bootstrap (999 draws) gives standard errors on the log odds ratio on average 6% larger than the analytic standard errors (1% to 12% by form). In a logistic model with mode, the standardized rest score, their interaction and the scheduled period, with standard errors clustered by child, 6 items remain significant after a 5% false-discovery-rate adjustment within form.

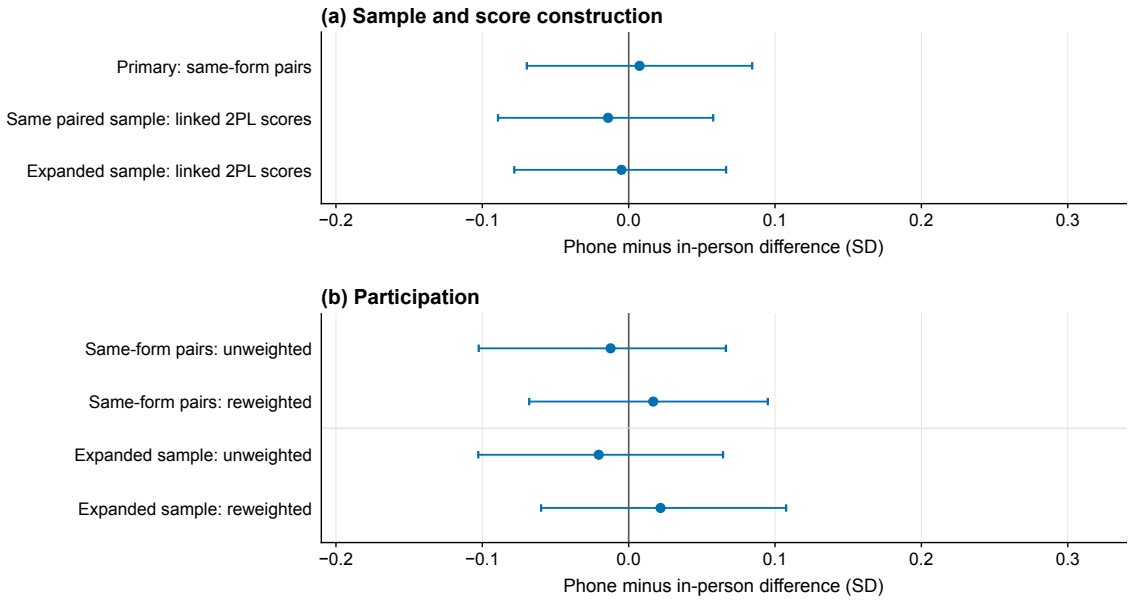


Figure A2: Sensitivity and Robustness of the Pooled Phone Effect

Notes: Points show pooled phone-minus-in-person estimates and 95% confidence intervals. Panel (a) compares standardized percentage-correct scores with linked 2PL scores on the same paired sample, then includes all valid assessments in the expanded specification. Shared items link forms within each subject; the expanded specification also changes the regression controls. Panel (b) compares unweighted and reweighted estimates on identical observations with screening records. Unweighted estimates give equal weight to each pair or subject assessment; reweighted estimates give greater weight to observations with lower predicted completion probabilities. The reweighted regressions use 333 children in the primary specification and 574 in the expanded specification, weighted to 1,044 randomized children with screening records. Primary confidence intervals use standard errors clustered by child; linked-score and reweighting confidence intervals use a child-level bootstrap with 999 replications. Reweighting addresses observed selection only.