Fiala, Lenka et al.

Working Paper

A Comment on "Delivering Remote Learning Using a Low-Tech Solution: Evidence from a Randomized Controlled Trial in Bangladesh"

I4R Discussion Paper Series, No. 241

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Fiala, Lenka et al. (2025) : A Comment on "Delivering Remote Learning Using a Low-Tech Solution: Evidence from a Randomized Controlled Trial in Bangladesh", I4R Discussion Paper Series, No. 241, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/321368

INSTITUTE for REPLICATION

No. 241
I4R DISCUSSION PAPER SERIES

A Comment on "Delivering Remote Learning Using a Low-Tech Solution: Evidence from a Randomized Controlled Trial in Bangladesh"

Lenka Fiala
Essi Kujansuu
David Valenta
Michael Wiebe
Abel Brodeur
Jack Fitzgerald
Derek Mikola
Juan P. Aparicio
Matthew D. Webb

July 2025

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 241

A Comment on "Delivering Remote Learning Using a Low-Tech Solution: Evidence from a Randomized Controlled Trial in Bangladesh"

Lenka Fiala[1,2], Jack Fitzgerald[3,4,2], Essi Kujansuu[5,6,2], Derek Mikola[1,2], David Valenta[1,2], Juan P. Aparicio[7,1,2], Michael Wiebe, Matthew D. Webb[8], Abel Brodeur[1,2]

[1]University of Ottawa/Canada
[2]Institute for Replication
[3]Vrije Universiteit Amsterdam/The Netherlands
[4]Tinbergen Institute, Amsterdam/The Netherlands
[5]University of Innsbruck/Austria
[6]University of Turku/Finland
[7]EAFIT University, Medellín/Columbia
[8]Carlton University, Ottawa/Canada

JULY 2025

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

A Comment on "Delivering Remote Learning Using a Low-Tech Solution: Evidence from a Randomized Controlled Trial in Bangladesh"

Lenka Fiala, Jack Fitzgerald, Essi Kujansuu, Derek Mikola, David Valenta, Juan P. Aparicio, Michael Wiebe, Matthew D. Webb, Abel Brodeur

July 4, 2025

Abstract

Wang et al. (2024) report that Bangladeshi students randomly given access to lessons on a phone server saw significant learning gains during COVID-19 school closures. We identify three sets of anomalies. First, this experiment shares participants with another experiment conducted simultaneously in the same region, but test scores for the same children systematically differ between the two experiments. Second, test scores for treated participants exhibit a uniform upward shift that is completely insensitive to the number of lessons children complete. Third, numerous documentation inconsistencies (e.g., concerning survey materials, randomization procedures, etc.) cast doubt on the study's data.

KEYWORDS: Reproduction, school closures, remote education, COVID-19, randomized controlled trial, Bangladesh

JEL CODES: B41, C12, C93, I21, I24

Brodeur (corresponding): University of Ottawa and Institute for Replication. E-mail: abrodeur@uottawa.ca. Fiala: University of Ottawa and Institute for Replication. Fitzgerald: Vrije Universiteit Amsterdam, Tinbergen Institute, and Institute for Replication. Kujansuu: University of Innsbruck, University of Turku, and Institute for Replication. Mikola: University of Ottawa and Institute for Replication. Valenta: University of Ottawa and Institute for Replication. Aparicio: EAFIT University, University of Ottawa and Institute for Replication. Wiebe: Independent Researcher. Webb: Carleton University. Fiala, Fitzgerald, Kujansuu, Mikola, Valenta and Wiebe contributed to the analysis and writing. Aparicio and Webb contributed to the analysis. Brodeur contributed to the writing. We thank RushTranslate for providing official translations of the Bengali text in the raw data into English; all other errors are our own. We thank Carl Bonander, Niklas Jakobsson, and Anders Kjelsrud for feedback and suggestions.

1 Introduction

Wang et al. (2024a) – henceforth REM24 – evaluate the effectiveness of an interactive phone-based remote learning program for primary school children in Bangladesh during COVID-19 school closures. In their randomized controlled trial, the authors assign 1763 primary school children in first through fourth grades to either a control group or one of two treatment groups. In the first treatment ("standard"), the children are given access to a 15-week program of numeracy and literacy lessons over an interactive voice response (IVR) system.[1] In the second treatment ("leadership"), the children are given the "standard" treatment and access to additional leadership lessons targeting non-cognitive skills. Lessons are intended to be delivered with the support of caregivers (usually parents) and no direct input from teachers or other trained professionals.

The results indicate that the intervention significantly improved student learning outcomes, with treated children achieving a 0.6 standard deviation increase in test scores relative to the control group ($p < 0.005$). The program was particularly effective for students with initially lower test scores, and for students with less educated parents. In contrast, the authors find no significant improvements in leadership skills, or non-cognitive skills such as grit, empathy, growth mindset, and impulsivity.

The authors compare their intervention against the existing literature and report that they find larger effect sizes than those obtained from numerous similar interventions (Gortazar et al. 2024, Carlana and La Ferrara 2021, Kraft et al. 2022, Angrist et al. 2022, Crawfurd et al. 2023). These interventions include live, online, or phone interactions with a teacher or a tutor who can immediately correct mistakes or provide additional exercises. Looking at interactive audio instruction specifically, REM24's effect sizes are on the higher end of the spectrum as estimated by the meta-analysis of Ho and Thukral (2009) for such programs in developing coun-

[1] IVR is an over-the-phone system that allows students to listen to lessons that can be selected from a predesigned menu, akin to phone trees typically used in automated customer service systems.

tries. These effects are slightly smaller than the effect of a 13-week phone-based mentoring program provided to primary schoolaged children in rural Bangladesh by volunteers during COVID-19 school closures (Hassan et al. 2024b), which is written by a largely overlapping group of authors and for which a similar comment has been written by Aparicio et al. (2025).

In this comment prepared for the Institute for Replication (Brodeur et al. 2024), we examine the reproducibility, robustness, and credibility of REM24. Our analysis relies on REM24's publicly available replication repositories and an additional folder containing raw data, clarifications, and code provided by the authors to the editors of this journal. The editors forwarded this folder to us in late February 2025. All our analyses were successfully reproduced by multiple coauthors. Section 2 explains the materials we have access to and when we obtained them.

Section 3 discusses the relationship between REM24 and a related study by Hassan et al. (2024b). Both studies examine randomized educational interventions but have overlapping samples, timelines, and test materials. However, inconsistencies emerge in test score distributions and participant overlap. We find that 132 individuals appear in both datasets, with some receiving treatment in both studies, raising concerns about interaction effects between interventions. Additionally, there are irregularities in individual test scores across both papers. This is particularly true for the students treated in REM24 but untreated in Hassan et al. (2024b), whose scores appear to differ systematically. These findings cast doubt on the credibility of the data.

Section 4 examines whether the lessons discussed in REM24 genuinely improve student test scores. First, we find that treated children exhibit a uniform upward shift in test scores compared to untreated children, with test scores being completely insensitive to the number of lessons completed. I.e., this uniform upward shift is just as pronounced for children who report completing a handful of lessons as it is for children who report completing all 30. Second, we investigate the reliability of self-reported lesson completion data, noting that the number of lessons reportedly

accessed far exceeds what the server data would suggest is possible. Third and finally, we uncover collected data on a pre-registered 'general knowledge' test score that REM24 claim was never collected, despite being analyzed in a prior working paper version of REM24. Despite REM24 conceding that the IVR treatments did not target general knowledge, the IVR intervention's estimated effects on general knowledge are virtually identical to those on numeracy and literacy scores, which were directly targeted by the IVR lessons.

Section 5 highlights documentation inconsistencies and data irregularities in the REM24 study. The reported sample size varies across sources, and there are numerous discrepancies between the Bangla survey materials, Kobo forms, and their English translations, casting doubt on reporting accuracy. Additionally, the study fails to disclose key procedural details, such as stratification of the randomization and the use of SMS reminders. Data exhibit several irregularities, such as inconsistencies in parental ages between surveys. Taken together, these findings raise concerns about the validity of the reported results and suggest that test score improvements may not be driven by the IVR lessons described in the paper.

2 Data

REM24's original replication package was published in February 2024 (Wang et al. 2024b). The authors published a more detailed replication package in February 2025 following our queries (Wang et al. 2025). We further requested additional clarifications, codes, materials, and raw data, resulting in a third replication package being provided to journal editors in February 2025. This third package is not publicly available. We also examine data from the experiment reported in Hassan et al. (2024b), specifically its V1 replication package (Hassan et al. 2024a).

We additionally examine REM24's pre-registration and pre-analysis plan. This experiment was registered on August 30th, 2021 through the AEARCT registry.[2] The platform also links to a pre-analysis plan, dated August 27th, 2021. This

[2] https://www.socialscienceregistry.org/trials/7931

pre-analysis plan was thus registered after the intervention began (in June 2021).

3 Overlap With Hassan et al. (2024b)

REM24 and Hassan et al. (2024b) – henceforth TEL24 – exhibit key overlaps in experimental design, sample and test score measures.[3] In the following subsection, we elaborate on these parallels. Leveraging this overlap, we then demonstrate that while individual baseline test scores are identical across both studies, endline scores recorded in the REM24 dataset are, on average, significantly higher than those observed in TEL24. These discrepancies are notably concentrated within the REM24 treatment group.

3.1 Overlaps in Experimental Design, Sample and Test Score Measures

Both REM24 and TEL24 report educational interventions where children and their families are randomly assigned to different treatments. REM24 have one control and two treatment arms (standard and leadership) while TEL24 have one control group and one treatment group (telementoring). Both experiments overlap in time: REM24 began in June 2021 and ended in October 2021, with in-home endline assessments and surveys being completed in November 2021, while TEL24 report that the second in-home endline assessments and surveys were completed in December 2021. TEL24's replication data contradicts the claim that its second endline assessments were completed in December 2021. In TEL24's replication repository, raw survey data file `Endline2-Survey-Data.xlsx` contains survey date variable `tm_e2_survey_date`, which records no participant receiving the endline survey at any date later than November 24th, 2021. Only six observations have dates later than November 13th, 2021. Survey dates are not available in any of the replication repositories furnished for REM24. If the timeline in REM24 and the replication data in TEL24 are to be believed, then the endline assessments for these experiments

[3] Aparicio et al. (2025) was prepared for the Institute for Replication to reproduce TEL24. Parts of Section 3 first appeared there.

were administered at about the same time.[4]

Both studies use the same questions for their numeracy and literacy assessments in November 2021. Appendix Figures A6 and A7 reproduce images of the questions asked in the endline assessment of REM24 and the second endline assessment of TEL24. The ten literacy questions for REM24's endline assessment are identical to the six English literacy and four Bangla literacy question from TEL24's second endline assessment, and the numeracy questions in REM24's endline assessment are identical to the mathematics questions in TEL24's second endline assessment.

Both interventions also appear to have the same individuals. We successfully match 132 individuals using `CHILD_ID`, `VILLAGE_ID`, and `FAMILY_ID` from both replication datasets.[5] A verification check confirms identical baseline literacy and numeracy scores across both datasets; see Figure 1, which reproduces Appendix Figure A1 in Aparicio et al. (2025).[6] Among the 132 matched children, treatment assignments overlap between the two studies. For instance, some students who received treatment in REM24 were also treated in TEL24. This overlap between the two studies contradicts REM24's AEARCT registry, which clearly states that "This trial does not extend or rely on any prior RCTs." Though the interactions between interventions are potentially concerning (Muralidharan et al. 2025), our primary concern lies with individual test score discrepancies.

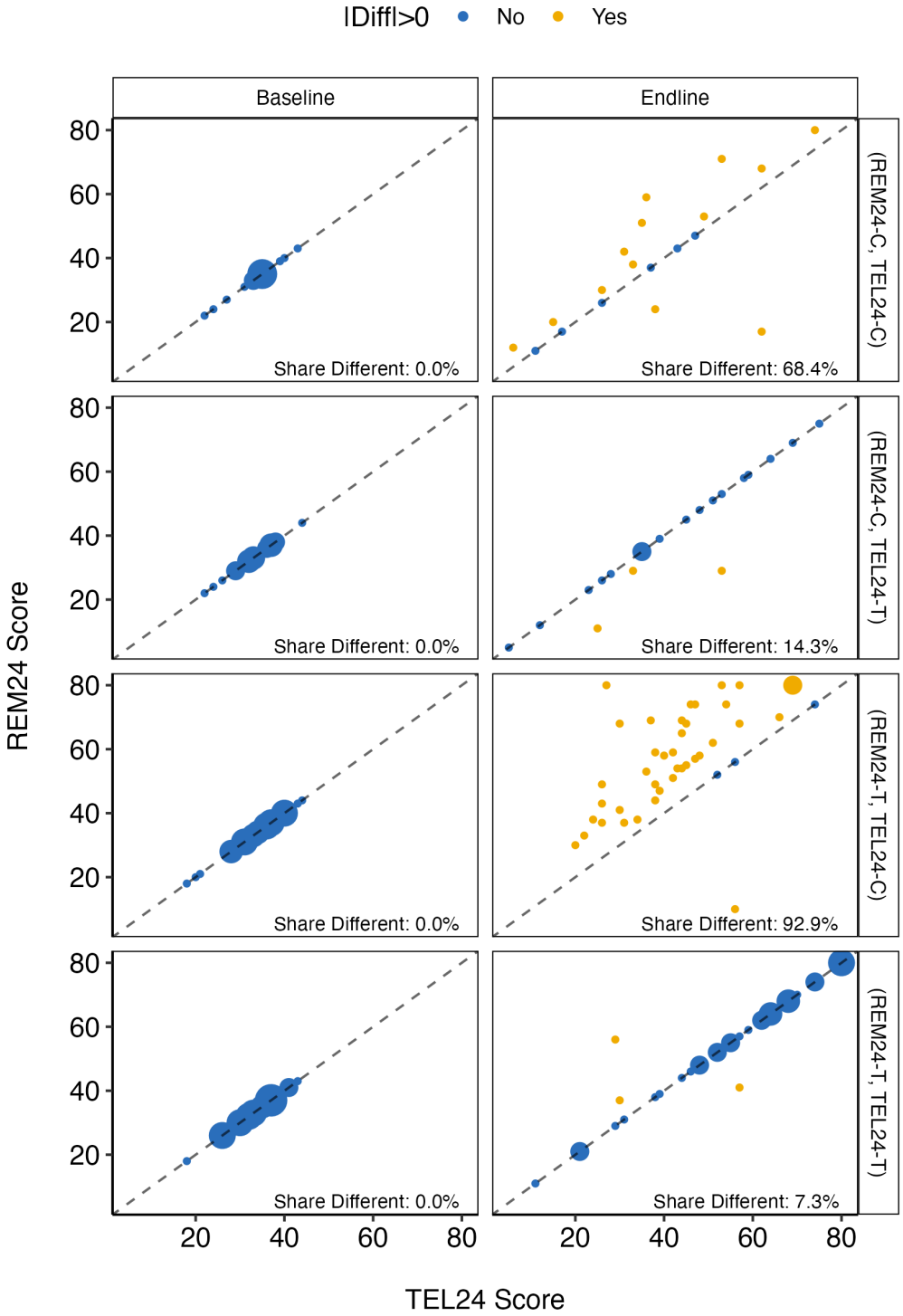
3.2 Discrepancies in Test Scores

Figure 1 reproduces Figure A1 in Aparicio et al. (2025). It plots total test scores (the sum of numeracy and literacy scores) for the baseline assessment, the REM24 endline, and the second TEL24 endline, each split by treatment group. The original baseline files contain 132 individuals, but this sample is reduced to 123 individuals by the final endline measurement.

[4]The exact sequence of data collection remains unclear. See Aparicio et al. (2025) for more details on the timeline inconsistencies.

[5]A less conservative approach matching only on `CHILD_ID` identifies 135 individuals.

[6]However, discrepancies exist in baseline covariates, e.g., some children's genders are inconsistently recorded.



Note: This figure is a reproduction of Figure A1 in Aparicio et al. (2025). The left column compares the baseline scores for both papers. All observations are on the main diagonal, meaning there is a perfect correspondence between baselines scores across both TEL24 and REM24 data. The right column shows the endline scores for the same group of individuals across TEL24 and REM24. An aggregated version of this figure is available as Figure A5.

Figure 1: Comparing Baseline Scores and Endline Scores Across TEL24 and REM24

There are considerable inconsistencies between the REM24 and TEL24 scores of students who appear in both samples. Students generally score higher in REM24 than in TEL24, particularly those in REM24's treatment groups and TEL24's control group, where 93% deviate from the 45-degree line.

3.3 Discrepancies in Psychological Measures

Both REM24 and TEL24 collect psychological measures, sometimes using the exact same measures or questions. For example, impulsivity is measured using the Impulsivity Scale for Children (ISC), an eight-item scale of five-point Likert items, and behavioral difficulties are measured using the Strengths and Difficulties Questionnaire (SDQ), a 25-item scale of three-point Likert items.

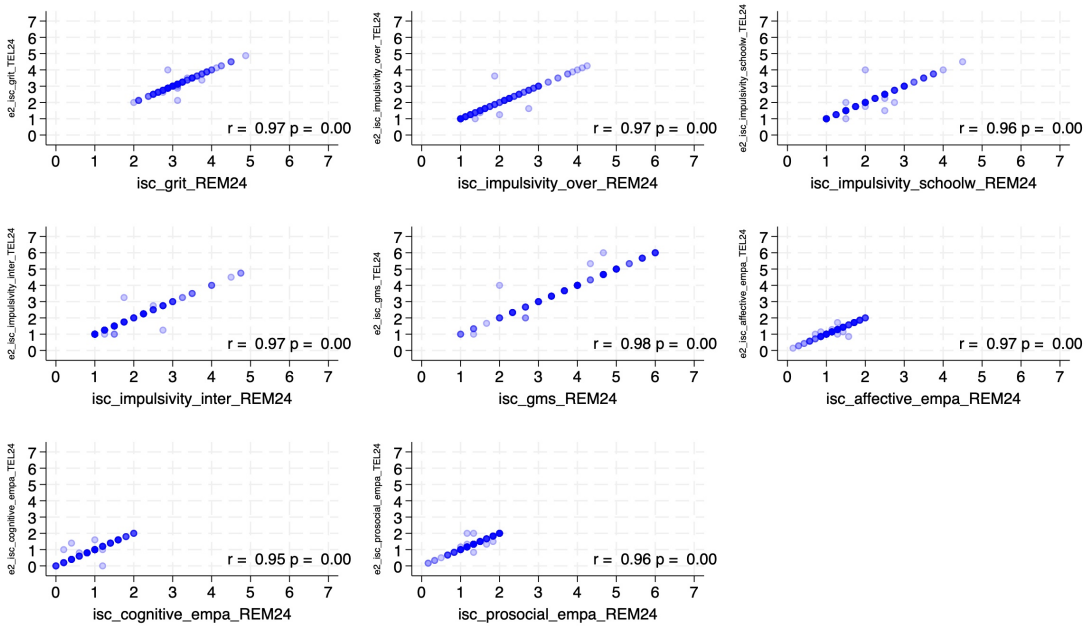
For the 123 children who remain in both studies, ISC scores are nearly identical (differing for eight individuals), but SDQ scores differ substantially. Figures 2a and 2b illustrate that across all eight items in the ISC scale, responses across studies follow a near-perfect diagonal trend in scatter plots, implying that the children responded consistently across both datasets. However, the SDQ responses are widely dispersed. Several items show no statistically significant correlation across datasets collected just weeks apart.

3.4 Response of REM24 Authors

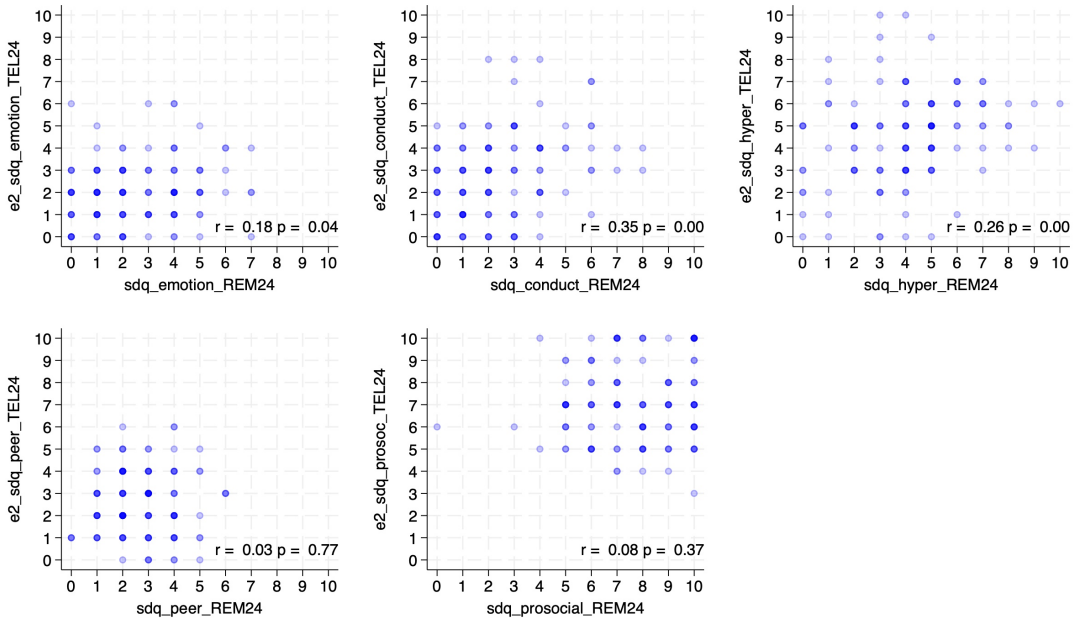
We approached the authors of REM24 with an earlier version of this report in late March 2025 to give them a chance to respond. Three of the four authors – Hashibul Hassan, Asad Islam, and Liang Choon Wang – provided a response. Their response to the irregularities pointed out so far is threefold. First, they acknowledge that some students were part of both studies and some students took the same assessment multiple times. The responding authors offer the following explanation, attributing responsibility to a project manager at the Global Research and Development Initiative (GDRI):

> "Hassan et al. (2025) investigated the reason for why overlapping chil-

(a) ISC Responses in REM24 and TEL24



(b) SDQ Responses in REM24 and TEL24



Note: Each subfigure represents a different outcome variable in both REM24 and TEL24. All scores here take on discrete values which vary by question. Figure (a) corresponds to the ISC scale items while Figure (b) corresponds to the SDQ subscales. Each y-axis is the score recorded in the TEL24 dataset and each x-axis is the score recorded in the REM24 dataset. The darker the colour, the more observations on that part of the grid. We report the correlation between the outcomes across the two papers and its significance in the lower right corner.

Figure 2: Comparing ISC and SDQ Outcomes Across REM24 and TEL24

> dren in the treatment group of TEL24 were more likely to complete only one test in late 2021.[7] Following Aparicio et al.'s (2025) observation, GDRIs project manager was contacted. According to the project manager, when planning the second endline of TEL24 in late 2021, they realized – based on names and phone numbers – that some parents and children would need to complete REM24's endline survey followed by TEL24's second endline survey. Concerned that these families might refuse to respond to the latter, the project manager decided that (a) parents would be required to do both surveys as the survey questions were different, but (b) overlapping children would only have to complete the earlier assessment instrument for REM24 endline since TEL24 had the same assessment instrument. The project manager did not inform us of this decision, reflecting a grave misjudgment by the project manager."

The responding authors argue that all children whose endline scores are on the 45-degree line in Figure 1 only took one test, with the reuse of their REM24 scores for TEL24 explaining their perfect one-to-one alignment between the two experiments. The responding authors go on to argue that this explains the differences between ISC and SDQ scores, as SDQ scores were filled out by parents (who took endline assessments twice) whereas ISC scores were filled out by children (who may have only been given endline assessments once).

Second, the responding authors contend that off-45-degree observations are concentrated in TEL24's treatment group because the project manager was more familiar with these families due to arranging and providing telementoring to families treated in TEL24. Third, the responding authors argue that off-45-degree students perform globally worse on the TEL24 assessment than on the REM24 assessment because students receive no feedback on the REM24 assessments, and are less mo-

[7] Hassan et al. (2025) refers to the authors' response to the Aparicio et al. (2025) comment. Concerning the score discrepancies, the authors' response to Aparicio et al. (2025) and this comment largely overlap.

tivated on the second attempt at an identical assessment.

The first part of the response is both inconsistent with the published version of REM24 and unverifiable. REM24 contend that students on the 45-degree line completed the REM24 assessments *before* the TEL24 assessments. To support this argument, the response provides a figure (Illustration 1) claiming that the endline assessments for REM24 were conducted in *October-November* 2021, whereas the one-year endline assessments for TEL24 are reported to take place in November 2021. This timeline contradicts the published version of REM24, wherein Figure 3 states that endline assessments were administered entirely in *November* 2021. This distinction is important because we know from timestamps in TEL24's replication repository that the vast majority of TEL24's assessments were completed in early November. This makes it difficult for all overlapping students' REM24 assessments to be taken before their TEL24 assessments if, as the published version of the paper claims, the REM24 assessments were all administered in November 2021. The new timeline in the response gives the sampling frame for REM24 more breathing room in October 2021. In addition, verifying that the TEL24 test scores for students on the 45-degree line are simply copy-pasted from those students' REM24 test scores would require timestamps, which could verify that such students' TEL24 assessments were taken on a date before their parents completed the parental survey for TEL24. These timestamps are not available in any of the three replication repositories provided by the REM24's authors.

Further, the authors' response does not explain systematic variation in test scores by treatment groups. Though the familial familiarity response would explain why off-45-degree observations are concentrated in *TEL24's* control group, it does not explain why off-45-degree observations are specifically concentrated in *REM24's* treatment group (for those already in TEL24's control).

3.5 Statistical Comparison of Test Scores

Aparicio et al. (2025) expand on concerns regarding data integrity by statistically

comparing test scores from REM24 and TEL24. Academic test responses are expected to follow systematic patterns, where students who perform well on initial questions are generally more likely to continue performing well. Any deviations from these expected patterns may indicate potential data inconsistencies or errors.

Aparicio et al. (2025) use a multivariate probit model, specifying and estimating it in Stan, to analyze response patterns and in comparison to expected distributions. Their findings reveal significant anomalies, particularly in TEL24. Specifically, a subset of treated students exhibit inflated scores, while control group participants perform below expected levels.

4 Do Lessons Improve Test Scores?

If the lessons described in REM24 improve student performance, then we should expect that more lessons yield better performance than fewer lessons. We evaluate whether this relationship holds by examining associations between test scores and the total number of lessons reported completed. Dataset `IVR_Data.dta` in the first replication repository stores variables `totlessons_num` and `totlessons_lit`, which respectively record the total number of numeracy and literacy lessons reported as completed by each student-caregiver dyad. 44% (46%) of students in the treatment groups completed all 30 numeracy (literacy) lessons. None of the 1047 (1058) students in the treatment groups with defined values for `totlessons_num` (`totlessons_lit`) reported completing zero numeracy (literacy) lessons. `totlessons_num` and `totlessons_lit` are both coded as missing for all students in the control group. In what follows, we analyze lesson completion variables that are recoded to equal zero, rather than taking on missing values, for students in the control group.

Though REM24 conduct a similar analysis, our analysis is more granular and reveals considerable anomalies. Specifically, REM24 examine dosage responses to lesson counts in Appendix Figure B7, finding null effects. As REM24 note, these dosage-response relationships after the extensive margin are not causal because les-

son completion intensity is not randomly assigned, and recent literature documents considerable uncertainty and noise in observational analyses compared to experiments (Bernard et al. 2024). However, our analyses in the rest of this section point to a considerable irregularity in estimated learning gains: a flat and uniform upward shift in test scores immediately after the extensive margin, which drives virtually the entire estimated effect of IVR lessons on student achievement.

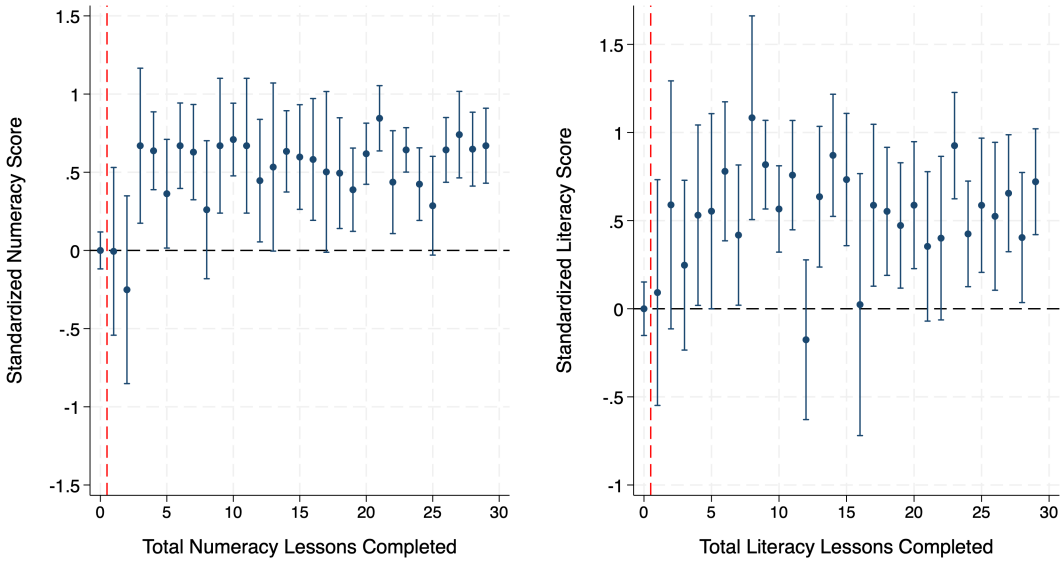
4.1 Nonparametric Analyses

Figure 3 shows a binscatter regression displaying average numeracy (literacy) test scores by number of numeracy (literacy) lessons completed. Test scores appear to jump immediately after receiving the first few lessons; additional lessons are not meaningfully associated with any additional benefit. Numeracy scores jump by 0.66 standard deviations after completing just three lessons, and literacy scores jump by 0.58 standard deviations after completing just two lessons. There is no evidence that students learn anything more from studying beyond these first few lessons.

Though these patterns could genuinely emerge if the first few IVR lessons cover a substantial share of the necessary materials for the assessments, we show in Appendix A that this is unlikely. The jump in test scores observed after the first handful of lessons does not appear to be driven by students quickly learning all necessary material for their assessments. Discontinuous jumps in test scores emerge immediately at the extensive margin of receiving any lessons regardless of how much of the assessment materials are covered in those first few lessons.

4.2 Linear Analyses

Alternative linear models can address two drawbacks of the nonparametric analyses in Section 4.1. First, one potential explanation for the flat relationships between lessons completed and test scores is that meaningful positive learning gains from lessons conflict with negative selection into more intense study schedules. That is, worse-performing students could self-select, or be selected by their caregivers,



Note: Nonparametric relationships between lesson completion variables and standardized endline test scores are produced by `binsreg` in Stata (Cattaneo et al. 2024). 95% standard errors are presented based on standard errors clustered at the village level. Observations with lesson completion variables equaling 30 (the mode) are dropped to prevent observations with zeros from being dropped. The dashed red vertical line indicates the extensive margin of observations who (are assigned to) take one or more lessons.

Figure 3: Endline Test Scores by (Highest) Lesson Completed

into higher study intensities if there is sufficient concern for the student's academic performance. Second, there is relatively low power to investigate average test scores at a specific number of lessons completed. A large plurality of treated students report completing all lessons. Below 30 lessons, only between 7-46 (7-36) students report completing each unique positive number of numeracy (literacy) lessons.

We address both issues with the nonparametric analysis using linear models that control for baseline test scores. Table 1 shows linear relationships between reported lesson completion and test scores, with and without controls for baseline test scores. When untreated students are kept in the sample, the regression suggests that completing a single IVR lesson out of 30 increases test scores by 0.017-0.018 standard deviations (Panel A). However, across all models, just excluding the untreated students who take zero lessons decreases estimated benefits of additional lessons by 87-93%, with no estimate being statistically significantly different from zero (Panel B). These results are precise and hold regardless of whether we control for baseline test scores.

	Standardized Endline Numeracy Scores (1)	Standardized Endline Numeracy Scores (2)	Standardized Endline Literacy Scores (3)	Standardized Endline Literacy Scores (4)
Panel A: Zero Lessons (Untreated Participants) Included				
Total Lessons Completed	0.017	0.017	0.018	0.018
	(0.002)	(0.002)	(0.003)	(0.003)
	[0]	[0]	[0]	[0]
N	1577	1577	1586	1586
Controls for Baseline Test Scores		✓		✓
Panel B: Zero Lessons (Untreated Participants) Excluded				
Total Lessons Completed	0.002	0.002	0.002	0.001
	(0.003)	(0.003)	(0.005)	(0.005)
	[0.541]	[0.656]	[0.72]	[0.812]
N	1015	1015	1024	1024
Controls for Baseline Test Scores		✓		✓

Note: Linear regression estimates are presented alongside standard errors clustered at the village level in parentheses and raw p-values in brackets. Total lessons completed and baseline test scores represent numeracy (literacy) lessons and scores in columns where the outcome is endline numeracy (literacy) scores.

Table 1: Linear Estimates of Dosage Response With and Without Untreated Participants

In other words, the relationship between lessons completed and test scores is functionally flat beyond the extensive margin. However, as shown in Figure 3, this flat line uniformly lays above the average test score in the control group. Consequently, the only variation in lessons that has any relationship with test scores is found at and shortly after the extensive margin, when scores uniformly shift upward.

4.3 Lesson Completion, Self-Reports, and Server Data

Across REM24's replication repositories, lesson completion is possible to measure using both self-reported and server data. The self-reported data comes from printed sheets where households were asked to keep track of lesson completion. This self-reported data is presumably captured in the aformentioned variables `totlessons_num`, `totlessons_lit`, `complete_num_1` through `complete_num_30`, and `complete_lit_1` through `complete_lit_30`. The server data is made available in REM24's third replication repository file `server_phone_quiz.dta`. REM24's dosage response analyses in Appendix Figures B4 and B7 are performed using the

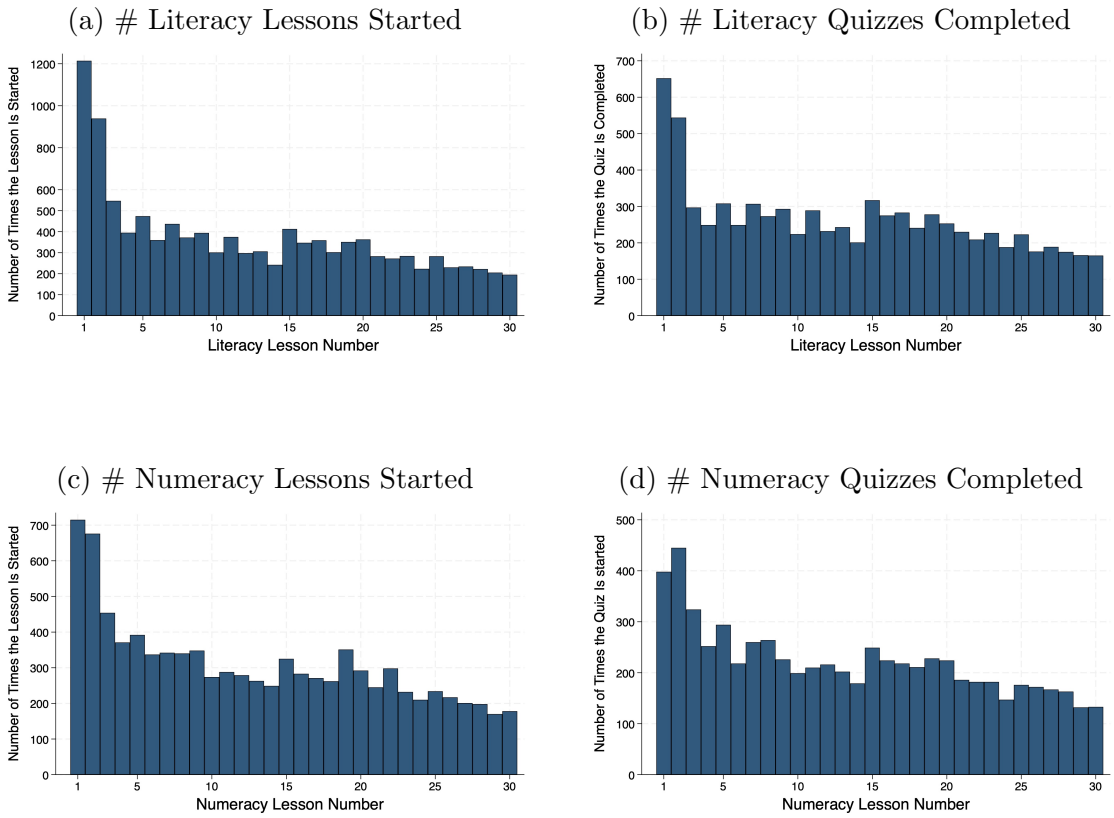
self-reported data, in part because a large portion of the calls made to the IVR server cannot be traced back to individual child IDs. Nearly 40% of the total call time in the server is concentrated in calls from phone numbers that cannot be matched to child IDs; self-reported lessons do not suffer from this misattribution property.

The server data allows us to develop an alternate measure of lesson completion. Dataset `server_phone_quiz.dta` contains variables `num_quizanswer_1` through `num_quizanswer_30` and `lit_quizanswer_1` through `lit_quizanswer_30`. A row of this dataset represents a single phone call. Apart from a brief pilot period in the first few weeks of the IVR program's rollout, it appears that these variables take string value "null" if a given lesson is not accessed on that call. Likewise, string value "" apparently indicates that a lesson was accessed on that call, but that no answer was provided to the quiz at the end of that lesson, whereas some nonempty string response other than "null" indicates the keypad response that was provided to the quiz at the end of that lesson. Taking the pilot rollout into account, we thus code a lesson in the server data as *taken* if the string value for that lesson is not "null", and code a lesson as *completed* if the string value for that lesson is neither "null" nor "". Leadership lessons have no surveys, so we cannot identify those individually. Appendix B offers an alternative analysis path for the leadership lessons.

Even if students cannot always be identified using the server data, the server data allows us to accurately estimate the aggregate number of times a given lesson has been started or completed on the IVR server. Figure 4 shows how often specific lessons have been started or finished in the server data. E.g., Figures 4a and 4b respectively show that the 30th and final literacy lesson is started 195 times, with the quiz being completed 165 times.

These completion rates are incompatible with the patterns presented in the original Appendix Figures B3 and B4, which display lesson completion numbers/rates for caregiver-child dyads.[8] Both figures indicate that a majority of the treated stu-

[8]We have 1763 unique `CHILD_IDs` and 1755 unique `RECORD_IDs`, the latter of which we believe identifies individual households. It is therefore unlikely that the discrepancy is explained by siblings listening to the same phone call.



Note: We estimate the number of times each lesson is started (quiz is completed) by looking at the associated quiz answers in the server data. If the answer is an empty string, we interpret the lesson as having been started, if the answer is a non-empty or "null", we interpret the quiz as completed. This approach is applied across all 30 literacy and numeracy lessons.

Figure 4: Lesson Completion, Constructed from Server Data

dents complete most of the lessons; the note under Figure B3 states: "...476, 454, and 276 dyads completed all the lessons for Literacy, Numeracy and Leadership, respectively." It is impossible that close to 500 children completed all lessons when fewer than 200 started the last literacy lesson, as shown in Figure 4a. Similar inconsistencies extend to numeracy and leadership lessons. Furthermore, we are largely able to reconstruct the time spent per week results, as depicted by the original Appendix Figure B5, despite the lack of replication code. The Figures (original and our reproduction from the server data) can be found in the Appendix Figure A3.

These differences between lesson completion in self-reports and the server are consistent with social desirability bias, and indicate meaningful risks of measurement error. Given that survey respondents are well-known to overreport positive characteristics about themselves, it is plausible that parents overreport the number

of lessons that their children complete, especially given the fact that parents are expected to complete those lessons alongside their children. This adds a caveat to our analyses earlier in this section, as such inaccurate reporting would constitute a source of measurement error, which would in turn attenuate estimated dosage responses to additional lesson-taking. These results also cast doubt on the high aggregate takeup levels claimed in the paper, despite the fact that more accurate aggregate takeup estimates were readily available for REM24 from the server data.

4.4 General Knowledge Scores and Reporting Accuracy

REM24 do not accurately describe aspects of the data collection. For instance, while the paper claims that data on 'general knowledge' was never collected, we observe it in the paper's second replication package, and treatment effects on this variable are analyzed in an earlier working paper draft of this article (Islam et al. 2022). From REM24 pg. 575:

> "One minor deviation from the preanalysis plan relates to the assessment of children's learning outcomes. In the preanalysis plan, we had specified the inclusion of a general knowledge component, as this is part of the national curriculum. However, as we finalized the material, we decided not to cover general knowledge questions in the assessment, as general knowledge was not covered in the intervention lesson plans. Consequently, the general knowledge component is not included in our measure of children's learning outcomes."

Data on general knowledge scores was provided in the second replication package.[9] A prior working paper draft of the published article also analyzes treatment ef-

[9]Repository dataset `Anushilon-Endline-Assessment.xlsx` contains variables `anu_e1_cog_5` through `anu_e1_cog_8`. On lines 474 and 475 of cleaning code file `Endline-data-clean-up.do`, these four variables are summed using Stata's `rowtotal()` function to produce variable `anu_e1_cog_gk`, which is labeled "All Grade - General Knowledge". When creating final analysis dataset `IVR_Data.dta`, cleaning code file `Data-Clean-Merge.do` runs a `keep` command on line 318 which drops all variables not included in the list. `anu_e1_cog_gk` is excluded from the list, dropping this general knowledge score from the `IVR_Data.dta` file presented in the first replication repository.

fects on general knowledge, finding treatment effects that are highly statistically significant and exceeding 0.5 standard deviations (Islam et al. 2022).

Figure 5 closely replicates the treatment effects on general knowledge reported by Islam et al. (2022). Treatment is estimated to increase general knowledge by more than 0.5 standard deviations, an effect size similar to that observed for other outcomes. A table version of this figure can be found in Appendix Table A1, showing that these effects are highly significant, with t-statistics exceeding seven. This is despite REM24's claim that "general knowledge was not covered in the intervention lesson plans" (pg. 575).

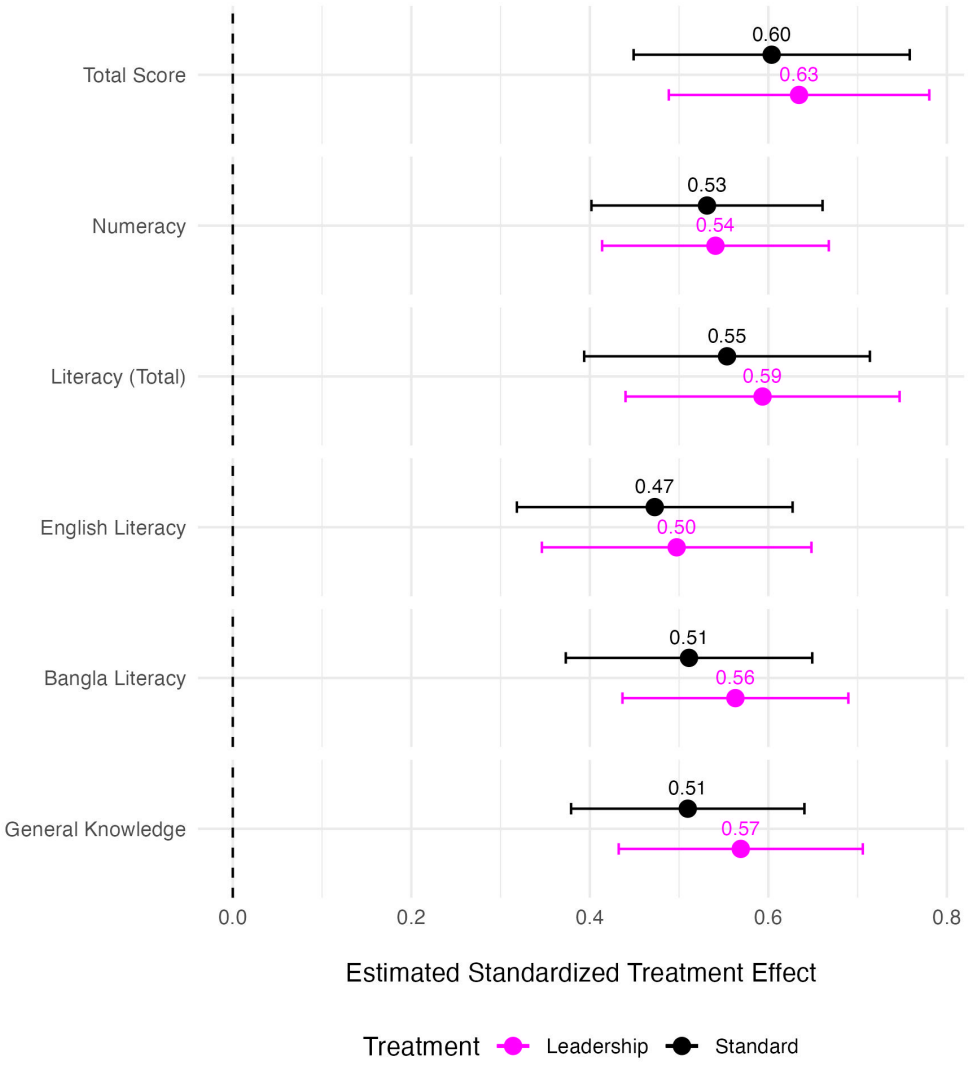
5 Documentation and Data Inconsistencies

We note two types of documentation inconsistencies that cast doubt on REM24. These concern i) study procedures and ii) data collection. Additionally, we highlight one large data inconsistency concerning parental ages.

5.1 Procedural Inconsistencies

First, the randomization file `Randomization-code-treatment-assignment.do` provided by the authors in the third replication package reveals that treatment assignment was stratified by parental education and family income, a procedure not disclosed in the published paper, pre-analysis plan, or other documentation. Due to a missing GDRI dataset from which the initial sample of 90 villages is drawn, and without the authors setting a seed in their randomization code, we cannot verify that the treatment assignment in the paper matches this randomization code.

Second, the published version of REM24 mentions SMS charges being part of the variable costs of the project (pg. 595); however, no use of text messages was pre-registered or acknowledged in the paper's documentation. When reached for clarification, the authors confirmed that the participants received regular text messages during the intervention. For details, see Appendix C.1. Thus, the paper fails to disclose the use of regular lesson reminders, which alters the interpretation of the



Note: Treatment effects are represented in units of outcome standard deviations in the control group. This figure replicates parts of Figure 4 from Wang et al. (2024a) and Figure 3 from Islam et al. (2022). Estimates are from a linear model with controls and fixed effects equivalent to those used in the specifications that produce Figure 4 Wang et al. (2024a). 95% confidence intervals are visualized based on standard errors clustered at the village level.

Figure 5: Estimated Treatment Effects on Standardized Outcomes

treatment.

5.2 Data Collection Inconsistencies

First, according to the published paper, the authors asked the households to "record the lessons they completed on a printed sheet to keep track of lesson completion" (pg. 573). However, this type of data collection is not mentioned anywhere in the pre-analysis plan. The pre-analysis plan only mentions using phone usage data as a measure of lesson completion, or relying on "assistants" to collect "some users'

usage information".

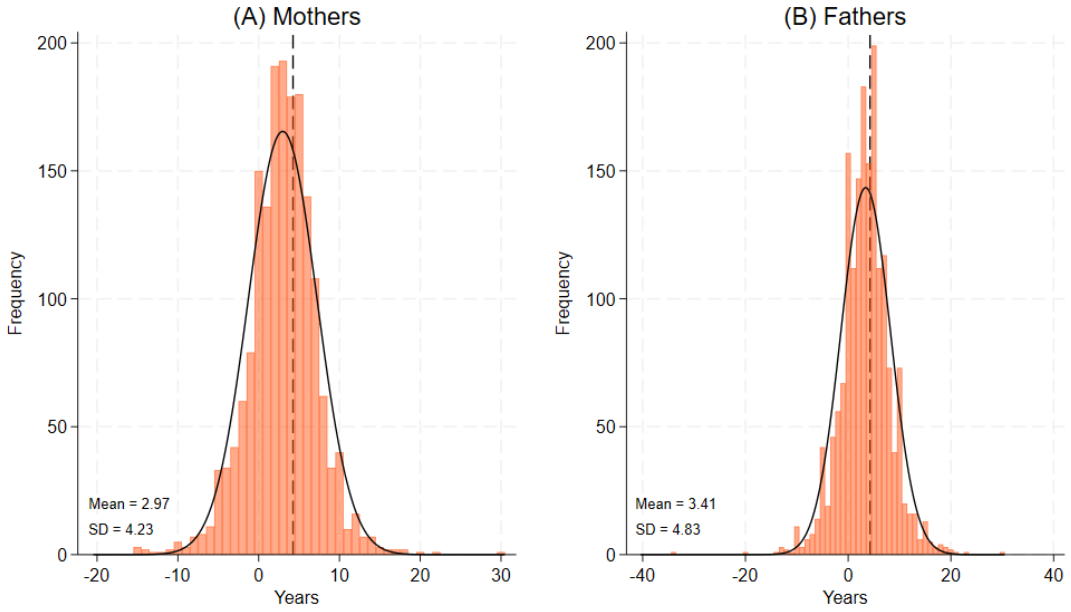
Second, the AEARCT registry pre-analysis plan, written after baseline data collection, mentions that the (baseline) sample consisted of 1741 children. This is contradicted by Table 2 in the pre-analysis plan, as columns 1, 2, and 3 do not sum up to column 4 as implied: $574+586+577 \neq 1741$. These counts and this total sum are likewise inconsistent with the baseline data in the replication packages, which contain 1763 observations (split 596 - 586 - 581).

Third, we document multiple inconsistencies between the survey materials written in Bangla, Kobo forms, and their English translations/data documentation in the paper, its appendix, pre-analysis plan, and data sheets. Broadly, these fall into two categories: 1) The question text or answer options in the Kobo forms and/or survey materials do not match each other or mismatch other documentation in ways that significantly changes their meaning or interpretation or suggests the documentation is incorrect. 2) Errors in translation. See Appendix Tables A3, A4, and A5 for details. As a related issue, the psychological scales used may be age-inappropriate, as they are traditionally used on adolescents or adults. In some cases, the original questions seem ill-suited for children as young as six years old, raising concerns about how to interpret the responses. Moreover, our translations of the Bangla materials suggest that several questions were modified – presumably to increase age-appropriateness – but these changes were not documented in the paper or its appendices. For details, see Appendix C.3.

Fourth and finally, the KoboToolbox survey file and the raw data provided for the endline assessment are not consistent with one another. For details, see Appendix C.2.

5.3 Data Inconsistencies

REM24 use two sources of data for their baseline test scores and demographic variables. One is a baseline survey collected in May 2021 and referred to as the Rapid Baseline Survey in the replication folder. The other is collected in February



Note: This figure shows the distribution of the differences between (A) mother's and (B) father's reported ages for observations with the same `CHILD_ID` reported in the 2021 Rapid Baseline Survey and in the 2017 data of the IIOF project. We exclude six observations where the mother's age was imputed and 24 where the father's age was imputed. We append the figure with an approximate normal distribution. The vertical dashed line is placed at 4.25, indicating the number of years that have actually passed between the IIOF survey and the Rapid Baseline Surveys.

Figure 6: Parents' Age Differences Between Rapid Baseline and Previous Project

2017 for a previous project called Investing in our Future (IIOF), of which the REM24 data are a subsample.[10] Notably, both datasets include information about mother's and father's age.

Comparing the parents' ages from the two sources highlights a discrepancy. While one would expect the parents to age by the number of years roughly equal to the time difference between the two surveys (about four years), this is not what we observe. Figure 6 shows histograms of the individuals' age difference between the two surveys for mothers and fathers. These patterns of aging in the data are

[10]The demographic information for IIOF was collected in February 2017, and the endline test scores for IIOF were collected in January 2019. While the authors do not specify when exactly the information from IIOF was collected, the demographic data can be traced to a 2017 socioeconomic survey dataset resulting from a previous project and published at UK Data Service (Islam (2022)). The test scores can be traced to endline files in the same data package. In a working paper by Guo et al. (2024), the baseline of this project is further specified to have happened in February 2017, and the endline in January 2019. To our knowledge, this working paper is no longer publicly available as of late March 2025. Raw data for the REM24 Rapid Baseline Survey and selected variables from the previous project's data are provided in the second replication package (Wang et al. 2025). The raw data from the previous project were not provided in the first replication package.

	(1) Mother's Age (2017)	(2) Father's Age (2017)	(3) Mother's Age (2021)	(4) Father's Age (2021)
Standard Treatment	0.79	0.74	-0.27	0.06
	(0.352)	(0.422)	(0.392)	(0.460)
	[0.027]	[0.085]	[0.489]	[0.892]
Extended Treatment	0.71	0.71	-0.29	0.15
	(0.417)	(0.487)	(0.390)	(0.512)
	[0.090]	[0.147]	[0.461]	[0.774]
Constant	26.62	33.82	30.28	37.73
	(0.277)	(0.312)	(0.276)	(0.332)
	[0.000]	[0.000]	[0.000]	[0.000]
N	1689	1689	1684	1667

Note: Linear regression estimates are presented alongside robust standard errors clustered at the village level in parentheses and p-values in brackets. Columns 1 and 2 report baseline balance for the age of mother and father collected during the IIOF project in 2017. Columns 3 and 4 report baseline balance for the age of mother and father collected during the Rapid Baseline in 2021. Columns 3 and 4 exclude six and 24 observations for which the age was imputed. Consistent with REM24's practice in reporting baseline balance, only participants who completed the endline are included.

Table 2: Age Discrepancies and Variation with Treatment Assignment

difficult to explain even if rounding or estimating one's age are common in the sample. Hundreds of people report that they have not aged at all or that they have become younger, despite the fact that four years and three months have passed. Only about 20% of the mothers and fathers report to be four or five years older, and the mean age difference is 2.97 years for mothers and 3.41 years for fathers.

Baseline treatment imbalance in parental age differs depending on whether one uses the data from the Rapid Baseline Survey or from the IIOF survey. When using the newer data from the Rapid Baseline Survey, the parents' ages are balanced at baseline across the treatment conditions.[11] On the other hand, when using the data from the IIOF project, parental ages are no longer fully balanced across the treatment groups. In particular, mothers in the control group are 0.79 years younger than those in the standard treatment ($p = 0.027$), and are 0.71 years younger than those in the extended treatment ($p = 0.09$). These results are reported in Table 2.

[11]REM24 (pg. 574) report baseline balance across 13 variables. They only find statistically significant differences between some of the treatment groups for two variables, religion and access to private tuition, both of which were collected as part of the Rapid Baseline Survey. Baseline balance of parents' ages across treatments is not reported.

	(1) Mother's Age Difference	(2) Father's Age Difference
Standard Treatment	-1.11	-0.69
	(0.373)	(0.393)
	[0.004]	[0.082]
Extended Treatment	-1.02	-0.67
	(0.372)	(0.380)
	[0.008]	[0.082]
Constant	3.69	3.87
	(0.279)	(0.266)
	[0.000]	[0.000]
N	1756	1738

Note: Linear regression estimates are presented alongside robust standard errors clustered at the village level in parentheses and p-values in brackets. Column 1 presents the effect of the treatment on the difference in mother's reported age between the 2021 Rapid Baseline and the 2017 data from the IIOF project. Column 2 reports the treatment effect on the difference for father's age. We exclude six observations where the mother's age was imputed and 24 where the father's age was imputed.

Table 3: Parental Age Discrepancies and Treatment Assignment

When comparing the age difference between the two data sources by treatment group, we also find statistically significant differences. The mother's age difference between the rapid baseline and the previous project data is larger by about 1 year for the control group than it is for both the standard and the extended treatment. These differences are statistically significant with $p = 0.004$ and $p = 0.008$ (respectively). These estimates are reported in Table 3. The father's age difference is about 0.7 years larger for the control group than for the standard and extended treatments $p = 0.082$. Because both of the age variables used to construct the difference were collected before the treatment was administered, it is not plausible that the treatment itself influenced the age reporting.[12]

6 Conclusion

Taken together, the array of anomalies detailed in this paper collectively undermines its credibility. This conclusion extends beyond the mere tallying of specific

[12]In Appendix Table A2, we show that these age discrepancies are not statistically significantly related with the education of mothers or fathers, nor with household income.

problems; it relates fundamentally to the broader principle of scientific reliability at every stage from data collection through final analysis. Unfortunately, the cumulative weight of inconsistencies, discrepancies, and unexplained irregularities identified in this analysis fundamentally compromises that trust, thereby weakening confidence in the conclusions drawn from this study.

References

Angrist, N., Bergman, P. and Matsheng, M.: 2022, Experimental evidence on learning using low-tech when school is out, *Nature Human Behaviour* **6**(7), 941–950.

Aparicio, J. P., Cook, N., Mikola, D., Rogeberg, O., Wiebe, M., Valenta, D. and Brodeur, A.: 2025, Comment on "Telementoring and homeschooling during school closures: A randomized experiment in rural Bangladesh" by Hassan et al. I4R Discussion Paper 240.

Bernard, D. R., Bryan, G., Chabé-Ferrett, S., De Quidt, J., Fliegner, J. C. and Rathelot, R.: 2024, How much should we trust observational estimates? accumulating evidence using randomized controlled trials with imperfect compliance. Working Paper.

Brodeur, A., Mikola, D., Cook, N. et al.: 2024, Mass reproducibility and replicability: A new hope. I4R Discussion Paper 107.

Carlana, M. and La Ferrara, E.: 2021, Apart but connected: Online tutoring and student outcomes during the COVID-19 pandemic. CEPR Discussion Paper No. DP15761.

Cattaneo, M. D., Crump, R. K., Farrell, M. H. and Feng, Y.: 2024, On binscatter, *American Economic Review* **114**(5), 14881514.

Crawfurd, L., Evans, D. K., Hares, S. and Sandefur, J.: 2023, Live tutoring calls did not improve learning during the COVID-19 pandemic in Sierra Leone, *Journal of Development Economics* **164**, 103114.

Duckworth, A. L. and Quinn, P. D.: 2009, Development and validation of the Short Grit Scale (GRIT–S), *Journal of Personality Assessment* **91**(2), 166–174.

Dweck, C. S., Chiu, C.-y. and Hong, Y.-y.: 1995, Implicit theories and their role in judgments and reactions: A word from two perspectives, *Psychological Inquiry* **6**(4), 267–285.

Gortazar, L., Hupkau, C. and Roldán-Monés, A.: 2024, Online tutoring works: Experimental evidence from a program with vulnerable children, *Journal of Public Economics* **232**, 105082.

Guo, K., Islam, A., List, J., Vlassopoulos, M. and Zenou, Y.: 2024, Early childhood education, parental social networks, and child development, *CDES Working Paper No.* **15/24**.

Hassan, H., Islam, A., Siddique, A. and Wang, L. C.: 2024a, Replication package for: "Telementoring and homeschooling during school closures: A randomized experiment in rural Bangladesh".
URL: *https://doi.org/10.5281/zenodo.10696443*

Hassan, H., Islam, A., Siddique, A. and Wang, L. C.: 2024b, Telementoring and homeschooling during school closures: A randomised experiment in rural Bangladesh, *The Economic Journal* **134**(662), 2418–2438.

Ho, J. and Thukral, H.: 2009, Tuned in to student success: Assessing the impact of interactive radio instruction for the hardest-to-reach, *Journal of Education for International Development* **4**(2), 34–51.

Islam, A.: 2022, The Early Childhood Intervention and Parental Involvement in Bangladesh, 2016-2021. [Data Collection].

Islam, A., Wang, L. C. and Hassan, H.: 2022, Delivering Remote Learning Using a Low-Tech Solution: Evidence from an RCT during the Covid-19 Pandemic. https://docs.edtechhub.org/lib/FE3VBQQW (accessed February 16, 2025).

Kraft, M. A., List, J. A., Livingston, J. A. and Sadoff, S.: 2022, Online tutoring by college volunteers: Experimental evidence from a pilot program, *AEA Papers and Proceedings*, Vol. 112, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, pp. 614–618.

Muralidharan, K., Romero, M. and Wüthrich, K.: 2025, Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments, *Review of Economics and Statistics* pp. 1–16.

Overgaauw, S., Rieffe, C., Broekhof, E., Crone, E. A. and Güroglu, B.: 2017, Assessing empathy across childhood and adolescence: Validation of the empathy questionnaire for children and adolescents (EmQue-CA), *Frontiers in Psychology* **8**, 1–9.

Rammstedt, B., Grüning, D. J. and Lechner, C. M.: 2022, Measuring growth mindset, *European Journal of Psychological Assessment* **40**(1).

Tsukayama, E., Duckworth, A. L. and Kim, B.: 2013, Domain-specific impulsivity in school-age children, *Developmental Science* **16**(6), 879–893.

Wang, L. C., Vlassopoulos, M., Islam, A. and Hassan, H.: 2024a, Delivering remote learning using a low-tech solution: Evidence from a randomized controlled trial in Bangladesh, *Journal of Political Economy: Microeconomics* **2**(3), 562–601.

Wang, L. C., Vlassopoulos, M., Islam, A. and Hassan, H.: 2024b, Replication Data for: "Delivering Remote Learning Using a Low-tech Solution: Evidence from a Randomized Controlled Trial in Bangladesh".
URL: *https://doi.org/10.7910/DVN/GB61G4*

Wang, L. C., Vlassopoulos, M., Islam, A. and Hassan, H.: 2025, Additional Replication Data for: "Delivering Remote Learning Using a Low-tech Solution: Evidence from a Randomized Controlled Trial in Bangladesh".
URL: *https://doi.org/10.7910/DVN/IIOU6U*

Online Appendix

A Lesson Completion and Scores by Grade

REM24's Appendix Table A2 describes the first five lessons of the numeracy module as covering (1) counting, (2) counting and addition, (3) addition, (4) subtraction, and (5) addition and subtraction. The sixth lesson is the first to go beyond addition and subtraction, introducing decomposition. Based on REM24's Appendix Table A4, for first-grade students, three of the five numeracy questions on the endline assessments test integer addition and subtraction, whereas only one of the five numeracy questions are on integer addition or subtraction for the higher grades.[1] Therefore, we should not expect sharp jumps in test scores after the first handful of lessons for students in second through fourth grade. However, Figure A1 shows that test scores in all grades exhibit sharp jumps before students complete the fifth lesson, with no meaningful improvement in test scores after completing any further lessons. This figure is constructed with a 'highest lesson completed' variable, constructed by taking the highest lesson number a student completed based on variables `complete_num_1` through `complete_num_30`.

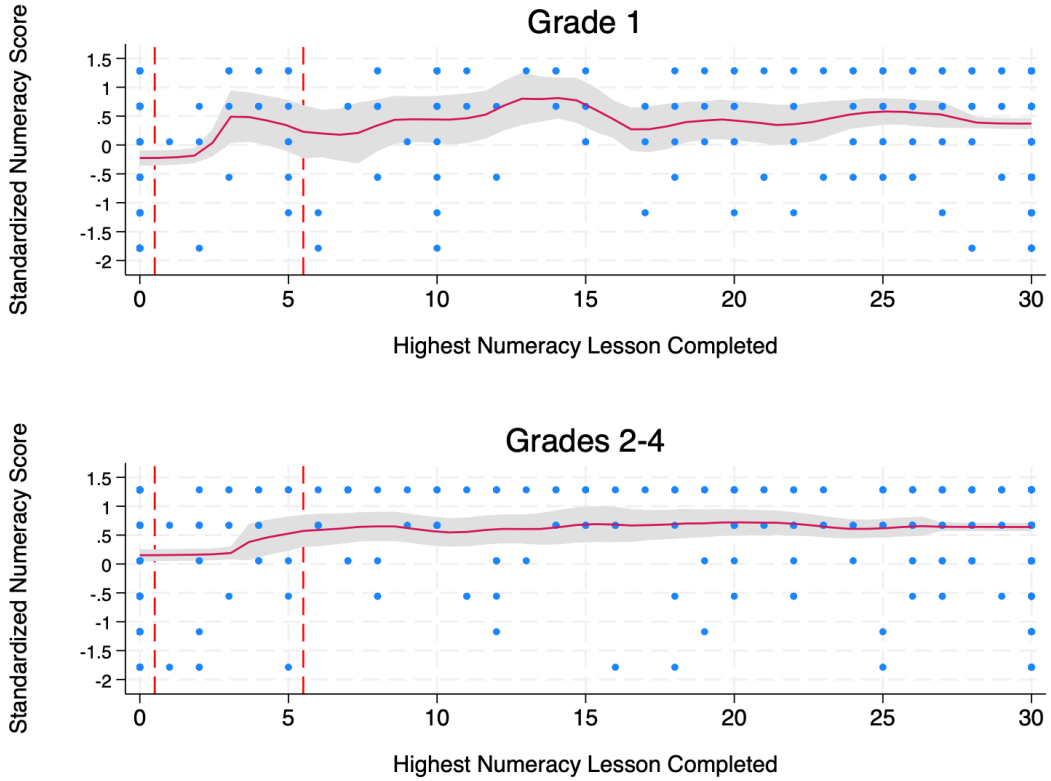
B Leadership lessons

We lack quiz data to investigate students' completion of leadership lessons. The best we can do is to investigate the amount of time spent on lessons per week. This has meaning if the students largely follow the suggested schedule, which we can check with the literacy and numeracy lessons. Literacy and numeracy lessons are assigned for each week in two-lesson intervals – e.g., lessons one and two are recommended for the first week, lessons three and four for the second week, etc. The server suggests that many dyads followed this recommended schedule; for example, in week 14, lessons 27 and 28 make up 64% of all literacy listens.

Therefore, relying on students to roughly follow the plan, we assume that all

[1] The specific question assessed to second graders is a word problem. Third and fourth graders are given a second addition problem, but answering it requires knowledge of fractions.

Figure A1: Highest Lesson Number Completed and Endline Numeracy Scores by Grade

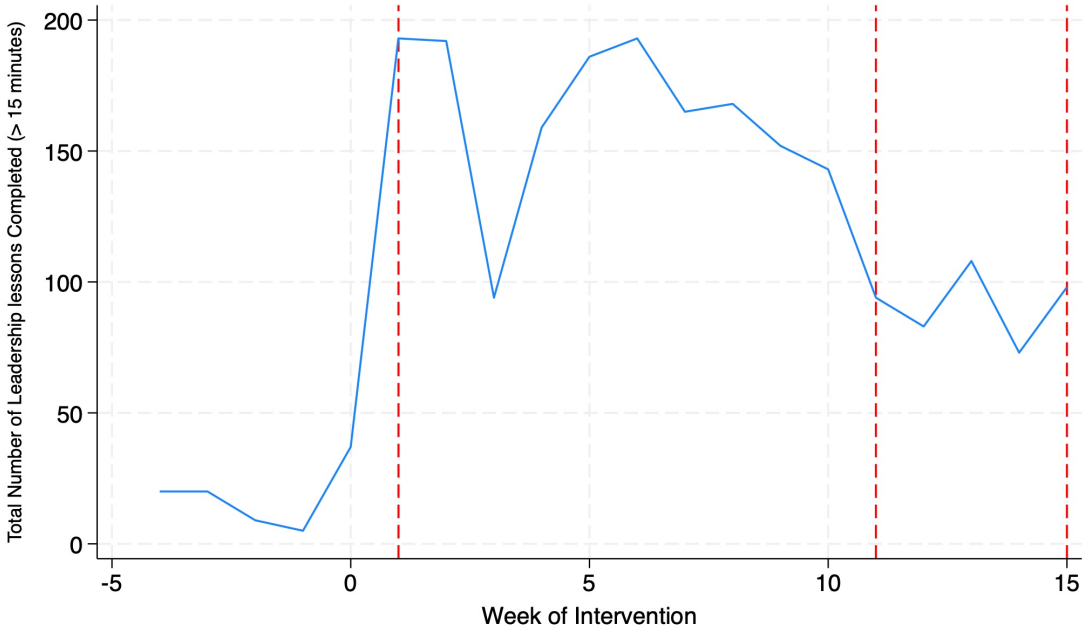


Note: Local polynomial estimates of the relationships between highest completed numeracy lesson and standardized endline numeracy scores are displayed alongside 95% confidence intervals in samples restricted by grade. Dots indicate observed combinations of highest completed numeracy lesson and standardized endline numeracy scores. The dashed red vertical lines are placed at 0.5 and 5.5, immediately before the extensive margin of lesson completion and the completion of the fifth lesson (respectively). The left and right dashed red vertical lines respectively indicate (1) the extensive margin of observations who (are assigned to) take one or more lessons and (2) observations who report completing more than the fifth lesson.

the remaining calls that are unassigned to any literacy or numeracy unit and which last at least 900 seconds (15 minutes) are then leadership lessons.[2] As depicted by Figure A2, we observe a very similar pattern to that of the total duration of calls per week: there are more listens in the beginning of the intervention and fewer towards the end of it. We observe approximately 100 listens in the last week, which corresponds to about half of the literacy and numeracy listens in the same week. Given half of the treated were assigned to the treatment with leadership lessons,

[2]We need to assume a minimum length because phone calls where no lesson has yet selected also look unassigned. There are many short calls in the dataset, and the lessons are described to be 16-18 minutes in length in the original paper.

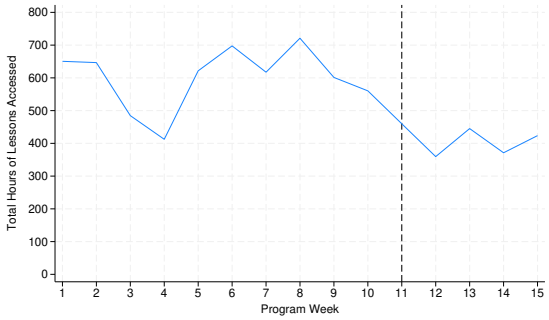
Figure A2: Estimated Number of Leadership Lessons Finished by Week



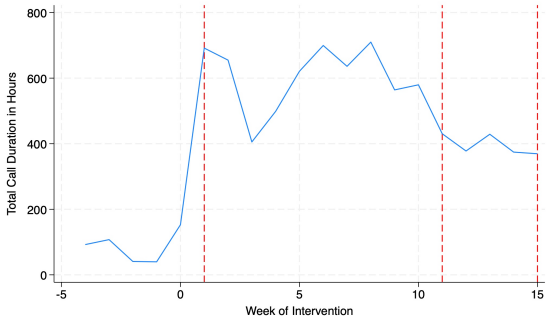
Note: We estimate the number of leadership lessons taken per week by counting unassigned lessons that last more than 15 minutes. This figure is based on the server data.

Figure A3: Original and Reconstructed Duration of Calls per Program Week

Panel A: First Replication Package



Panel B: Server Data



Note: Panel A: Total lesson hours by week, computed using the same code used to reproduce Figure B5 in Wang et al. (2024a). Panel B: Total call duration by week, using the server data from the third replication package. Note that the server data starts five weeks before the intervention – we have added dashed vertical lines at weeks 1, 11, and 15 to aid the comparison.

the pattern seems plausible.

C Additional Inconsistencies

C.1 Intervention Description When reached for clarification regarding the use of SMS in the intervention, the authors stated:

SMS messages were used during the intervention. There were three types

of SMS messages sent on Friday and Wednesday during program weeks:

1. *The first day of each program week (Friday): SMS message informing participants about the lesson numbers of each module that week.*
2. *The sixth day of each program week (Wednesday): SMS message informing participants about the closing for answering quizzes to be eligible for the quiz prize.*
3. *The day after each program week (Friday): SMS message informing quiz winners that they won the quiz and mobile credit would be disbursed via bKash (MFS) in the coming days.*

Note that this explanation is consistent with the text in the working paper (Islam et al. 2022), which additionally mentions reminder *phone calls*, particularly targeting caregivers whose children participated in the program on an "irregular basis" (page 48). Taken together, the program description in the published paper is incomplete.

C.2 Mismatches Between Kobo Forms and Raw Data The KoboToolbox survey file and the raw data provided for the endline assessment are not consistent. The raw data file `Anushilon-Endline-Assessment.xlsx` includes the variable `CHILD_ID`, whereas the Kobo form in `Endline_Assessment_Kobo_Form.xlsx` does not include this variable. Therefore, the raw data file could not have been produced by the form provided. Omission of `CHILD_ID` is not explained by standard removals of personal information. Furthermore, the survey does not include any other type of respondent or participant ID that could be directly used to merge this data in a straightforward way to the remainder of the project data. This was verified using an official translation of the survey questions included in the provided Kobo form.

C.3 Assessments Use Drawing on the appendix of the original paper, we find that there are doubts about the suitability of four assessment scales for children in this study, all of whom are in first through fourth grade, and most of whom are between ages six and nine. In particular, these scales were designed and validated for older children, adults, or teachers, and their use thus deviates from typical use in the literature.

However, as we note in Appendix Tables A3, A4, and A5, it appears that the authors implemented different versions of some of these questions. As a result, even if the formulations are more age-appropriate, the subsequent scales are no longer validated or consistent with the literature.

Impulsivity Scale for Children. Tsukayama et al. (2013) investigate impulsivity in middle school students, with a mean age of 12 to 13 years. It is unclear whether seven-year-old children would be able to answer the questions with adequate self-reflection, similar to older children. Furthermore, many questions are set in the classroom, which may be particularly unsuitable during pandemic conditions.

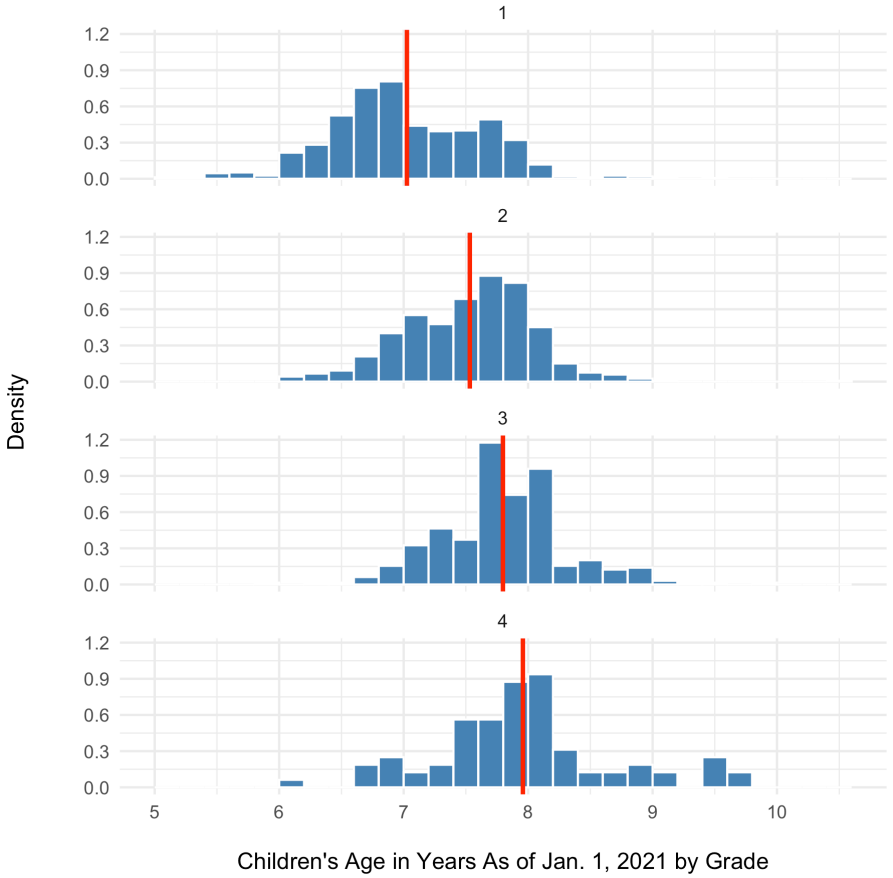
Grit. Duckworth and Quinn (2009) validate the shorter eight-item grit questionnaire on adolescents at youngest. Based on our translations, several grit questions are systematically modified, changing "ideas/projects" to "tasks/games" or "activity/sport" (questions 1, 3, 6) and adding "tasks and games" in the otherwise vaguely formulated question 7. These activities have very different time-spans compared to the original formulations.

Growth Mindset Scale. Dweck et al. (1995) develop the three-item scale and study its properties in a group of people that is not specified to be children. We thus consider it reasonable to assume that it was developed for adults. The three questions in this scale are again difficult for the target group of this study, seven to ten year old children. Rammstedt et al. (2022) validate the measure on adults and, at the youngest, 14-year-old adolescents.

Empathy Questionnaire for Children and Adolescents. Overgaauw et al. (2017) validate the questionnaire with 10-15 year old children, stating: "Taken together, we show that the EmQue-CA is a reliable and valid instrument to measure empathy in typically developing children and adolescents aged 10 and older." This does not indicate validity of the scale for the youngest cohorts of REM24's participants.

C.4 Grades and Student Age We also note that the children in this data do not seem to age in accordance with grade progression. Using grades recorded during 2021, the average age (measured on January 1 2021) in grades 1, 2, 3, and 4 is 7.0,

Figure A4: Distribution of Student Age by Grade



Note: Distribution of children's age in years by grade. Grades recorded in 2021 are displayed at the top of each subfigure. Vertical bars denote average ages in each grade.

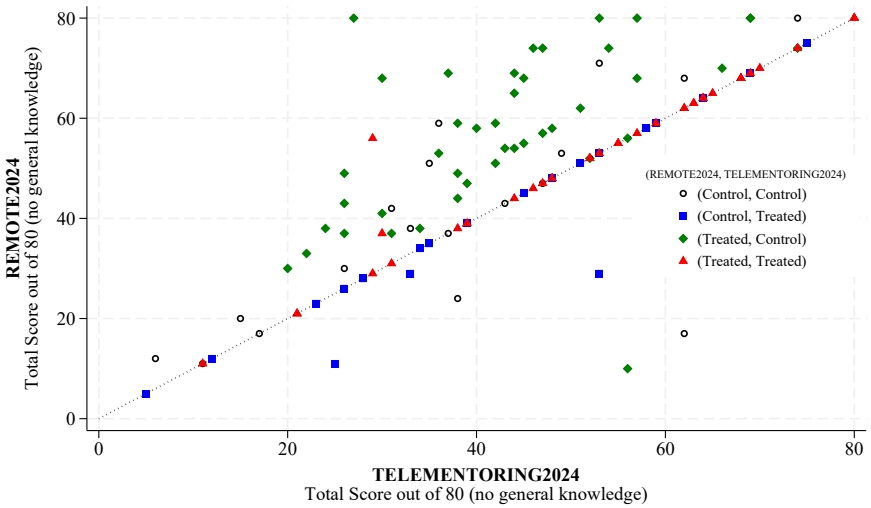
7.5, 7.8, and 8.0 (respectively); see Figure A4.[3] Hence, students in adjacent grades differ in age by 0.2 to 0.5 years, on average. This difference is far smaller than the expected age gap of one year between adjacent grades. Note that this pattern is not explained by students repeating a grade, as a repeater increases the average age in both the repeated grade and in the following grade (when they do advance).

[3]Grades were recorded during the May 2021 baseline survey; school closures were in effect since March 2020. Because the school year starts in January, it is not clear how a student's enrolled grade is determined. In particular, it is unclear whether students who missed most of 2020 were advanced to the next grade or forced to repeat it.

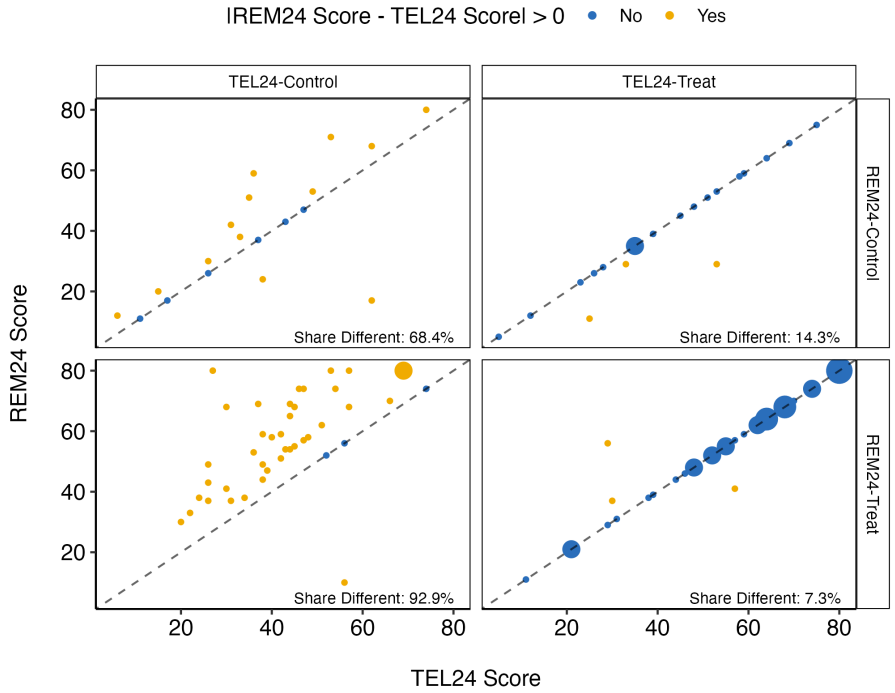
D Appendix Figures

Figure A5: Comparing Baseline and Endline Scores Across TEL24 and REM24

(a) Test Score Variation Across Papers by Treatment Groups



(b) Separated by Treatment Groups and Showing Density of Individuals



Note: This figure reproduces Figure 1 in Aparicio et al. (2025). Scores from the endline of REM24 and the second endline of TEL24 are used. Y-axes represent the total score from the REM24 replication data; x-axes represent the same score reported in the TEL24 replication data. Each marker in the top figure represents a person who appears in both datasets. The bottom figure concentrates similar score levels into bubbles: the larger the bubble, the more individuals with that score level. Treated participants in REM24 are those in either the standard or leadership treatments.

Figure A6: REM24 Endline Assessments

Table A4. Endline assessment test questions

Subject	No	Level 1 (Grade 1)	Level 2 (Grade 2)	Level 3 (Grade 3 & 4)
Literacy	1.	Read aloud the following letters (the first 4 letters from the Bengali alphabet)	Make two words using the Bangla letter ----.	Read aloud this following paragraph (Bangla).
	2.	Fill in the gaps (5 Bangla letters with 2 gaps).	Fill in the gap (a line in Bangla from the textbook)	What is the antonym of the Bangla word (FREEDOM)?
	3.	Make a word with the Bengali letter -----.	What is the spelling of the word (Sundarbans)?	What is the spelling of the word (Bangla of freedom fighter)?
	4.	What is the spelling of (Bengali word)?	What is the antonym of the Bangla word (high)?	What is the meaning of this Bangla word (Bangla word from the textbook)?
	5.	Read the following word (CAP).	Read the following word (FARMER)?	Read aloud this following paragraph (English).
	6.	Answer this English question: What is your name?	Answer this English question: How old are you?	Answer this English question: What month is it now?
	7.	Say the English of Bangla word – (DOOR).	Say the English of Bangla word – (WINDOW).	Say the English of Bangla word – (FARMER).
	8.	Say the English of Bangla word – (BOOK).	Say the English of Bangla word – (UMBRELLA).	Say the English of Bangla word – (WEDNESDAY).
	9.	Say the English of Bangla word – (DOG).	Say the English of Bangla word – (BREAKFAST).	Say the English of Bangla word – (FLAG).
	10.	Spell your name in English.	Read and say the name of these shapes (picture of the square, circle, triangle, and rectangle).	Match the appropriate description with this picture (match from 4 options).
Numeracy	11.	Which number comes after 6? Does it even or odd?	Name the even numbers between 1 and 10.	Sort these three numbers, smallest to the largest (20, 73, 10, 78).
	12.	What is the result of 3+4=?	Sort these three numbers, smallest to largest (23, 17, 38).	There are 6 notes of 20 BDT. How much money is there?
	13.	What is the result of 8-3=?	In a class, there were 16 students. The teacher sends 5 of them for gardening. How many students are left in the classroom?	What is the result of 13+11=?
	14.	How many minutes in 60 seconds?	How many sides a triangle has?	What is the result of 2/4+2/4=?
	15.	What is the result of 6+0=?	There are three fruits on a plate. How many fruits there are in 4 plates?	The price of 5 eggs is BDT 30. How much does it cost to buy 2 eggs?

Note: The test was conducted on a one-on-one basis.

Notes: This image was cropped as the assessment presented in the appendix of Wang et al. (2024a) spanned two pages.

Figure A7: TELEMENTORING Second Endline Assessment

Table B3. Children's learning assessment at the one-year endline

Subject	No	Level 1	Level 2	Level 3	Marks
English Literacy	1.	Read the following word (CAP)?	Read the following word (FARMER)?	Read aloud this following paragraph (English)?	6
	2.	Answer this English question: What is your name?	Answer this English question: How old are you?	Answer this English question: What month is it now?	6
	3.	Say the English of Bangla word – (Door).	Say the English of Bangla word – (Window).	Say the English of Bangla word – (FARMER).	4
	4.	Say the English of Bangla word – (Book).	Say the English of Bangla word – (UMBRELLA).	Say the English of Bangla word – (WEDNESDAY).	4
	5.	Say the English of Bangla word – (Dog).	Say the English of Bangla word – (BREAKFAST).	Say the English of Bangla word – (FLAG).	4
	6.	Spell your name in English.	Read and say the name of these shapes (picture of the square, circle, triangle, and rectangle).	Match the appropriate description with this picture (match from 4 options).	6
Mathematics	7.	Which number comes after 6? Does it even or odd?	Name the even numbers in between 1 and 10.	Sort these three numbers, smallest to the largest (20, 73, 10, 78).	6
	8.	What is the result of 3+4=?	Sort these three numbers, smallest to the largest (23, 17, 38).	There are 6 notes of 20 BDT. How much money is there?	6
	9.	What is the result of 8-3=?	In a class, there were 16 students. The teacher sends 5 of them for gardening. How many students are left in the classroom?	What is the result of 13+11=?	6
	10.	How many minutes in 60 seconds?	How many sides a triangle has?	What is the result of 2/4+2/4=?	6
	11.	What is the result of 6+0=?	There are three fruits on a plate. How many fruits there are in 4 plates?	The price of 5 eggs is BDT 30. How much does it cost to buy 2 eggs?	6
Bangla Literacy	12.	Read aloud the following letters (first 4 letters from Bangla alphabets)	Make two words using the Bangla letter ----.	Read aloud this following paragraph (Bangla)?	5
	13.	Fill in the gaps (5 Bangla letters with 2 gaps).	Fill in the gap (a line in Bangla from the textbook)	What is the antonym of the Bangla word (FREEDOM)?	5
	14.	Make a word with Bangla letter -----.	What is the spelling of the word (Sundarbans)?	What is the spelling of the word (Bangla of freedom fighter)?	5
	15.	What is the spelling of (Bangla word)?	What is the antonym of the Bangla word (high)?	What is the meaning of this Bangla word (Bangla word from the textbook)?	5
General Knowledge	16.	How many days there are in a week?	Give an example of three red coloured flowers.	On which date of 1952, there was a march for the Bangla language?	5
	17.	What are the days come after Saturday?	What is the first month of Bangla year?	What is victory day in Bangladesh?	5
	18.	Give examples of three fruits.	Which season is best for homemade cakes?	Mostafa Kamal is an ---- (textbook problem).	5
	19.	What is the national bird of Bangladesh?	What was the pet name of the national poet of Bangladesh?	How many days there are in the month 'March'?	5

Second endline assessment for Hassan et al. (2024b).

E Appendix Tables

Table A1: Treatment Effects on Standardized Outcomes

Outcome	Standard	Leadership
Total Score	0.604	0.634
	(0.079)	(0.074)
Numeracy	0.531	0.541
	(0.066)	(0.065)
Literacy (total)	0.554	0.594
	(0.082)	(0.078)
English Literacy	0.473	0.497
	(0.079)	(0.077)
Bangla Literacy	0.511	0.563
	(0.070)	(0.065)
General Knowledge	0.510	0.569
	(0.067)	(0.070)

Note: Estimated treatment effects represented in units of outcome standard deviations in the control group. This figure replicates estimates from Figure 4 from Wang et al. (2024a) and Figure 3 from Islam et al. (2022). Estimates are from a linear model with controls and fixed effects equivalent to those used in the specifications that produce Figure 4 Wang et al. (2024a). Standard errors clustered at the village level are reported in parentheses.

Table A2: Effect of Parental Characteristics and Treatment on Age Differences

	Father's Education (1)	Mother's Education (2)	Family Income (3)	Any Treatment (4)
Panel A:				
Difference in Mother's Age				
Coefficient	−0.002	−0.002	0.00003	−1.065
	(0.026)	(0.035)	(0.00002)	(0.329)
	[0.934]	[0.960]	[0.235]	[0.002]
N	1756	1756	1756	1756
Panel B:				
Difference in Father's Age				
Coefficient	0.008	0.017	−0.00003	−0.680
	(0.027)	(0.038)	(0.00002)	(0.332)
	[0.755]	[0.655]	[0.710]	[0.043]
N	1738	1738	1738	1738

Note: Each column presents results from a separate regression of either mother's or father's reported age difference between the 2021 Rapid Baseline and the 2017 data from the IIOF project on one covariate. All regressions include a constant (omitted from table). Robust standard errors clustered at the village level are reported in parentheses, and p-values in brackets. We exclude six observations where the mothers age was imputed and 24 where the fathers age was imputed.

Table A3: Summary of Inconsistencies in Bangla Surveys: Rapid Baseline Survey

Variable	Explanation
`bs21F6`	The documentation allows for a father's occupation to be a "housewife"; this option is not present in the original survey or the Kobo form.
`bs21M6`	The documentation allows for a mother's occupation to be a "cobbler" or "barber"; this option is not present in the original survey or the Kobo form.
`child_pvt_tuition`, `child_pt_no`, `child_pt_exp`	The documentation does not mention private tutoring groups which the original survey and Kobo form allow for.
`study_helper`	The original survey and Kobo form ask whether the child regularly studies at home. This differs from the documentation ("Does anyone *help* this child in study regularly?").
`helper_rel`	The original survey and Kobo form omit an option for cousins, in contrast to the documentation.
`no_helper`	The original survey and Kobo form question does not match the documentation: "Why doesn't the child sit and read?" vs. "Why no one help this child in study?"

Unless noted otherwise, documentation refers to `Baseline-Rapid-Survey-Data-Layout.pdf`. Original survey refers to `Baseline-Rapid-Survey-Questions-Bangla.pdf`, i.e., the survey text provided in the second replication package (Wang et al. 2025). Kobo form refers to `Baseline-Survey-Kobo-Form.xlsx` that was provided in the third (unpublished) replication package and that we had translated into English.

Table A4: Summary of Inconsistencies in Bangla Surveys: Endline Assessment

Variable	Explanation
anu_e1_cog1_1	The pre-analysis plan question ("Give an example of one Bangla vowel letter.") differs from the question in the appendix/original survey/Kobo form ("Read aloud the following letters (the first 4 letters from the Bengali alphabet)").
anu_e1_cog1_4	The pre-analysis plan question ("What is the English of —— (common flower name)?") differs from the question in the appendix/original survey/Kobo form ("What is the spelling of (Bengali word)?").
anu_e1_cog1_7	The pre-analysis plan question ("Give an example of three flowers.") differs from the question in the original survey/Kobo form ("Can you name any three fruits?"). The paper claims the question was not administered.
anu_e1_cog1_8	The pre-analysis plan question ("What is the national animal of Bangladesh?") differs from the question in the original survey/Kobo form ("What is the name of the national bird of Bangladesh?"). The paper claims the question was not administered.
anu_e1_cog1_9	The pre-analysis plan question ("Make a word with 'C'.") differs from the question in the appendix/original survey/Kobo form ("Read this word out loud.").
anu_e1_cog1_10	The Kobo form differs from the original survey in enumerator instructions in what constitutes a sufficient answer.
anu_e1_cog1_11	The pre-analysis plan question ("Tell the English of Bangla word (Hand).") differs from the question in the appendix/original survey/Kobo form (" I will now tell you some Bengali words; you will tell me their English. Okay? What is the English of 'door'?").
anu_e1_cog1_14	The original survey asks the child to state their name in English; the appendix/pre-analysis plan/Kobo form specify the child should *spell* their name in English.
anu_e1_cog2_1	The original survey and Kobo form differ in the letter that is provided. The appendix and pre-analysis plan are vague and thus allow both options.
anu_e1_cog2_2	The pre-analysis plan question ("Give an example of a word written with joint letters.") differs from the question in the appendix/original survey/Kobo form ("Can you tell which word will be in this blank space?").
anu_e1_cog2_4	The original survey question ("Now you tell me the opposite of higher?") differs from the appendix/pre-analysis plan/Kobo form ("What is the opposite word for 'high'?").
anu_e1_cog2_5	The pre-analysis plan question ("Give an example of five flowers.") differs from the question in the original survey/Kobo form ("Can you name any three flowers of red color?"). The paper claims the question was not administered.
anu_e1_cog2_7	The pre-analysis plan question ("Which season is best for homemade cakes?") differs from the question in the original survey/Kobo form ("When is the festival of eating "Pitha-puli"?"). The paper claims the question was not administered.

anu_e1_cog2_9	The pre-analysis plan question ("Make a word with 'M'.") differs from the appendix/original survey/Kobo form ("Read this word out loud (farmer)").
anu_e1_cog2_10	The original survey contains a copy-paste error: the question duplicates text of a subsequent question.
anu_e1_cog2_12	The pre-analysis plan question ("Tell the English of Bangla word (Rose).") differs from the appendix/original survey/Kobo form ("What is the English of 'umbrella'?").
anu_e1_cog2_13	The original survey and Kobo form indicate the child is supposed to translate from English to Bengali; the appendix and pre-analysis plan indicate the opposite translation.
anu_e1_cog2_14	The pre-analysis plan question ("Spell the English word 'Mother'."). differs from the appendix/original survey/Kobo form ("Now you say from this picture, which shape is it? Speak in English.").
anu_e1_cog2_16	The pre-analysis plan question ("Whether the sum of 3 and 4 is an even or odd number?") differs from the appendix ("Sort these three numbers, smallest to largest (23, 17, 38)."), which further differs from the original survey/Kobo form ("Which of the following numbers is the smallest?").
anu_e1_cog3_1	The pre-analysis plan question ("Make one word and a sentence from that word using the Bangla letter (—)."). differs from the appendix/original survey/Kobo form ("Look at the picture below and read it out loud.").
anu_e1_cog3_9	The pre-analysis plan question ("Make two words with 'C'.") differs from the appendix/original survey/Kobo form ("Read aloud this following paragraph (English).").
anu_e1_cog3_12	The pre-analysis plan question ("Tell the English of Bangla word (Umbrella).") differs from the appendix/original survey/Kobo form ("What is the English of the word 'Wednesday'?").
anu_e1_cog3_14, anu_e1_cog3_15	Questions are flipped in the original survey: The first is a numeracy and the second is a literacy question, contrary to the documentation. Both questions differ between the pre-analysis plan ("Spell 'English Teacher' in English." and "Which number is bigger in 525 and 495?") and the Kobo form/appendix/original survey ("There are four options in this picture. Say which option will replace the question mark? People of what profession do we see in this picture?" and "Arrange the four numbers in the picture in order from smallest to largest.").
anu_e1_cog3_17	The appendix question ("What is the result of 13+11=?") differs from the pre-analysis plan, original survey, and Kobo form ("Whether the sum of 13 and 11 is an even or odd number?").
anu_e1_cog3_18	The pre-analysis plan question ("How many sides a rectangle has?") differs from the appendix, Kobo form, and original survey ("What's the answer to this sum problem?" [Answer: 1]). The Kobo form contains incorrect instructions to enumerators ("Instructions: Show the question sheet to the child. [Answer: 24 and even number.]").
anu_grit2	The pre-analysis plan contains a copy-paste error.
anu_grit6	The original survey/Kobo form question omits that this should relate to "projects that take more than a few months to complete".

anu_isc3	The original survey asks whether the child abused others, while the appendix/pre-analysis plan question asks whether the child was rude. The Kobo form translation is closer to the appendix formulation.
anu_isc8	The original survey and Kobo form question differs from the appendix/pre-analysis plan formulation ("arguing" vs "talking back").
anu_EmQue9	The original survey/Kobo form question seems incorrectly translated from the pre-analysis plan/appendix.
anu_EmQue10	The original survey/Kobo form refers to *happiness* as opposed to laughing in the pre-analysis plan/appendix.
anu_childsocial desirability1	The original survey seems incorrectly translated; the question reads: "Do you wish to abuse someone?" In comparison, the appendix/pre-analysis plan/Kobo form refer to saying something harsh or unkind.
anu_childsocial desirability2	The original survey/Kobo form omits the second half of the question about keeping one's room tidy that is mentioned in the appendix/pre-analysis plan.
anu_childsocial desirability3	The original survey/Kobo form omits the second half of the question that mentions the child might act this way even if not sick. This is in contrast to the appendix/pre-analysis plan.
anu_childsocial desirability10	The original survey question is identical to anu_childsocialdesirability1; this seems to be a copy-paste error.

Unless noted otherwise, documentation refers to Endline-Assessment-Data-Layout.pdf. Original survey refers to Endline-Assessment-Questions-Bangla.pdf, i.e., the survey text provided in the second replication package (Wang et al. 2025). Kobo forms refers to Anushilon-Endline-Assessment.xlsx that was provided in the third (unpublished) replication package and that we had translated into English.

Table A5: Summary of Inconsistencies in Bangla Surveys: Endline Parent Survey

Variable(s)	Explanation
anu_edu_1	Documentations asks about satisfaction with the partial opening of schools; original survey and Kobo form ask about satisfaction with the *decision* to partially open schools.
anu_edu_5	Documentation lists answer option 2 as "Forget to read & write"; this is different from the original survey/Kobo form.
anu_edu_8	Documentation question text: "Do you think your child could recover the learning loss?" The Kobo form refers to "damage caused by long school closure" rather than learning loss.
anu_edu_10, anu_edu_11	Documentation refers only to private tutors; original survey and Kobo form explicitly include private study groups ("batches").
anu_edu_14	Answer options in the original survey and Kobo form are more precise than in the documentation (e.g., "weekly one or two days" vs "occasionally").
anu_edu_15	The original survey and Kobo form contain instructions for enumerators that mention a conversion into minutes using a table; this table is not provided and the answer options are qualitative (matching the documentation).
anu_edu_18, anu_edu_19, anu_edu_20, anu_edu_21, anu_edu_22	The original survey and Kobo form specify this is *only* "alone" study time of children.
anu_pi_1, anu_pi_2, anu_pi_3	Instructions to enumerators in the original survey and Kobo form count ordering children to study on their own as "teaching them".
anu_pi_9, anu_pi_16	The original survey and Kobo form provide additional instructions to enumerators how and in what cases to validate the parents' answers.
anu_pi_17, anu_pi_18	The questions are swapped in the documentation relative to the survey and Kobo form.
anu_pi_19, anu_pi_20	The questions are swapped in the documentation relative to the survey and Kobo form.
anu_pi_21, anu_pi_22	The questions are swapped in the documentation relative to the survey. The text of the questions differs: the documentation asks about ability to *help* child study, but the original survey and Kobo form ask about ability to teach.
anu_negP*	The original survey and Kobo form provide instructions that answers should consider the events over the last two months. The Kobo form contains additional instructions to encourage "correct" answers.
anu_negP6	The documentation provides more nuance to the question text, suggesting a reason for beating a child.
anu_PA11	The documentation asks about talking to a child in a friendly manner, but the original survey and Kobo form are broader and ask about treating a child as a good friend.
anu_ren_lead2	The appendix refers to being respected by classmates, but the original survey and Kobo form ask about being praised by friends and classmates.
anu_ren_lead5	The appendix refers to the ability to bring structure to things, people and situations, but the original survey/Kobo form only ask about organizing items or leading friends.

`anu_ren_plan1`	The original survey/Kobo form question seems mis-translated, as it reads "Can properly manage what is needed to do a job." as opposed to "determines what information or resources are necessary for accomplishing a task." in the appendix.
`anu_ren_plan7`	The original survey/Kobo form question seems mis-translated, as it reads "Expert in a particular game like chess" as opposed to "is good at games of strategy where it is necessary to anticipate several moves ahead." in the appendix.
`anu_sds1`	The original survey/Kobo form question refers to motivation as opposed to encouragement in the appendix.

Unless noted otherwise, documentation refers to `Endline-Parent-Survey-Data-Layout.pdf`. Original survey refers to `Endline-Parent-Survey-Questions-Bangla.pdf`, i.e., the survey text provided in the second replication package (Wang et al. 2025). The Kobo form refers to `Endline_Assessment_Kobo_Form.xlsx` that was provided in the third (unpublished) replication package and that we had translated into English.