

Bonander, Carl et al.

**Working Paper**

"Try to Balance the Baseline": A Comment on "Parent-Teacher Meetings and Student Outcomes: Evidence from a Developing Country" by Islam (2019)

IZA Discussion Papers, No. 17781

**Provided in Cooperation with:**

IZA – Institute of Labor Economics

*Suggested Citation:* Bonander, Carl et al. (2025) : "Try to Balance the Baseline": A Comment on "Parent-Teacher Meetings and Student Outcomes: Evidence from a Developing Country" by Islam (2019), IZA Discussion Papers, No. 17781, Institute of Labor Economics (IZA), Bonn

This Version is available at:

<https://hdl.handle.net/10419/316736>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

DISCUSSION PAPER SERIES

IZA DP No. 17781

**“Try to Balance the Baseline”:  
A Comment on “Parent-Teacher Meetings  
and Student Outcomes: Evidence from a  
Developing Country” by Islam (2019)**

Carl Bonander  
Olle Hammar  
Niklas Jakobsson  
Gunther Bensch  
Felix Holzmeister  
Abel Brodeur

MARCH 2025

## DISCUSSION PAPER SERIES

IZA DP No. 17781

### **“Try to Balance the Baseline”: A Comment on “Parent-Teacher Meetings and Student Outcomes: Evidence from a Developing Country” by Islam (2019)**

**Carl Bonander**  
*University of Gothenburg*

**Olle Hammar**  
*Linnaeus University and Institute for  
Futures Studies*

**Niklas Jakobsson**  
*Karlstad University*

**Gunther Bensch**  
*RWI*

**Felix Holzmeister**  
*University of Innsbruck*

**Abel Brodeur**  
*University of Ottawa, Institute for Replica-  
tion and IZA*

MARCH 2025

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

**IZA – Institute of Labor Economics**

Schaumburg-Lippe-Straße 5–9  
53113 Bonn, Germany

Phone: +49-228-3894-0  
Email: [publications@iza.org](mailto:publications@iza.org)

[www.iza.org](http://www.iza.org)

## ABSTRACT

---

### **“Try to Balance the Baseline”: A Comment on “Parent-Teacher Meetings and Student Outcomes: Evidence from a Developing Country” by Islam (2019)\***

Islam (2019) reports results from a randomized field experiment in Bangladesh that examines the effects of parent-teacher meetings on student test scores in primary schools. The reported findings suggest strong positive effects across multiple subjects. In this report, we demonstrate that the school-level randomization cannot have been conducted as the author claims. Specifically, we show that the nine included Bangladeshi unions all have a share of either 0% or 100% treated or control schools. Additionally, we uncover irregularities in baseline scores, which for the same students and subjects vary systematically across the author’s data files in ways that are unique to either the treatment or control group. We also discovered data on two unreported outcomes and data collected from the year before the study began. Results using these data cast further doubt on the validity of the original study. Moreover, in a survey asking parents to evaluate the parent-teacher meetings, we find that parents in the control schools were more positive about this intervention than those in the treated schools. We also find undisclosed connections to two additional RCTs.

**JEL Classification:** B41, C12, I25

**Keywords:** reproduction, student outcomes, field experiments, Bangladesh

**Corresponding author:**

Abel Brodeur  
University of Ottawa  
75 Laurier Ave E  
Ottawa, ON K1N 6N5  
Canada  
E-mail: [abrodeur@uottawa.ca](mailto:abrodeur@uottawa.ca)

---

\* Bonander, Hammar, and Jakobsson contributed equally to the analysis and writing. Bensch contributed to the analysis. Holzmeister and Brodeur contributed to the writing. We are grateful to Bangladeshi colleagues who want to remain anonymous and to Lenka Fiala, Jack Fitzgerald, Anders Kjelsrud, Andreas Kotsadam, Essi Kujansuu, Ole Rogeberg, and David Valenta for comments and suggestions. Errors are ours.

*“Randomized controlled trials (RCTs) are experiments in which participants are randomly assigned to either intervention or control groups.”—Islam (2024, p. 1)*

## 1 Introduction

Islam (2019) evaluates the impact of a randomized intervention aimed at improving educational outcomes through structured parent-teacher meetings in rural Bangladesh. The treatment involved parents attending multiple one-to-one meetings with teachers across two years, where they received detailed information about their children’s academic progress and were encouraged to support their education at home actively. The study compares outcomes between treated and untreated schools in terms of intention-to-treat (ITT) estimates and analyzes various standardized test scores across multiple subjects. The main results indicate significant improvements in students’ academic performance in the treatment group: *“The regression results suggest that the students in grade 5 at the end-line gained in all subjects, with the highest increases occurring in math (0.42 SD) and English (0.41 SD), respectively. Considering all subjects (Math, English, Science, and Bengali), the average increase in test scores at the end of year 2 is 0.38 SD for grade 5 students who were in the program for two successive years.”* (Islam 2019, p. 284)

In this comment prepared for the Institute for Replication (I4R) (Brodeur et al. 2024), we reproduce and assess the analyses using the publicly available replication package from the original authors (published as *Appendix E. Supplementary Materials* of the original article; Islam 2019). All our analyses were successfully reproduced by multiple coauthors.

When examining the replication package, we took notice of an annotation by the author on line 1 in one of the two .do files (EER\_Baseline\_midline\_2010\_11.do) available in Islam (2019)’s replication package, which reads *“try to balance the baseline”*. Our subsequent analysis uncovered several data irregularities that severely question the validity and, thus, the credibility of the original study’s findings. First and foremost, we find that the school-level treatment allocation has a clear non-

random spatial pattern. We demonstrate using permutation tests that the observed allocation is extremely unlikely to have occurred by the school-level randomization described in [Islam \(2019\)](#). Second, baseline test scores for the same students differ systematically across datasets, with discrepancies varying between the treatment and control groups. We also find irregular patterns in the distribution of baseline test scores, including clusters of values that are unique to either the treatment or control group. In the author’s replication files, among other things, we also find unreported pre-treatment test score data and outcome data on two unreported subjects (social science and religion), positive evaluations of the treatment among parents in the control schools, and note several additional data irregularities and inconsistencies throughout.

This comment is structured as follows. In [Section 2](#), we provide an overview of the intervention and its context and discuss interconnections with other field experiments in which the author has been involved. [Section 3](#) demonstrates that the school-level randomization cannot have been conducted as the author claims. [Section 4](#) identifies several data irregularities, including inconsistencies in baseline test scores, unreported pre-treatment data, and issues with student IDs and survey sampling. [Section 5](#) presents robustness checks, including regressions with alternative sets of baseline test scores as controls and a panel data analysis using pre-treatment test scores from 2010.

## **2 Intervention and context**

The experiment was carried out in 76 government primary schools in rural areas in the Khulna and Satkhira districts in Bangladesh. 40 schools were randomly selected for the treatment group and 36 for the control group. The experiment involved monthly one-on-one parent-teacher meetings over two academic years. Meetings lasted about 15 minutes and included personalized feedback on academic performance, attendance, and study habits. Report cards were prepared for both treatment and control schools, but only treatment school parents received them during

the meetings. Control school parents received no report cards and were not invited to one-to-one meetings. The intervention was implemented by the Global Development Research Initiative (GDRI) with approval from Bangladesh’s Department of Primary Education, and teachers in treatment schools received training prior to the intervention.

The parent-teacher meetings began in April 2011 with an initial information session for parents. Monthly meetings followed, starting in May and June 2011, with reminders sent in advance. The intervention lasted two academic years. Year 1 (2011) involved students in grades 4 and 5, with five meetings held between May and October. In Year 2 (2012), eight meetings were held from March onward, adding grade 3 students (formerly grade 2). Only grade 4 students from Year 1 participated in both years.

A standardized baseline test was conducted in March 2011 before the intervention was initiated in April 2011, followed by midline (December 2011) and endline (December 2012) tests in math, English, science, and Bengali. Year 2 assessments included reading, writing, and general knowledge. Tests were developed by education professionals and graded by retired teachers, with no involvement from current teachers. Grade 5 students were not assessed separately but took the nationwide Professional Skills Course (PSC) exams, a mandatory high-stakes test for secondary school entry. Official PSC exam scores (CGPA) were collected for all students in treated and untreated schools. Students were also surveyed on time use, non-cognitive skills, and behavior. Over a year after the program ended, a household survey assessed parental time allocation and perceptions of the meetings’ impact on their children’s education.

## **2.1 Interconnection with other field experiments**

While conducting this replication, we noted similarities with two other field experiments conducted by the same author in the same districts (Khulna and Satkhira) in Bangladesh: [Begum et al. \(2018\)](#) and [Begum et al. \(2022\)](#), with the latter being

an extension of the former. While Begum et al. (2018) investigates parental gender bias using an allocation experiment, Begum et al. (2022) examines the relationship between parental gender bias and investment in children’s education and health. Neither of these two studies nor the study by Islam (2019) was pre-registered.

For Begum et al. (2018), we have not been able to re-analyze the data since the publication does not include any publicly available replication package. For Begum et al. (2022), however, a replication package is available at the journal’s website (published as *Supplementary material* alongside the article). While the replication package only includes the final dataset and *Stata* codes for generating the main tables (i.e., no raw data or cleaning codes), we found that one of the ID variables (*sch\_id*) matched perfectly to the data used in Islam (2019). This same variable—including identical value labels—indicates schools in Islam (2019) and villages in Begum et al. (2022). More importantly, there is a perfect correlation in treatment status per school/village across the two papers, which suggests that the same randomization has been (re)used in the two field experiments.<sup>1</sup>

While the exact timeline is not clearly specified, the authors write that the interventions in Begum et al. (2018) and Begum et al. (2022) took place in 2012, that is, at the same time as the intervention in Islam (2019) was ongoing. Despite this, neither Begum et al. (2018) nor Begum et al. (2022) is mentioned in Islam (2019); likewise, Islam (2019) is neither mentioned in Begum et al. (2018) nor in Begum et al. (2022).<sup>2</sup>

---

<sup>1</sup>In Islam (2019), there are 76 schools in total (labeled numerically with 1 through 77, with ID 52 having been omitted), of which school IDs 1–40 are in the treatment group, and school IDs 41–77 are in the control group. In Begum et al. (2022), there are in total 55 schools/villages, which constitute a proper subset of the schools in Islam (2019), with the exact same *sch\_id*, value labels, and treatment status. See Table A1.

<sup>2</sup>In an earlier working paper version of Begum et al. (2018), the following is mentioned in a footnote: “A baseline survey for a different research project had been conducted by the third author. [...] The other project is a multiyear RCT (Randomized Controlled Trial) involving the school children in the locality; this study was undertaken independently from that project except that we used the basic household information.” (Begum et al. 2014, p. 11) However, reusing the same randomization is not mentioned, and even this vague reference to another intervention has been removed from the published version of the paper.



### 3 A non-randomized intervention

A first existential issue with Islam (2019) is that the intervention does not appear to be randomized at all. According to the author: “A total of 40 schools were selected randomly for the treatment after baseline tests had been conducted. The remaining 36 schools served as the control group throughout the two years of intervention.” (Islam 2019, p. 277). Looking at the geographical distribution of these schools, however, all treatment schools are located in five unions:<sup>3</sup> Amadi, Chandkhali, Garuikhali, Laskar, and Raruli. All control schools are located in four other unions: Anulia, Bagali, Baradal, and Khajra. Even at the *upazila* level, all included schools in Paikgachha are treated, while all included schools in Assasuni are untreated. Koyra is the only *upazila* that includes schools in both the treatment and control groups (but still with perfect separation at the union level, with treatment in Amadi and control in Bagali). With the exception of Bagali, all treated schools are in the Khulna district, whereas all control schools are in Satkhira. That is, instead of randomly assigning the 76 included schools into the treatment and control conditions, either *all* schools in certain unions were treated or *all* schools in certain unions were untreated. This non-random treatment allocation of the schools is shown graphically in Figure 1. As emphasized by a permutation test randomly permuting the treatment assignment 10,000 times (Figure 2), the probability that the treatment allocation in Islam (2019) would occur due to chance is infinitesimal.

Additionally, while the two districts of Khulna and Satkhira cover 16 *upazilas*, all schools included in the study are located in only three of them (Assasuni, Koyra, and Paikgachha; see Figure 3). In these three *upazilas*, there are a total of 227 government primary schools (GPS) as of February 2025 (IPEMIS 2025) (see Table 1). We interpret these schools as being the “set of more than 200 schools in those regions” referred to in Islam (2019, p. 277). However, while there are a total of 28 unions in these three *upazilas*, as mentioned above, the schools included in Islam

---

<sup>3</sup>Bangladesh is geographically divided into eight divisions (NUTS1 areas; called *bibhags*), 64 districts (NUTS2 areas; called *zilas*), 495 sub-districts (NUTS3 areas; called *upazilas*), and 4,596 unions (NUTS4 areas).

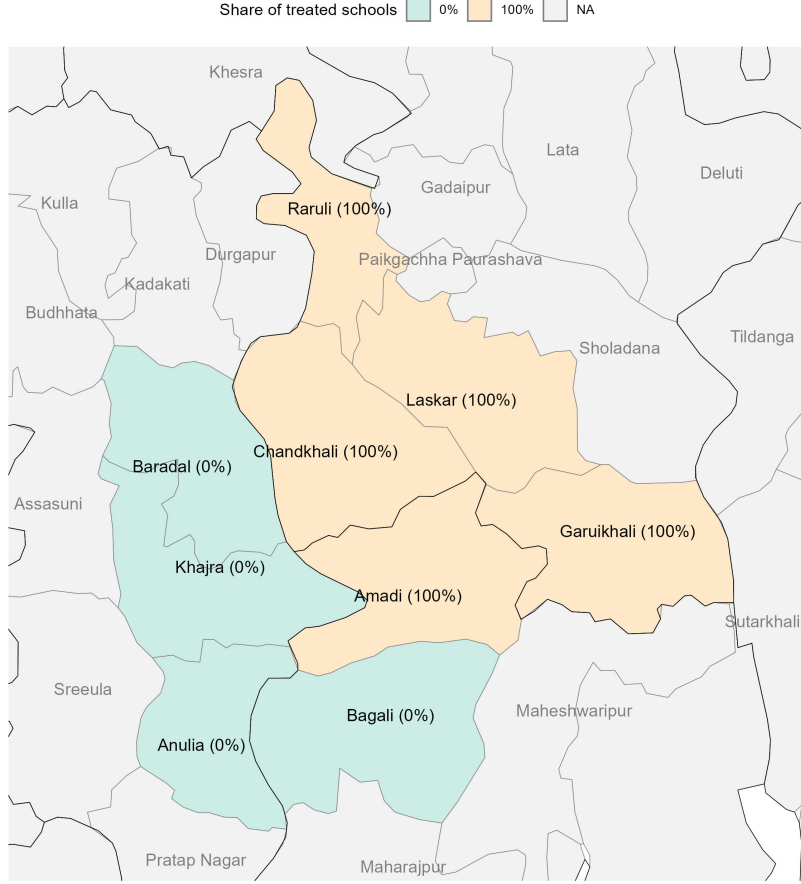


Figure 1: Allocation of treated and untreated schools on the union level in [Islam \(2019\)](#). Union-level shares of treated schools are presented in parentheses.

(2019) are only located in nine of these unions. In these nine unions, however, 95 percent of all schools (that is, 76 out of 80 schools) are included in the study. That is, instead of randomizing among the more than 200 schools in these *upazilas*, almost all schools in a few selected neighboring unions were included. This non-random location of the schools implies that the following claim by the author is not true: “*The field experiment was carried out in 76 government primary schools in rural areas in two districts of Bangladesh. These schools were chosen randomly from a set of more than 200 schools in those regions.*” ([Islam 2019](#), p. 277). In other words, opposite to what is stated in the paper, the author neither randomized which schools to include in the study nor which schools were treated.

It should be noted that since the same “randomization” was used in [Begum et al. \(2018\)](#) and [Begum et al. \(2022\)](#) (see Section 2.1), this major concern of non-

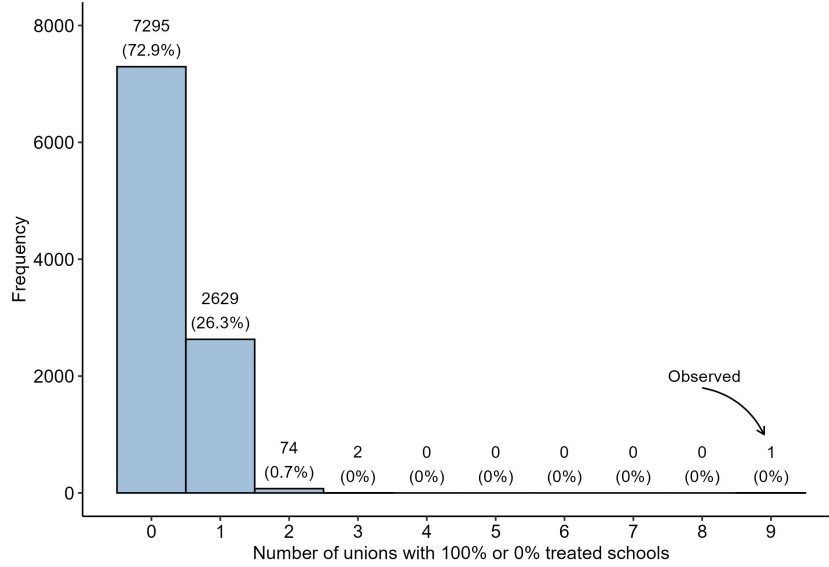


Figure 2: Results of a permutation test randomly permuting the school-level treatment variable 10,000 times, counting the number of unions with 0% or 100% treated schools in each iteration. Numbers above each bar show the number and percent of iterations that arrive at a specific value, including the observed count from Islam (2019), which is 9. The highest count we obtain in the random permutations is 3.

randomization applies to these papers as well.

## 4 Additional data irregularities

In this section, we discuss a number of additional data irregularities found when conducting the reproduction. For a detailed description of the replication package for Islam (2019), see Appendix A.

### 4.1 Inconsistencies in reported data collection timing and data labels

As detailed in Section 2, Islam (2019) reports that the baseline tests were conducted in March 2011, with midline outcomes measured in December 2011 and endline outcomes measured in December 2012. In the author’s replication package, we find multiple variables labeled 2010, indicating that they are from one year before the study was initiated. Specifically, we discover final mark variables named `[subject]_10` and `[subject]_11`, with *Stata*’s attached variable labels indicating that they are from 2010 (“final mark of [subject] 2010”) and 2011 (“final mark of [subject] in 2011”) in `hh2013a_meet2011.dta`. We also identify multiple sets of

Included in sample ☐ No ☒ Yes

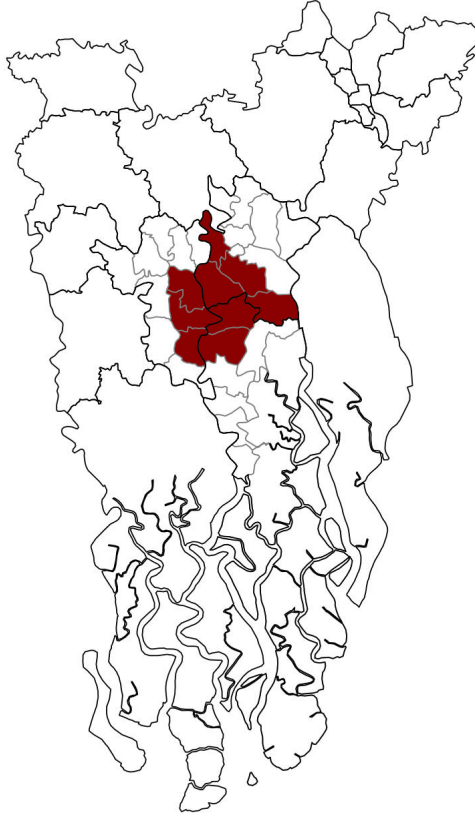


Figure 3: Map of the three *upazilas* (Assasuni, Koyra, and Paikgachha) highlighting the nine unions sampled in [Islam \(2019\)](#).

baseline test scores that appear to be from different years scattered across various files (Table 2).

#### 4.2 Missing, non-distinct, and non-matching student IDs

In the main analysis file for 2011 (`testscore2011_2010_FINAL.dta`), there are 4,759 observations, 2,643 (55.5%) of which have missing student IDs (`st_ID`). The student ID variable in this dataset is labeled as “*student matching ID base on 2012 file*”, suggesting that it is meant for matching students moving from grade 4 in 2011 to grade 5 in 2012. However, attempting to match individuals with non-missing and distinct IDs to the 2012 analysis file (`testscore2012_Final.dta`) yields severe

Table 1: Total number of government primary schools (GPS) per union in February 2025 ([IPEMIS 2025](#)), and number of GPS included in [Islam \(2019\)](#) per union and treatment status.

District	Upazila	Union	Total number of schools	Number of schools in <a href="#">Islam (2019)</a>	Treatment status
Khulna	Paikgachha	Deluti	6	0	
Khulna	Paikgachha	Gadaipur	7	0	
Khulna	Paikgachha	Haridhali	9	0	
Khulna	Paikgachha	Kapilmuni	8	0	
Khulna	Paikgachha	Lata	6	0	
Khulna	Paikgachha	Sholadana	6	0	
Khulna	Paikgachha	Other (wards)	2	0	
Khulna	Paikgachha	Chandkhali	11	11	Treatment
Khulna	Paikgachha	Garuikhali	6	3	Treatment
Khulna	Paikgachha	Laskar	8	8	Treatment
Khulna	Paikgachha	Raruli	8	8	Treatment
Khulna	Koyra	Amadi	11	10	Treatment
Khulna	Koyra	Dakshin Bedkashi	6	0	
Khulna	Koyra	Koyra	8	0	
Khulna	Koyra	Maharajpur	9	0	
Khulna	Koyra	Maheshwaripur	8	0	
Khulna	Koyra	Uttar Bedkashi	5	0	
Khulna	Koyra	Bagali	11	11	Control
Satkhira	Assasuni	Anulia	7	7	Control
Satkhira	Assasuni	Baradal	8	8	Control
Satkhira	Assasuni	Khajra	10	10	Control
Satkhira	Assasuni	Assasuni	9	0	
Satkhira	Assasuni	Budhhata	9	0	
Satkhira	Assasuni	Durgapur	6	0	
Satkhira	Assasuni	Kadakati	6	0	
Satkhira	Assasuni	Kulla	6	0	
Satkhira	Assasuni	Pratap Nagar	9	0	
Satkhira	Assasuni	Sobhnali	11	0	
Satkhira	Assasuni	Sreeula	11	0	
			227	76	

*Note:* Number of schools per union might have changed since 2012.

mismatches on the student sex, indicating a problem with the ID variable that hampers individual matching across the two main analysis files.

For students with non-missing and distinct IDs, however, we can perfectly match students across files within the same year (see [Appendix C](#) for details). This allows us to compile two datasets—one for the 2011 analyses and one for the 2012 analyses—that include a greater set of variables than available in the authors main analysis files, including alternative baseline variables (see [Section 4.1](#)).

### 4.3 Unreported pre-treatment and outcome data

[Table 2](#) shows that there are multiple sets of baseline test score variables, some of which are only available in the household follow-up survey data files. These are

Table 2: Baseline variables available for the four studied subjects (math, English, Bengali, and science) identified in various datasets in Islam (2019)’s replication package.

Dataset	Variable name	Variable label
testscore2011_2010_FINAL.dta hh13a_meet_teac_sch11_final.dta	[subject]10b	RAW test score of [subject] at BASELINE
hh2013_FINAL1.dta hh2013a_meet2011.dta	[subject]_b	Baseline test mark of [subject] 2011 (out of [10/15])
testscore2012_FINAL.dta hh13a_meet_teac_sch12_final.dta hh2013a_report_meeting2012.dta	std[subject]12b1	Standardized test score of [subject] at BASELINE
hh13a_meet_teac_sch12_final.dta hh2013a_report_meeting2012.dta	[subject]_b	Baseline test mark of [subject] 2012 (out of [10/15])

*Note:* Baseline tests for Bengali and Science have a maximum score of 10. Baseline tests for English and Mathematics have a maximum score of 15.

named \*\_b, available for all four studied subjects, and clearly labeled in *Stata* by year (2011 or 2012). Islam (2019)’s analyses rely on the baseline variables named \*10b as controls in the main regressions and variables named \*12b1 in the corresponding 2012 analyses. Given the naming conventions for other variables and the naming of the datasets, one could infer that these refer to 2010 and 2012 data (Table 2), although the documentation is not entirely clear.

Importantly, the final mark variables from 2010 appear to constitute unreported, pre-treatment versions of the midline outcome data. Using our matched dataset for 2011, we find that the final mark variables from 2011 are equivalent to the midline outcomes that Islam (2019) uses in the 2011 analyses for students in Grade 4.<sup>4</sup> Comparing the distribution of the 2010 and 2011 final mark variables suggests that these data were collected in a similar manner; one that is distinct from how the baseline tests were measured (Figure 4). This is inconsistent with the description of the data collection process provided in Islam (2019) (see Section 2).<sup>5</sup>

<sup>4</sup>For students in grade 5, 2011, Islam (2019) instead relies on grade point averages from a high-stakes test conducted by all schools: “The grade 5 students were not assessed separately as part of the project, but sat for the nationwide competitive exams (PSC exams) at the end of grade 5.” (Islam 2019, p. 282).

<sup>5</sup>In footnote 15, Islam (2019, p. 278) notes that “[t]he tests were administered separately by the implementing NGO (GDRI). However, PSC exams are conducted by education boards. The boards appoint external graders who are anonymous. For grade five students, the estimates are based on the PSC test scores. For students in other grades the estimates are based on the project specific test conducted by GDRI. Tests conducted by school teachers were used only for the report card, and we

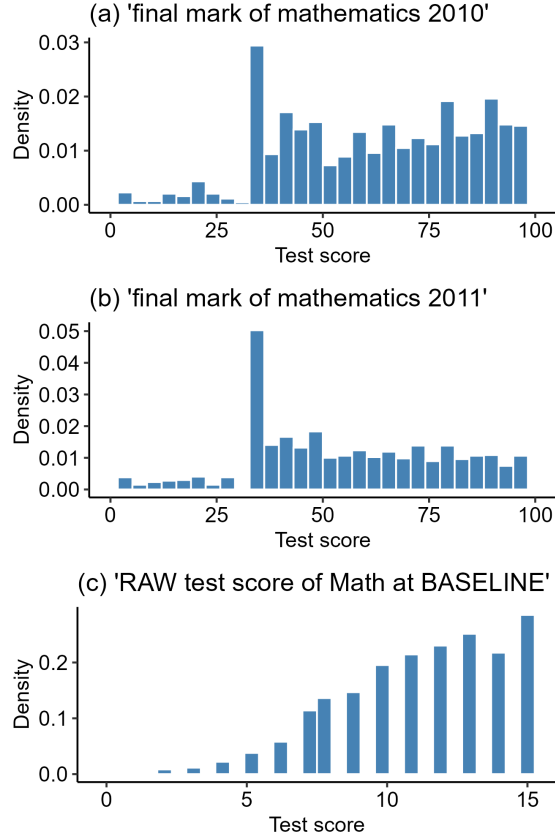


Figure 4: Distributions of variables labeled “final marks in 2010” and “final marks in 2011” as well as baseline test scores from the 2011 data in mathematics.

The data on final marks from 2010 and 2011 also contain two unreported subjects (social science and religion). We note that the author does not mention these subjects in the paper, except that one of them (social science) is shown to be included in report cards shared with the parents during the parent-teacher meetings (see *Appendix C: Sample Report Card* in Islam 2019, p. 303).<sup>6</sup> However, if these final marks were—as Islam (2019) states—collected by GDRI solely for evaluation and not conducted by the schools, it appears that it was initially intended to evaluate the effects on these subjects as well. However, no explanation is given in the paper

---

*did not use school administered test scores for the purpose of evaluation.*” Moreover, in footnote 19, Islam (2019, p. 280) states that “[t]he exams conducted by the schools differ across schools, so we did not consider them in our analysis. For the purposes of this study, we conducted the same tests in all treatment and control schools. We also used nationwide, externally-administered public exam results for the grade 5 students.”

<sup>6</sup>In addition to the four subjects analyzed in the paper (math, English, Bengali, and science) and the subject for which data is available despite not being mentioned in the article (social science), the sample report card lists an additional subject which does neither appear in the manuscript nor the data: general science.

as to why these outcomes were omitted or why these data were collected already in 2010 before the study commenced. We find no corresponding data for the two additional subjects in the 2012 datasets.

#### 4.4 Inconsistencies in baseline test scores for the same students

We identify systematic inconsistencies in the baseline test scores within the same subject and student when comparing the variables named `*_b`—not used in [Islam \(2019\)](#)’s analysis—with those used in the main analyses in [Islam \(2019\)](#).

For grade 4, 2011, we see that these values differ greatly for a substantial number of students (Figure [A3](#)). This could be inferred to be due to these variables potentially referring to different years (`*10b` possibly being from 2010; `*_b` being labeled as being from 2011). However, there is a clear clumping of individuals with the lowest possible Bengali baseline test score in the `*10b` variable that is clearly not present in the `*_b` variable. We note that many of these students have a perfect score in the `*_b` variable, so even if these are from different years, it would appear unlikely that they would shift from having the lowest possible score in one year to having the highest possible score in the next year.

In the 2012 data, available for students in grades 3 and 5, we identify systematic shifts in values occurring only in the control group (see Figure [5](#) for grade 3, and Figure [A4](#) for grade 5). For the treatment group, the correlations between baselines used by [Islam \(2019\)](#) and the alternative set of baselines are perfect within each subject, which corroborates the conjecture that the `*_b` variables (not used by [Islam 2019](#)) and `*_12b1` variables (used by [Islam 2019](#)) are from the same year and the same test. The observation that only the control group values on a baseline measure are shifted indicates that the data has been changed. We note that similar systematic shifts can also be seen in some subjects in grade 4, 2011 data (see, e.g., Bengali [control group] and science [treatment group] in Figure [A3](#)), although these systematic patterns are less obvious in the scatter plots for those data due to the many sporadic mismatches.



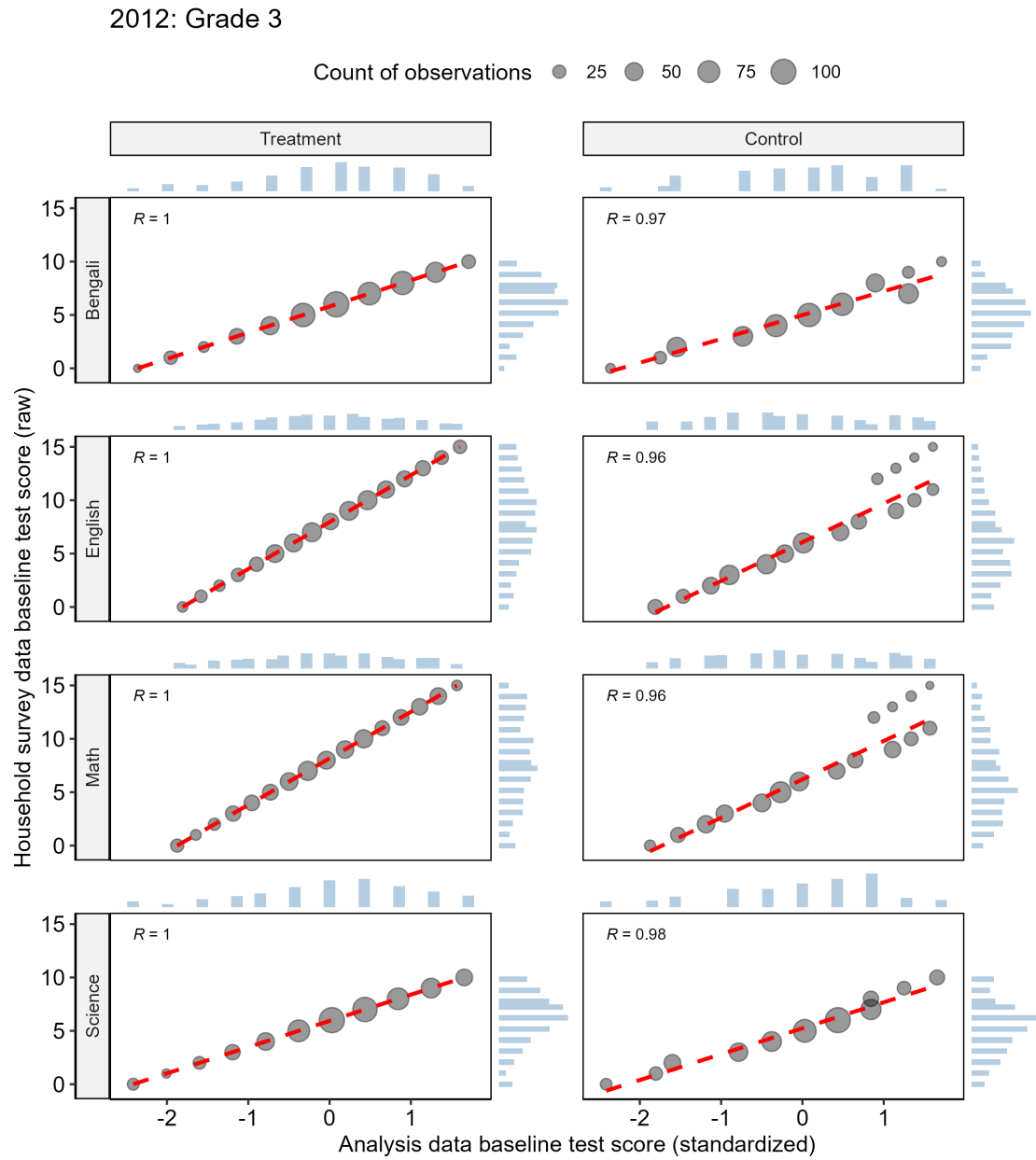


Figure 5: Comparison of baseline test scores in grade 3, 2012, from [Islam \(2019\)](#)'s analysis files (variables named \*12b1) and alternative baseline variables identified in the household survey files (named \*\_b) available in the author's replication package.

Investigating the histograms of the analysis file baseline test scores (see Figure 6 for grade 3, Figure A5 for grade 4, and Figure A6 for grade 5), we also discovered that some relatively common values were unique to either the treatment or the control group. This appears to be a data irregularity that should not typically occur for baseline data in a randomized trial, especially with discretely measured test scores, unless data for a particular school in either group happened to be reported with systematic bias.

To investigate this possibility, we performed a permutation test to assess the likelihood of the observed data pattern (i.e., baseline test scores being unique to one of the two groups) occurring by chance in these data (see Appendix D for details). We performed this test for the baselines used in Islam (2019)’s analysis (\*10b/\*12b1) and the alternative set of baselines uncovered in the replication data (\*\_b; see Section 4.3 for details). A low  $p$ -value in these tests indicates that it is unlikely that the observed patterns could have occurred by chance. As shown in Table A2, it is very unlikely that the patterns observed in the data could have occurred by chance for baseline test scores in the analysis files (\*10b; \*12b1) for several subjects, particularly in the 2012 data. We do not identify similar issues in the alternative baseline set from the household survey files (\*\_b).

#### 4.5 Irregular grouping among controls in the 2011 GPA data

Islam (2019, p. 302) presents only a single figure showing how the test scores are distributed: a kernel density plot for test scores in standardized GPA (grade 5, 2011; see Figure A7). Reproducing the figure as a histogram without smoothing reveals an irregular bunching of GPA test scores that is limited to the control group (Figure 7).

#### 4.6 Being in the household survey sample is strongly correlated with treatment and outcomes

Islam (2019) reports that approximately 50% of households were randomly sampled

2012: Grade 3

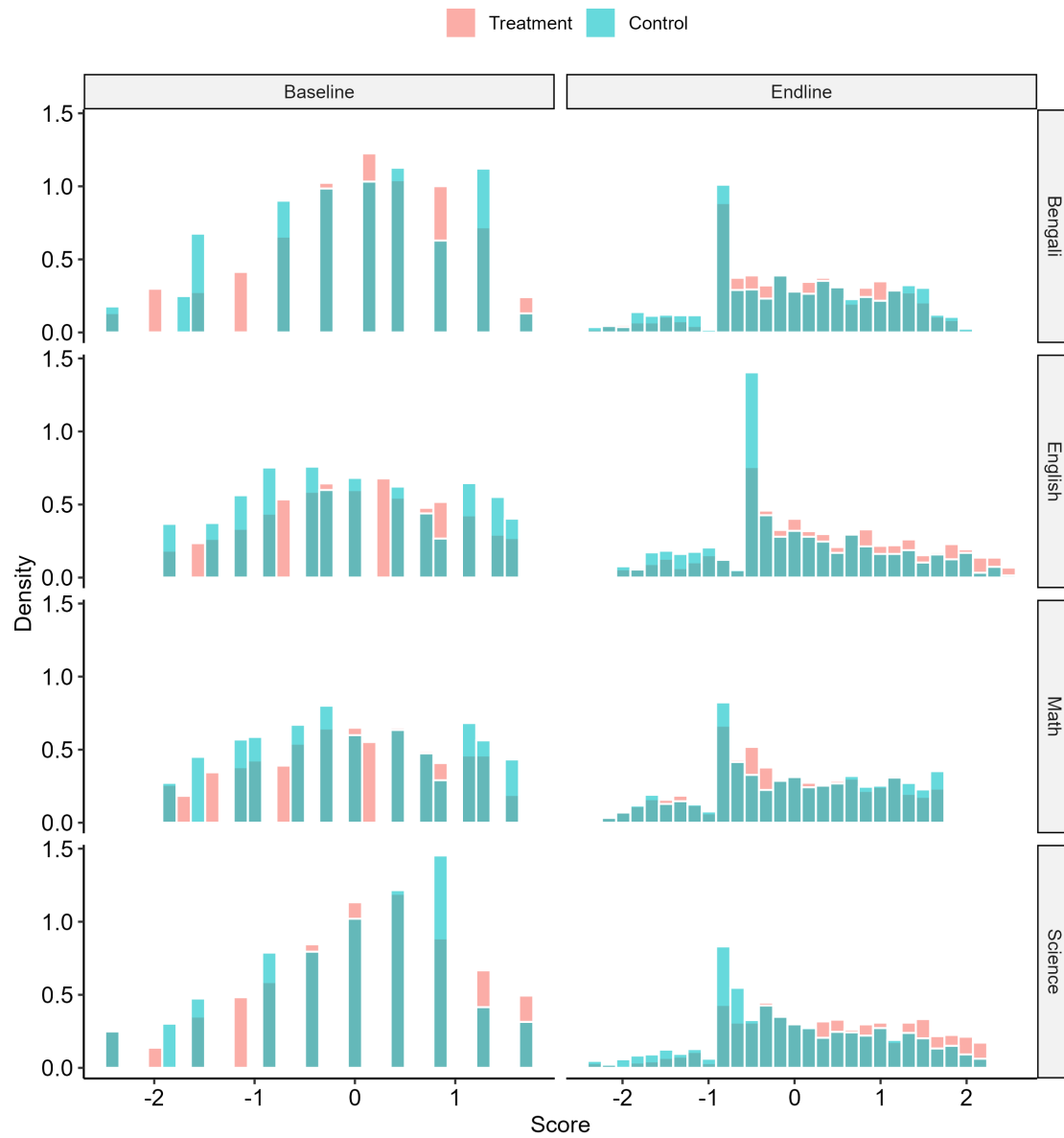


Figure 6: Histograms of the baseline and midline test scores for grade 3, 2012, by treatment group and subject.

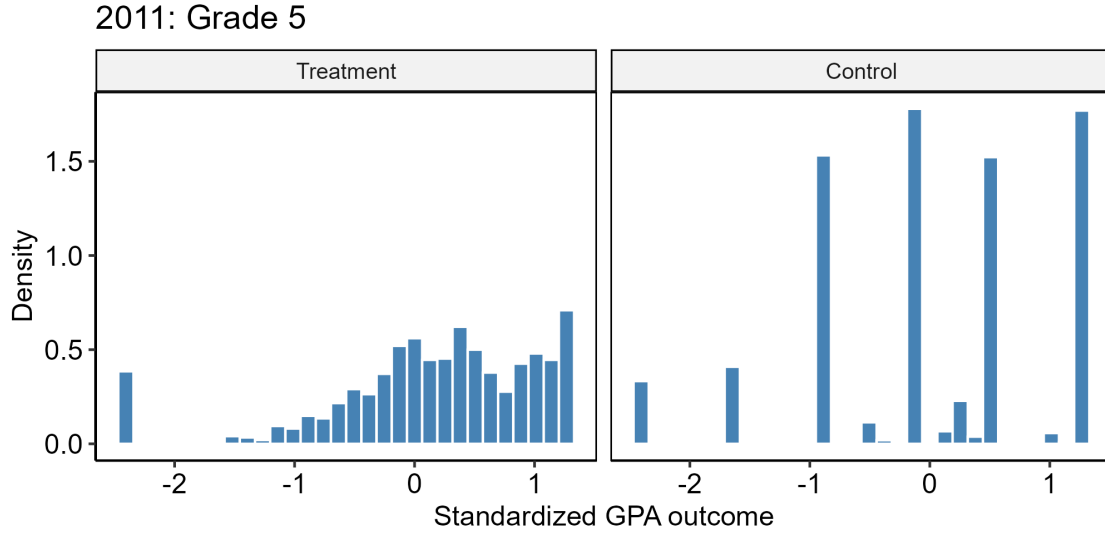


Figure 7: Replication of Figure A4 in [Islam \(2019\)](#) without smoothing.

for a follow-up survey in 2014. However, the paper provides very limited details about the sampling strategy, other than a footnote indicating that it was based on odd or even roll numbers: *“We attempted to visit either odd or even-numbered students by their class roll numbers, which are based on their classroom rankings”* ([Islam 2019](#), p. 289). [Islam \(2019\)](#) also assures readers that test scores do not systematically differ between sampled and non-sampled students: *“The test scores of the children in the households that were surveyed do not differ from those in the households that were not”* ([Islam 2019](#), p. 289), but does not present data to support this claim.

To investigate, we first examine whether individuals in the analysis data files appear in the household survey files, broken down by year, class, and treatment group. For grade 4, 2011, 79% of treated students are included in the household survey file compared to only 53% of control students. For grade 3, 2012, the difference is smaller, with 46% of treated students and 43% of controls included. For grade 5, 2012, 75% of treated students appear in the household data, compared to just 44% of controls. These discrepancies point to significant issues with the survey sampling that are not disclosed in the paper.<sup>7</sup>

<sup>7</sup>One possible explanation is that treated students may have performed better due to the intervention, making their parents more likely to respond to the follow-up survey. However, [Islam \(2019\)](#) provides no information on response rates, and the paper states that the follow-up

To further explore this issue, we re-run the main regressions from Islam (2019), adding a dummy variable to indicate whether an individual is included in the household survey. The results are striking: in many cases, the “effect” of being in the survey sample is similar in size to the effect of being in the treatment group (see Table 3). These findings suggest either substantial failures in the random sampling process or severe, undisclosed non-response biases that differ by treatment group.

Table 3: Comparison of regression estimates with treatment and household survey sampling dummies, based on our matched data files.

Year	Grade	Subject	Treatment	Household Sample
2011	4	Bengali	0.262** (0.101)	0.291*** (0.064)
2011	4	English	0.323** (0.127)	0.247*** (0.077)
2011	4	Math	0.201* (0.110)	0.223*** (0.070)
2011	4	Science	0.124 (0.109)	0.242*** (0.060)
2012	3	Bengali	-0.006 (0.107)	0.066* (0.039)
2012	3	English	0.317*** (0.091)	0.062 (0.043)
2012	3	Math	-0.087 (0.079)	0.057 (0.040)
2012	3	Science	0.332*** (0.107)	0.046 (0.037)
2012	5	Bengali	0.309*** (0.100)	0.456*** (0.058)
2012	5	English	0.413*** (0.126)	0.438*** (0.091)
2012	5	Math	0.420*** (0.129)	0.290*** (0.082)
2012	5	Science	0.339*** (0.106)	0.339*** (0.079)

*Note:* Cluster robust standard errors in parentheses. Estimates in each column are coefficients from separate regressions. Estimates of the constant and the baseline test scores are omitted from the reporting. Treatment effect estimates differ slightly from Tables 5 and 6 in Islam (2019) due to individuals lost when cleaning out duplicated and missing student IDs before matching with the household survey files. \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

#### 4.7 A closer look at the follow-up survey: A multitude of inconsistencies

Taking a closer look at the follow-up survey (Panel A of Table 8 in Islam 2019, p. 289), which reports parents’ evaluations of their children one year after the end of the intervention, we note a number of inconsistencies.

First, according to Islam (2019), these results are from a follow-up survey conducted in 2014. However, the data for these results are scattered across four different datasets (`tab8_panelA_part1.dta`, `tab8_panelA_part2.dta`, `Table8_PanelA_B.dta`, and `roster_all_final.dta`). One of these datasets includes information about the

---

survey was not obviously related to the experiment: “we surveyed parents almost after a year of the completion of the intervention. We also used a different set of enumerators for the post-intervention household survey” (Islam 2019, p. 290).

date of the interview (variable `inter_date` in `roster_all_final.dta`), showing that those surveys were conducted in 2013 (August 10 through October 29), rather than in 2014.<sup>8</sup> For two of the datasets (`tab8_panelA_part1.dta` and `tab8_panelA_part2.dta`), no information about interview dates is included. Furthermore, the question “*Have private tutor*”, appears not to be from a follow-up survey in 2014 but taken directly from the same dataset as Panel B, (i.e., from the students’ evaluations in 2012). Data on the question “*Private tuition is very important for doing well in exams*” is not available at all. Moreover, there is a large variation in sample sizes between the different datasets,<sup>9</sup> which might suggest that the data does not originate from the same survey. There are also multiple observations for the same student IDs.<sup>10</sup>

Second, in an attempt to computationally reproduce the results reported in Table 8 of Islam (2019, p. 289) (i.e., running the same analysis code with the same data as in the original paper), the results for “*Whether child fails to progress to next grade*”, “*Did the child get a scholarship in the grade 5 PSC exam?*”, and “*Private tuition is very important for doing well in exams*” were not reproducible—see Table 4, columns (1) and (2). Interestingly, the treatment effect on “*Whether child fails to progress to next grade*” is positive but statistically insignificant in our reproduction (instead of negative and statistically significant, as reported in the paper). On the other hand, we find that the treatment effect on “*Did the child get a scholarship in the grade 5 PSC exam?*” is positive and statistically significant in our reproduction (instead of statistically insignificant, as reported in the paper).

Third, while no raw data, questionnaire, or codebook is provided for the sur-

---

<sup>8</sup>This dataset also includes a variable `hh_cov`, with value labels 1 *only baseline survey*, 2 *only evaluation survey*, 3 *both baseline & evaluation survey*. However, only observations with value 2 *only evaluation survey* are included in the data.

<sup>9</sup>In `roster_all_final.dta`, there are 13,896 individuals, of which 3,265–5,068 have answered the different survey questions. In `tab8_panelA_part1.dta`, there are 8,626 observations with complete records on the survey questions. In `tab8_panelA_part2.dta`, there are 5,259 individuals, of which 4,966 have answered the questions. Finally, in `Table8_PanelA_B.dta`, there are 9,137 observations, with 7,114 complete records on the survey items.

<sup>10</sup>In `roster_all_final.dta`, only 11,380 of the 13,896 student IDs are unique. In `tab8_panelA_part1.dta`, there are 8,626 observations for 5,055 students. In `tab8_panelA_part2.dta`, 5,073 out of 5,259 student IDs are unique. `Table8_PanelA_B.dta` does not involve duplicates in terms of student IDs. It should also be noted that none of these numbers corresponds to the 5,128 households reported in the paper.

Table 4: Robustness reproduction of *Table 8, Panel A: Parental self-report in 2014* in [Islam \(2019\)](#).

	Original	Reprod.	Adj. Reprod.
<b>tab8_panelA_part1.dta</b>			
Father is helping with study at home most of the time	0.041*** (0.006)	0.041*** (0.006)	0.042*** (0.010)
Mother is helping with study at home most of the time	0.031*** (0.007)	0.031*** (0.007)	0.032* (0.016)
Others (brother/sister) helping with study at home	0.045*** (0.007)	0.045*** (0.007)	0.045*** (0.015)
<b>Table8_PanelA_B.dta</b>			
Have private tutor	.0743*** (0.012)	0.074*** (0.012)	0.074** (0.034)
<b>roster_all_final.dta</b>			
Whether child fails to progress to next grade	-0.017*** (0.008)	0.042 (0.068)	0.027 (0.037)
Did the child get a scholarship in the grade 5 PSC exam?	0.027 (0.043)	0.031*** (0.010)	0.029*** (0.010)
Does student go to school now?	N/A	-0.023*** (0.008)	-0.026** (0.011)
Does there change of class roll of student?	N/A	-0.018** (0.007)	-0.007 (0.011)
Does there change of class of student?	N/A	-0.017** (0.007)	-0.007 (0.011)
Does student leave school?	N/A	0.197*** (0.018)	0.146*** (0.030)
<b>tab8_panelA_part2.dta</b>			
Child spends more time on household work than study	-0.017*** (0.004)	-0.017*** (0.004)	-0.017* (0.009)
Child cannot go to school regularly because of work	-0.032*** (0.005)	-0.032*** (0.005)	-0.032* (0.018)
Child hangs out with naughty boys/girls	-0.009 (0.006)	-0.009 (0.006)	-0.009 (0.015)
<b>No data available</b>			
Private tuition is very important for doing well in exams	-0.100*** (0.014)	N/A	N/A

*Note:* Column (1) shows the original results as reported in [Islam \(2019\)](#). Column (2) shows results from a computational reproduction using the same code and data as in the replication package. Column (3) shows the results from a reproduction that excludes observations from school ID 93, includes students in grade 4, and clusters the standard errors at the school level. Columns (1) and (2) report *t*-test results (T-C) with standard errors in parentheses. Column (3) reports results with cluster-robust standard errors in parentheses. \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

vey(s), the variable naming convention might suggest that the variables included in the survey data files are a selection from a broader set of variables (for example, in `tab8_panelA_part2.dta`, the three outcome variables included are named `q_6_4`, `q_6_5`, and `q_6_11`); in `roster_all_final.dta`, however, six potential outcome variables are included (`q1b_1`, `q1b_2`, `q1b_3`, `q1b_4`, `q1b_5`, and `q1b_6`), of which only two are analyzed in the paper (`q1b_2` and `q1b_5`). Interestingly, the treatment effects on the variables not reported in Islam (2019) point in the opposite direction of those reported in the paper. That is, approximately one year after the end of the intervention, the children in the treated schools were not significantly less likely to fail to progress to the next grade (as reported in the original article), but statistically significantly less likely to still go to school and significantly more likely to have left school—see column (2) in Table 4.

Fourth, the dataset `tab8_panelA_part1.dta` includes a mysterious school with school ID 93, which is not included in any of the other datasets or analyses. This school is allocated to the control group and includes 41 observations. In an adjusted reproduction—reported in Table 4, column (3)—we have excluded all observations from this school. Additionally, in the analysis using the dataset `roster_all_final.dta`, the author excludes all students in grade 4 (5,900 observations).<sup>11</sup> In the adjusted reproduction, we have included all students. Finally, in the adjusted reproduction we have also appropriately clustered the standard errors at the school level. As shown in column (3) in Table 4, these adjustments change the statistical significance from the 1% level to the 10% level for the following outcomes: *“Mother is helping with study at home most of the time”*, *“Child spends more time on household work than study”*, and *“Child cannot go to school regularly because of work”*.

---

<sup>11</sup>The reason for this, as per the annotation in the analysis script `EER_Endline2012_13.do`, is the following: *“report this variable dropping grade 4 students[...] as we do not have the record for them in time.”* We do not understand the reasoning behind this motivation.



#### 4.8 Different schools are defined as treated in the teacher absence analysis

In the analysis of teacher absences in 2011 (reported in Table A11 in Islam 2019), the treatment and control group allocation of schools differs from the other analyses and from how it is explained in the paper. In the dataset for this analysis (`Teacher Absence_2011_final.dta`), all schools with school ID 1–51 are defined as treatment schools (instead of school IDs 1–40), and all schools with school ID 54–77 are defined as control schools (instead of school IDs 41–77). That is, 11 schools switched from the control group to the treatment group for this analysis.<sup>12</sup> Additionally, the school with school ID 53 has observations in both the treatment (1 observation) and control (4 observations) groups. Another peculiarity with this dataset is that the variable that is supposed to measure absence in all visits in 2011 is called `abs_teach13` (and not `abs_teach11`, following the author’s variable naming convention to indicate the year of data collection as a suffix in variable names).<sup>13</sup>

Allocating schools to treatment and control in accordance with the other analyses, the treatment has a positive effect on teacher absence in June 2011 (statistically significant at the 10% level) as compared to the statistically insignificant result reported in Islam (2019), while the statistical significance for the negative effect found for August 2011 changes from the 1% to the 10% level (see Table A3).

#### 4.9 The control group was even more positive about the treatment

According to Islam (2019, p. 289), “[t]he household survey asked parents in the treatment group for their opinion of the parent-teacher meetings in this intervention. Most of the parents in the treatment schools thought that the parent-teacher meetings contributed to the students learning, and more than 90% believed that they should continue (Table A14)”. When we reproduce Table A14 (using the dataset `Table A14_EER.dta`), however, we find that these two questions were not only asked for

---

<sup>12</sup>The value labels for the variable indicating treatment status (`sch_type`) further indicates that this variable has been redefined. According to the value labels, 1 indicates *Treatment* while 2 indicates *Control*, but for this analysis, *Control* appears to have been recoded to 0.

<sup>13</sup>The variable that measures absence in all visits in 2012 is called `abs_teach12`.

the treatment schools (as claimed by the author) but also for the control schools, which should not have received these meetings at all.

As shown in Table 5, we find that the respondents in the control schools were even more positive about the treatment (i.e., parent-teacher meetings) than the treated respondents. While 93% in the treatment schools thought that the one-to-one meetings contributed to improving the child’s study, the corresponding number in the control schools was 98%. Similarly, 91% and 93% in the treated and untreated schools, respectively, thought that there should be a monthly parent-teacher meeting in school.

Table 5: Treated and control parents’ evaluations of parent-teacher meetings.

	% of parents	
	Treatment schools	Control schools
Do you think that the parent meetings help to improve a child’s study?	93.36	98.06
Do you think that there should be monthly parent meetings in school?	91.03	93.46
Missing or Not Applicable (NA)	7.66	5.55

*Note:* The results correspond to Table A14 in [Islam \(2019\)](#) but are extended with data from control schools, which are contained in the replication package. The number of observations is 3,132 for treatment schools and 2,127 for control schools.

If anything, one would also expect the number of missing observations or “*Not Applicable*” (NA) responses to these questions to be higher in control than in treatment schools. On the contrary, this number is higher in the treatment schools, with 240 missing/NA responses (8%), while the corresponding number in the control schools is 118 missing/NA responses (6%).

#### 4.10 A comment on the costs of the intervention

As a smaller non-technical comment, we also note that the author’s explanation of the intervention includes the following statement: “*Each meeting between a parent and a teacher was one-on-one and lasted about 15 min*” ([Islam 2019](#), p. 276). At the same time, the average number of students per treated class (according to the file `roster_all_final.dta`) is 59 students. This means that, for a teacher in a treated

class, almost 15 hours per month would have to be spent in these parent-teacher meetings (corresponding to an additional working time of more than 9%, assuming a normal working time of 40 hours per week)—excluding the extra workload due to additional tests, reporting, administration, etc. Still, they were only paid an extra \$2 per year (or 1.5% of the average yearly salary, according to the numbers provided in Islam 2019, p. 277) for all this additional work. In other words, the author’s conclusion that “[t]he intervention is remarkably low cost” (Islam 2019, p. 290) does not seem fair.

## 5 Robustness checks

### 5.1 Main analyses with alternative baselines as control variables

We conduct robustness checks by replacing the baseline test scores from the analysis files with the alternative baseline variables (`*_b`) as control variables in the main outcome regressions (see Section 4.3 for details). The results are summarized in Table A4. Column (1) presents the results using the baselines from Islam (2019) for reference.<sup>14</sup> Column (2) shows the results for the subset of individuals with non-missing values for both sets of baselines, again using Islam (2019)’s baseline variables. This intermediary step confirms that the sample restriction has minimal impact on the coefficients, which retain the same signs and similar magnitudes and remain robust in terms of statistical significance. Finally, Column (3) reports the robustness check using the same subset as in Column (2) but controlling for the alternative baseline variables (`*_b`).

The findings indicate that in grade 4, 2011, the results for Bengali become non-significant, while other outcomes remain largely unchanged. In grade 3, 2012, the result for math becomes significantly negative, and the coefficient for English is attenuated by approximately 40% while remaining statistically significant at the 10% level. In grades 5, 2012, the results are robust for all subjects.

---

<sup>14</sup>These results differ slightly from those in Tables 5 and 6 in Islam (2019) due to the exclusion of individuals with missing or duplicate IDs.

## 5.2 Panel data analysis of midline results using final marks from 2010 as pre-treatment data

Finally, we examine the 2010 and 2011 final marks variables for all six subjects identified in the 2011 household data files (see Section 4.1), including social science and religion. Given the data collection timing described in Islam (2019), we should be able to consider the 2010 data as pre-treatment versions of the midline outcomes. We exploit these data to construct a panel dataset for grade 4, 2011, students to follow the same individuals over time and perform a difference-in-differences (DID) analysis.

The results, presented in Table 6, reveal large baseline differences in standardized test scores for mathematics, social science, and science in 2010, with treatment group scores exceeding controls by 0.35 to 0.53 standard deviations. Column (2) shows midline results (2011), including the two previously unreported subjects. Notably, both social science and religion exhibit lower scores in treatment schools, though only religion is significantly different at the 10% level. Column (3) presents the DID results with individual and time fixed effects. We find significant negative changes for treatment schools in mathematics, science, social science, and religion between pre- and post-intervention tests, with the only positive change occurring in English. These results are in contrast with the main findings of the paper and—if taken at face value—would imply that the intervention is detrimental to students’ test scores. However, it appears reasonable to assume these results are biased due to the randomization failures described in Section 3.

## 6 Conclusion

The original study by Islam (2019) investigates the effects of structured parent-teacher meetings on student test scores in rural Bangladesh. Claiming to use a randomized controlled trial (RCT), the study reports substantial improvements in student performance across multiple subjects, with particularly strong gains in

Table 6: Panel data analysis for six subjects (Grade 4, 2011).

Subject	(1) Before treatment (2010)		(2) Midline (2011)		(3) DID Analysis	
	Std. Coef (SE)	N	Std. Coef (SE)	N	DID Coef (SE)	N
English	0.015 (0.121)	1434	0.361*** (0.130)	1529	0.293*** (0.104)	1400
Bengali	0.081 (0.116)	1435	0.311*** (0.109)	1530	0.164 (0.101)	1402
Mathematics	0.516*** (0.127)	1435	0.302** (0.115)	1527	-0.263** (0.104)	1399
Science	0.534*** (0.133)	1434	0.122 (0.117)	1528	-0.480*** (0.126)	1399
Social Science	0.351** (0.163)	913	-0.136 (0.129)	1110	-0.471*** (0.170)	748
Religion	0.186 (0.118)	1355	-0.223* (0.130)	1444	-0.438*** (0.109)	1240

*Note:* Reported are standardized coefficient estimates (*b*) and cluster robust standard errors (*se*; in parentheses). \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

mathematics and English. The intervention aimed to enhance parental engagement by providing detailed feedback on student progress, and the study concludes that such meetings are an effective tool for improving educational outcomes in developing countries.

Our reproduction effort, however, uncovered several significant data irregularities that challenge these conclusions. First, we uncovered randomization failures in both sampling and treatment allocation. Second, we identified inconsistencies in baseline test scores, with systematic discrepancies between datasets that varied by treatment status. Third, unreported pre-treatment data from 2010 suggest that the control and treatment groups were not initially balanced, which, when accounted for, renders several of the original study's key findings either insignificant or flips the sign on key effect estimates. Additionally, mismatches in student IDs, selection biases in household survey data, and anomalies in the reported timing of data collection further undermine the reliability of the dataset. We also note other inconsistencies and data irregularities throughout our commentary.

Overall, these irregularities are detrimental to the credibility of the study. The

non-randomized treatment allocation, systematic discrepancies in baseline scores, and the unreported pre-treatment data indicate that the positive treatment effects reported by Islam (2019) should be interpreted with caution.

## References

- Begum, L., Grossman, P. J. and Islam, A.: 2014, Identifying gender bias in parental attitude: An experimental approach. Monash University Department of Economics Discussion Paper 32/14 (Preprint).
- Begum, L., Grossman, P. J. and Islam, A.: 2018, Gender bias in parental attitude: An experimental approach, *Demography* **55**(5), 1641–1662.
- Begum, L., Grossman, P. J. and Islam, A.: 2022, Parental gender bias and investment in children’s health and education: Evidence from Bangladesh, *Oxford Economic Papers* **74**(4), 1045–1062.
- Brodeur, A., Mikola, D., Cook, N. et al.: 2024, Mass reproducibility and replicability: A new hope. I4R Discussion Paper 107 (Preprint).
- IPEMIS: 2025, Primary school directory. Integrated Primary Education Management Information System (IPEMIS). <https://ipemis.dpe.gov.bd/search-school> (accessed February 16, 2025).
- Islam, A.: 2019, Parent-teacher meetings and student outcomes: Evidence from a developing country, *European Economic Review* **111**, 273–304.
- Islam, A.: 2024, Dos and don’ts when implementing randomized controlled trials in developing countries. CDES Working Paper 02/24 (Forthcoming in the Handbook of Experimental Development Economics).

## Appendices

### A Replication package structure

Our reproduction and robustness analyses rely on the publicly available replication package for [Islam \(2019\)](#), provided as supplementary material accompanying the article. The package includes annotated *Stata* code but lacks general instructions or data management code to trace the creation of final datasets from raw data.

The package contains two *Stata* `.do` files for reproducing results in [Islam \(2019\)](#), divided into analyses for 2011 and 2012:

- `EER_Baseline_midline_2010_11.do`
- `EER_Endline2012_13.do`

Additionally, it includes 18 *Stata* `.dta` files, with student IDs enabling matching across files. We infer from the `.do` files that two datasets are central to reproducing main results:

- `testscore2011_2010_final.dta`
- `testscore2012_Final.dta`

[Islam \(2019\)](#), pp. 288–289) also reports a follow-up household survey based on a random sample of about 50% of the students included in the trial. Data from this survey is linked to individuals in year-specific household survey data files:

- `hh13a_meet_teac_sch11_final.dta`
- `hh13a_meet_teac_sch12_final.dta`
- `hh2013_FINAL1.dta`
- `hh2013a_meet2011.dta`
- `hh2013a_report_meeting2012.dta`

Additionally, the replication package contains a file named `roster_all_final.dta`, which includes information on the geographical locations of the included schools.

We provide a comprehensive data dictionary, including observations, variables, labels, unique values, and missing values for each of the 18 datasets, on our Open Science Framework (OSF) repository. The repository also contains code to reproduce all analyses in this report.

## **B Computational reproducibility**

Using the provided code in the author’s replication package, we successfully reproduce most of the main findings reported in Table 5 (2011) and Table 6 (2012) in [Islam \(2019\)](#). However, we identify errors in Table 5, column (6) for standardized GPA (grade 5, females, 2011). The reported estimate, 0.209 (SE 0.123), significant at the 10% level, matches the math results for grade 4 (Table 5, column 1) instead of the correct value, 0.020 (SE 0.112), which is statistically insignificant. The *Male-Female diff* coefficient in column (6) is also incorrect (reported: 0.139; correct: 0.330).

## **C Matching datasets**

To conduct additional analyses, we match datasets from the household survey and analysis files. The matching process is illustrated in the flowcharts in Figures [A1](#) and [A2](#). For 2011, we identify three household survey files. For 2012, two similar files are matched internally by year to create master household data files. While matching succeeds for 2012, duplicate IDs in 2011 files required removing non-distinct or missing IDs. The master household files are then matched to the main analysis files, with individuals with non-distinct or missing IDs being dropped.

## **D Permutation test for unique values**

In Section [4.4](#), we observed that baseline distributions in the analysis data files included common values that were unique to either the treatment or the control



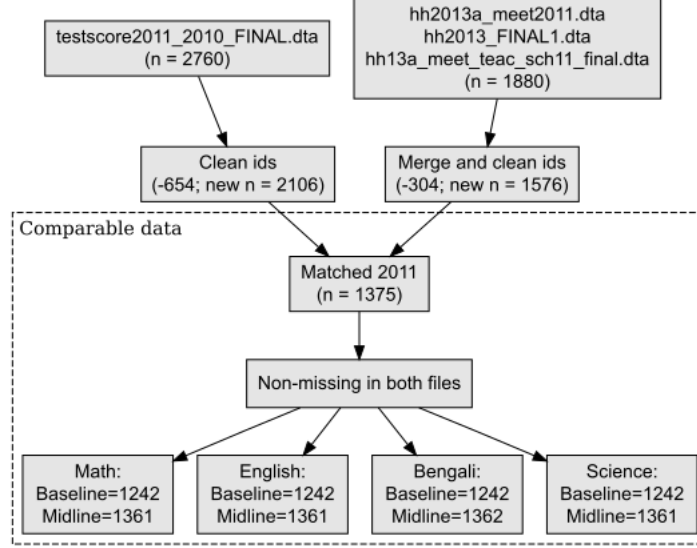


Figure A1: Flowchart for grade 4, 2011. Datasets refer to [Islam \(2019\)](#)’s replication package.

group.

To assess the likelihood of such patterns arising by chance, we conduct a permutation test. We randomly reassign treatment to the same number of schools as in [Islam \(2019\)](#) ( $n = 40$ ) multiple times, mimicking the original randomization strategy. Our test statistic is defined as:

$$S_{\text{obs}} = \frac{1}{n} \sum_{i=1}^n I(X_i \text{ is unique to either treatment or control})$$

We compute this statistic for each permutation  $S^{(b)}$  over  $B = 10,000$  iterations and calculate the  $p$ -value as:

$$p_S = \frac{1}{B} \sum_{b=1}^B I(S^{(b)} \geq S_{\text{obs}})$$

This  $p$ -value represents the proportion of permutations where the share of individuals with values unique to either the treatment or control group matches or exceeds that observed in [Islam \(2019\)](#)’s data.

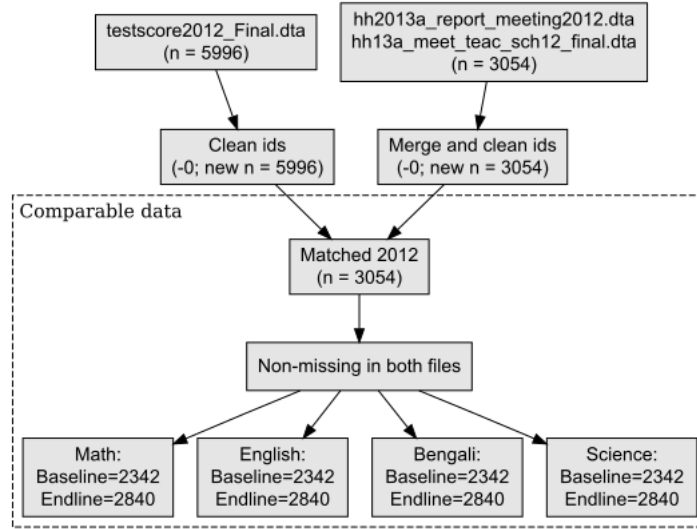


Figure A2: Flowchart for grade 3 and 5, 2012. Datasets refer to [Islam \(2019\)](#)'s replication package.

## E Additional figures and tables

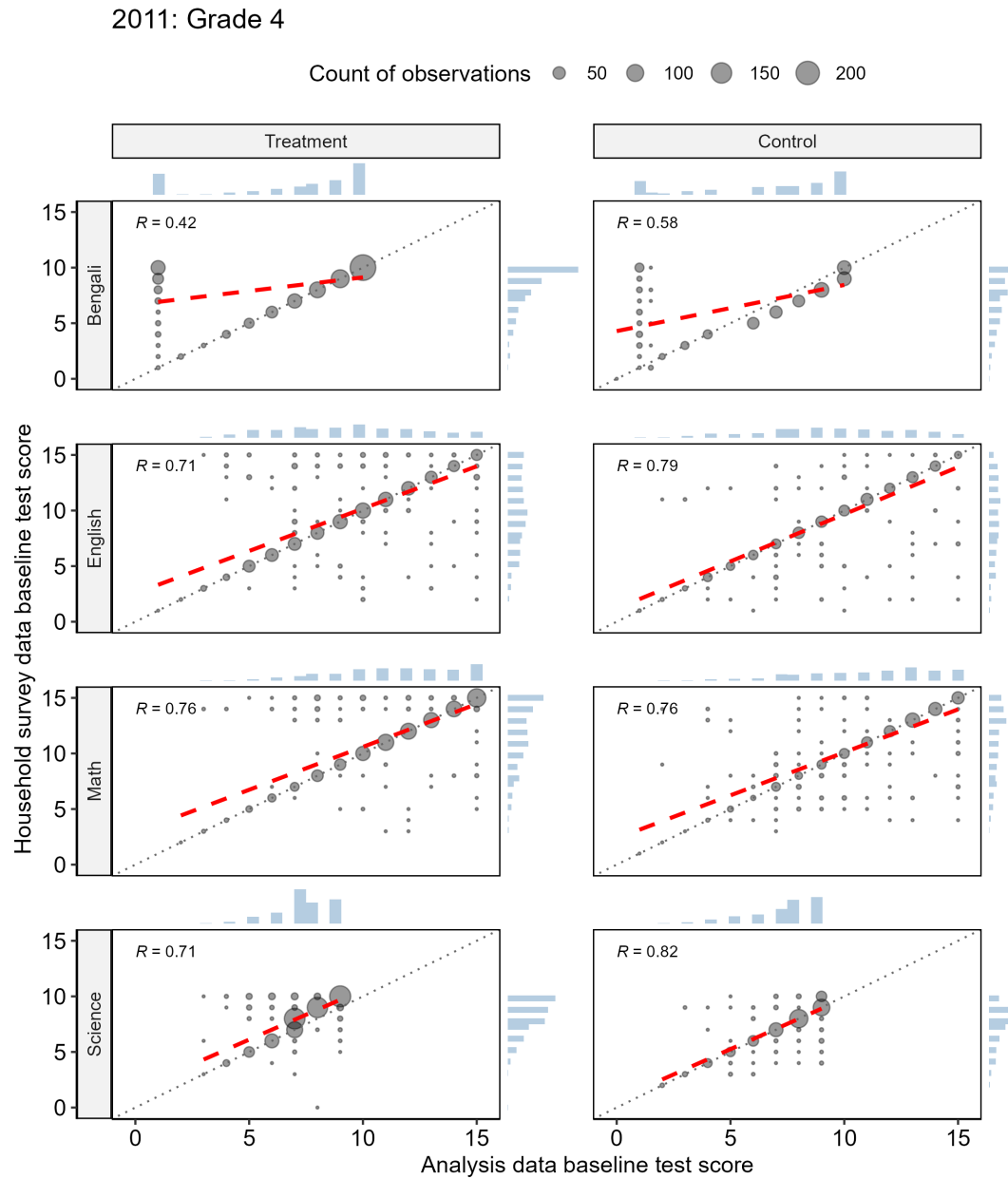


Figure A3: Comparison of baseline test scores in grade 4, 2011, from [Islam \(2019\)](#)'s analysis files (variables named \*10b) and alternative baseline variables identified in the household survey files (named \*\_b) available in the author's replication package.

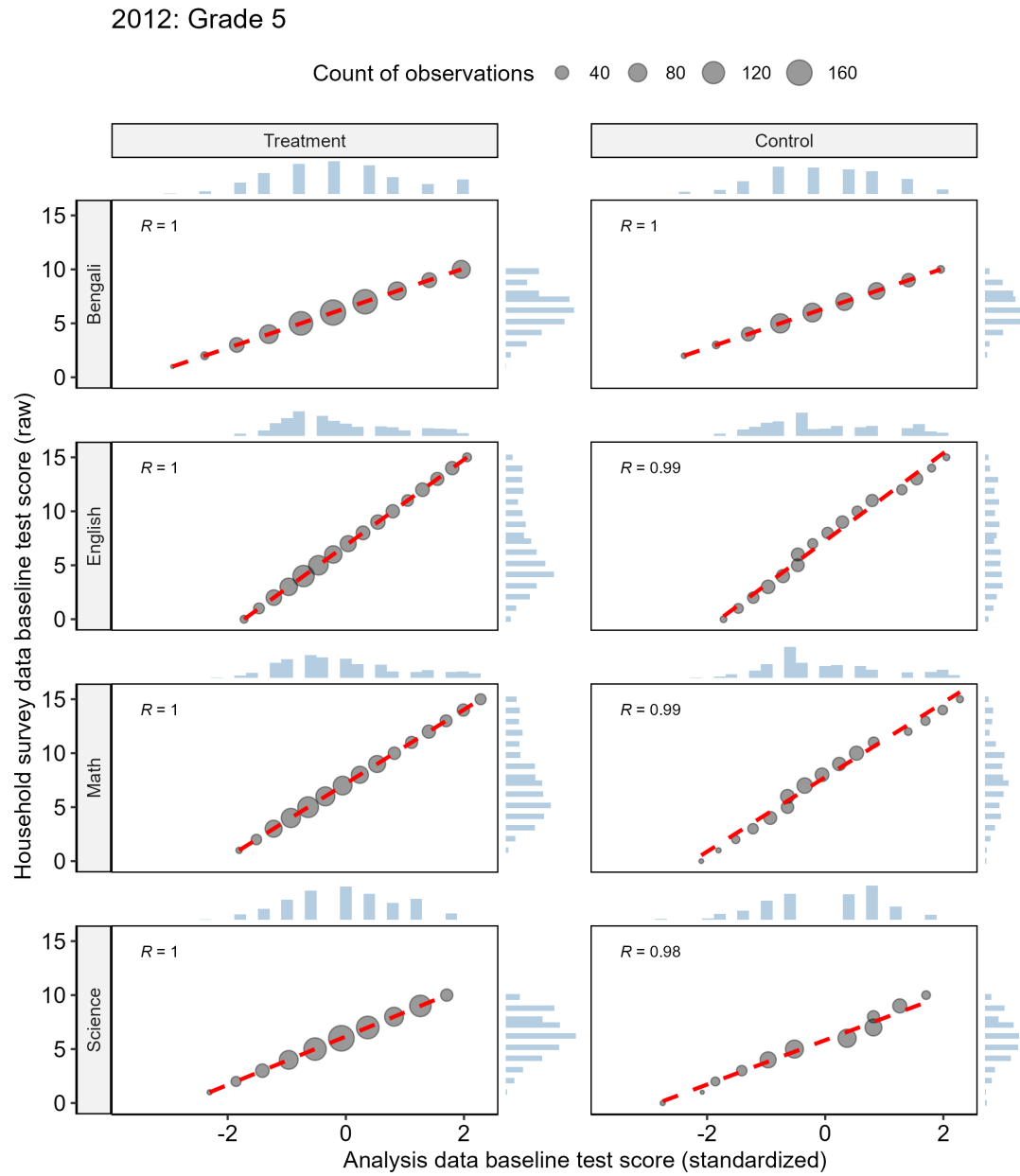


Figure A4: Comparison of baseline test scores in grade 5, 2012 from [Islam \(2019\)](#)'s analysis files (variables named \*12b1) and alternative baseline variables identified in the household survey files (named \*\_b) available in the author's replication package.

## 2011: Grade 4

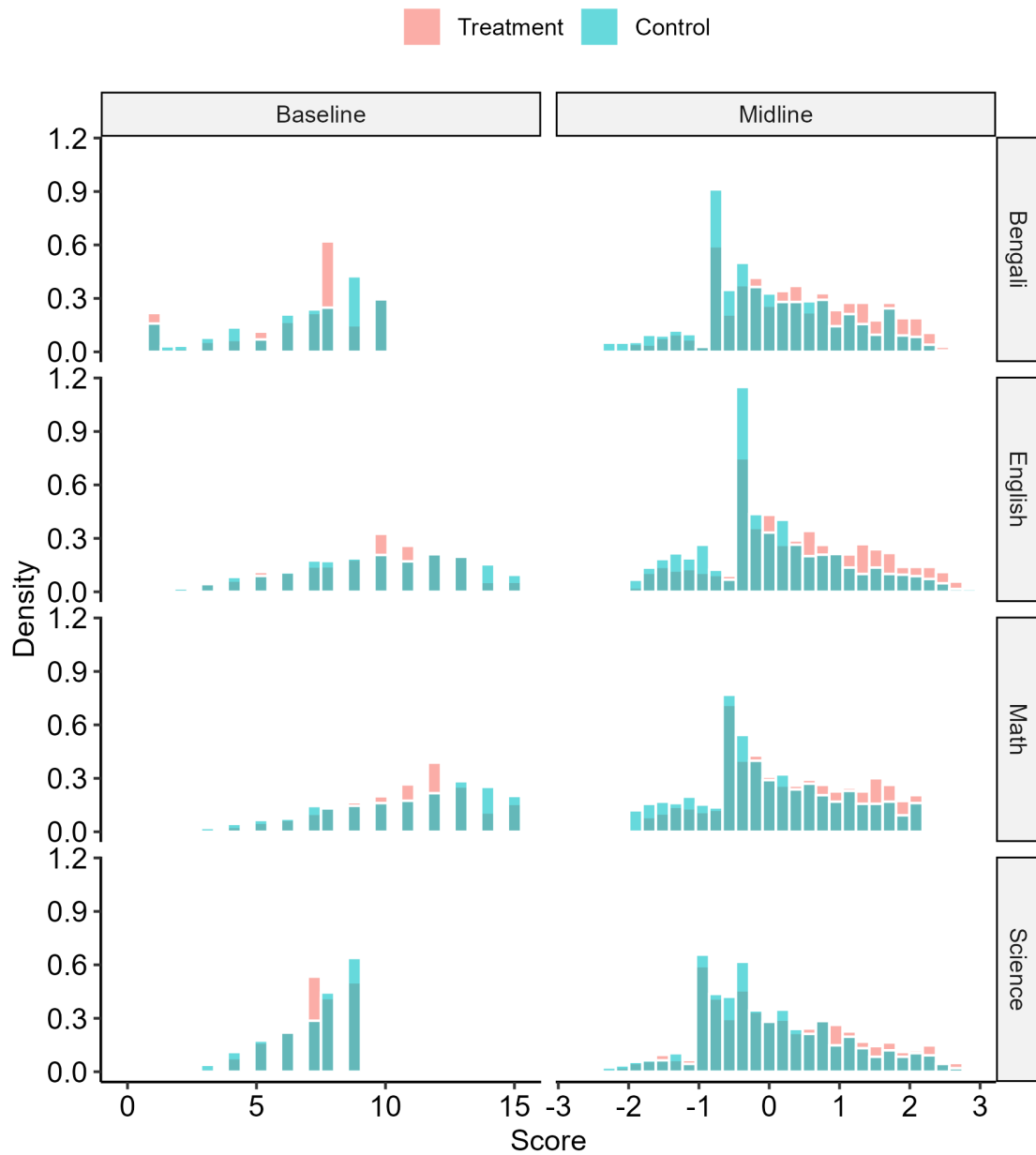


Figure A5: Histograms of the baseline and midline test scores for grade 4, 2011, by treatment group and subject.

2012: Grade 5

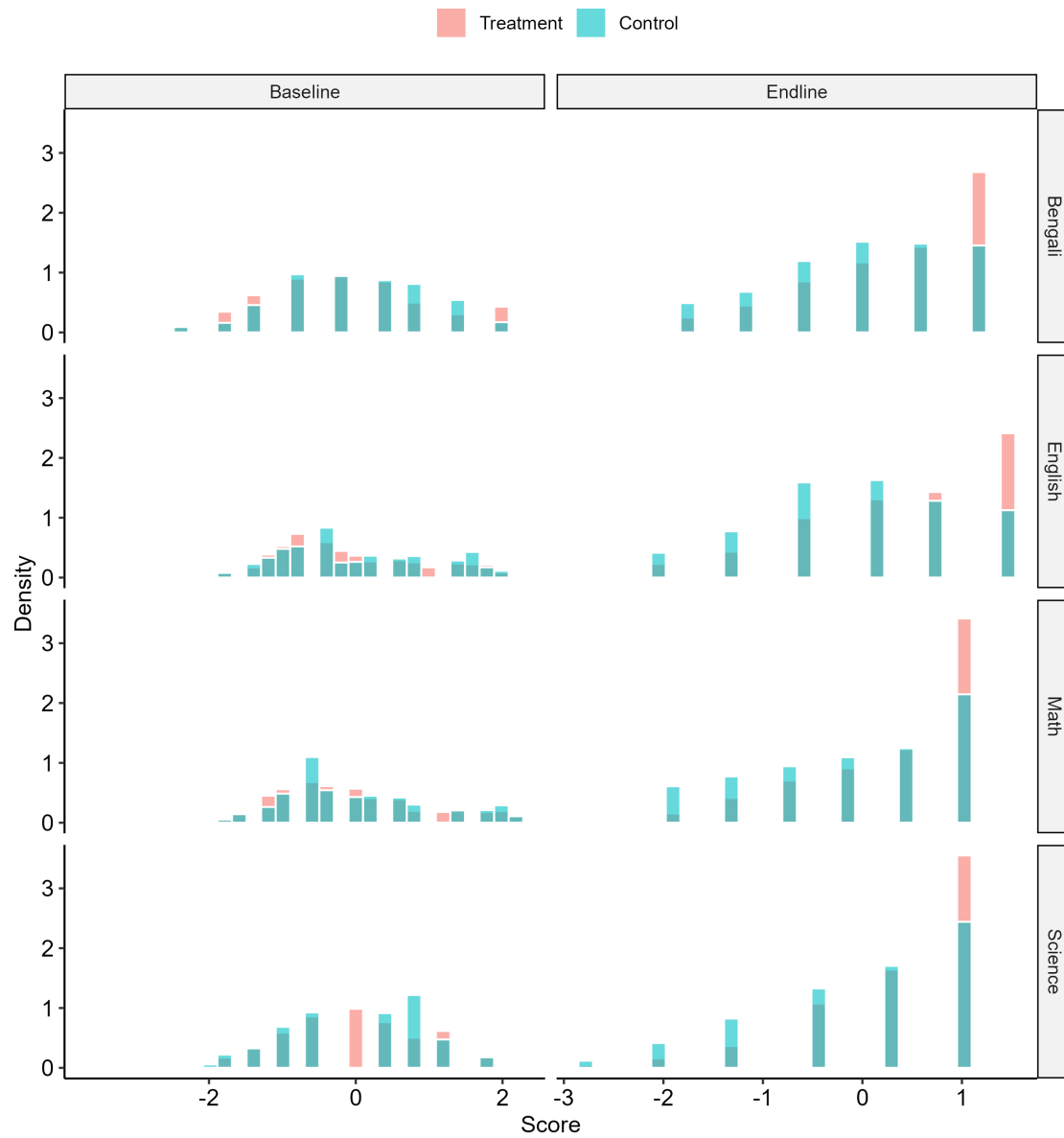
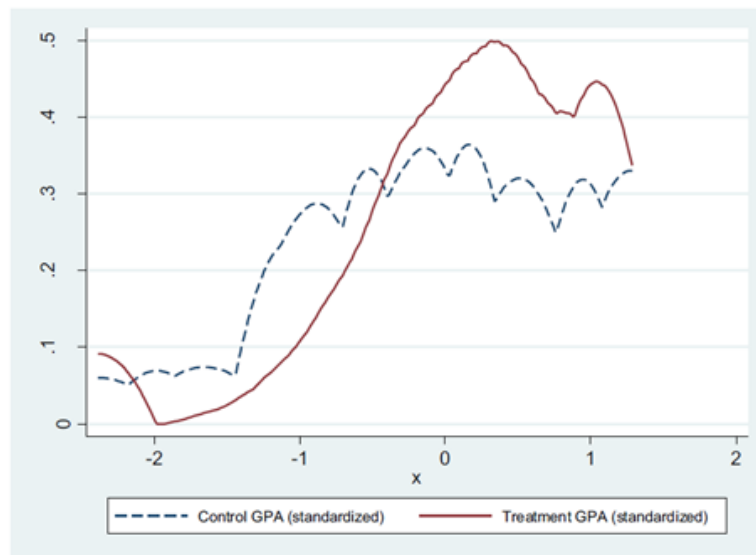


Figure A6: Histograms of the baseline and midline test scores for grade 5, 2012, by treatment group and subject.



**Fig. A4.** Distribution of Cumulative GPA Test Scores in Year 1.  
 Note: Figure shows GPA test scores distribution of treatment and control school children of grade 5 students at the end of year 1 of the intervention.

Figure A7: Screenshot of Figure A4 in [Islam \(2019\)](#).

Table A1: Treatment comparison between Islam (2019) and Begum et al. (2022).

Islam (2019)	Begum et al. (2022)	Islam (2019)	Begum et al. (2022)
<i>sch_id</i> school ID	<i>sch_id</i> id indicating village	<i>sch_type</i> type of school (1 treatment, 2 control)	<i>treat</i> Treatment group
1 Naksa GPS	1 naksa gps	1	
2 Masjidkuri GPS	2 masjidkuri gps	1	
3 Amadi GPS	3 amadi gps	1	
4 Shahid SA Gafur GPS	4 shahid sa gafur gps	1	
5 Hatiar Danga GPS	5 hatiar danga gps	1	
6 Kinukathi GPS	6 kinukathi gps	1	
7 Channir Chak GPS	7 channir chak gps	1	
8 Khaona GPS	8 khaona gps	1	
9 Vanderpul GPS	9 vanderpul gps	1	
10 Jargir Mohal GPS	10 jargir mohal gps	1	
11 Kalmibunia GPS	11 kalmibunia gps	1	1
12 Shapara GPS	12 shapara gps	1	1
13 Gajalia GPS	13 gajalia gps	1	1
14 Chandkhali GPS	14 chandkhali gps	1	1
15 Fatapur GPS	15 fatapur gps	1	1
16 Kaliar Abad GPS	16 kaliar abad gps	1	1
17 Mukhali GPS	17 mukhali gps	1	
18 Dhamrail GPS	18 dhamrail gps	1	1
19 Dabduor Sheikh GPS	19 dabduor sheikh gps	1	1
20 Kisnagar GPS	20 kisnagar gps	1	1
21 K.D S. GPS	21 k.d s. gps	1	1
22 Kharia Binapani GPS	22 kharia binapani gps	1	1
23 Karulia Somiuddin GPS	23 karulia somiuddin gps	1	1
24 Kharia Pallimongal GPS	24 kharia pallimongal gps	1	1
25 Kharia milan bithi GPS	25 kharia milan bithi gps	1	1
26 Laskor GPS	26 laskor gps	1	1
27 Kharia Khalpar GPS	27 kharia khalpar gps	1	1
28 Laxikhola GPS	28 laxikhola gps	1	1
29 Alamtola GPS	29 alamtola gps	1	1
30 Kathipara GPS	30 kathipara gps	1	1
31 Borhanpur GPS	31 borhanpur gps	1	1
32 Raruli GPS	32 raruli gps	1	1
33 Baka Bonipur GPS	33 baka bonipur gps	1	1
34 Shrikantipur GPS	34 shrikantipur gps	1	1
35 Baka Pachim GPS	35 baka pachim gps	1	1
36 Baka Purb GPS	36 baka purb gps	1	1
37 Purbkatipara GPS	37 purbkatipara gps	1	1
38 Basakhali GPS	38 basakhali gps	1	1
39 Gogla Rochak GPS	39 gogla rochak gps	1	1
40 Bagul Rochak GPS	40 bagul rochak gps	1	1
41 Hugli Shahpara GPS	41	2	
42 Bagali GPS	42	2	2
43 Ghugarkati GPS	43	2	2
44 Chirifultola GPS	44	2	
45 S. Islampur GPS	45	2	2
46 Bamia GPS	46	2	2
47 Sorishamut Gazinagar GPS	47	2	2
48 Baropota Chatkatola GPS	48	2	
49 Fatakati PS	49	2	
50 BogaG PS	50	2	
51 Basakhali GPS	51	2	
52 Bholanathpur GPS	52	2	
53 Khajra GPS	53	2	2
54 Laotara GPS	54	2	
55 Pirojpur GPS	55	2	2
56 Khalia GPS	56	2	2
57 Pashomari GPS	57	2	
58 Kashun Danga GPS	58	2	2
59 Gadaipur GPS	59	2	2
60 Dakhin Gadaipur GPS	60	2	2
61 Toara Danga GPS	61	2	2
62 Gugumari GPS	62	2	2
63 Uttar borodal GPS	63	2	2
64 Boro Dal Dakhin GPS	64	2	2
65 Baintola GPS	65	2	2
66 Mardia GPS	66	2	2
67 Jamal NagarGPS	67	2	2
68 Basun danga GPS	68	2	2
69 Champakhali GPS	69	2	2
70 Fakrabad GPS	70	2	2
71 kathbashia GPS	71	2	2
72 Anulia GPS	72	2	2
73 Bishat GPS	73	2	
74 Bholanathpur GPS	74	2	2
75 Aksor GPS	75	2	2
76 Nakla GPS	76	2	
77 Bagali GPS	77	2	2



Table A2: Results of a permutation test randomly permuting the school-level treatment assignment to baseline test scores 10,000 times.

Variable	Dataset year	Grade	$S_{obs}$	p-value
math10b	2011	4	0.002	0.512
eng10b	2011	4	0.001	1.000
beng10b	2011	4	0.036	0.000
sci10b	2011	4	0.001	1.000
math_b	2011	4	0.001	1.000
eng_b	2011	4	0.000	1.000
ben_b	2011	4	0.002	0.509
sci_b	2011	4	0.005	0.276
stdmath12b1	2012	3	0.138	0.000
stdeng12b1	2012	3	0.150	0.000
stdben12b1	2012	3	0.069	0.000
stdsci12b1	2012	3	0.066	0.000
math_b	2012	3	0.000	1.000
eng_b	2012	3	0.000	1.000
ben_b	2012	3	0.000	1.000
sci_b	2012	3	0.000	1.000
stdmath12b1	2012	5	0.020	0.002
stdeng12b1	2012	5	0.019	0.253
stdben12b1	2012	5	0.003	0.060
stdsci12b1	2012	5	0.118	0.000
math_b	2012	5	0.002	0.249
eng_b	2012	5	0.000	1.000
ben_b	2012	5	0.002	0.496
sci_b	2012	5	0.003	0.169

*Note:*  $S_{obs}$  is the observed share of individuals in the data that have a baseline test score value that is unique to either the treatment or the control group. The  $p$ -value shows how many of the 10,000 permutations yield shares that are greater than or equal to the observed value.

Table A3: Adjusted reproduction of *Table A11: Teacher absences in 2011*.

Teacher absences in 2011 (Year 1)	Original (1)	Adjusted reproduction (2)
April	0.03 (0.031)	−0.004 (0.029)
May	−0.01 (0.025)	−0.005 (0.023)
June	0.04 (0.026)	0.042* (0.025)
August	−0.09*** (0.033)	−0.059* (0.031)
Absence in all visits in 2011	−0.03 (0.065)	−0.025 (0.061)

*Note:* Column (1) shows the original results as reported in [Islam \(2019\)](#). Column (2) shows the results from an adjusted reproduction, where the schools have been defined as treatment and control in accordance with the other analyses. Both columns report  $t$  test results (T–C) with standard errors in parentheses. \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

Table A4: Robustness checks with alternative baseline score sets (original baselines, complete case subset with non-missing values on both baseline variable sets, and alternative baselines).

Year	Grade	Subject	(1) Orig. BL		(2) Orig. BL, CC		(3) Alt. BL, CC	
			$b$ ( $SE$ )	$N$	$b$ ( $SE$ )	$N$	$b$ ( $SE$ )	$N$
2011	4	Math	0.201* (0.110)	1804	0.252** (0.116)	1231	0.246** (0.116)	1231
2011	4	English	0.323** (0.127)	1806	0.332** (0.133)	1231	0.320** (0.125)	1231
2011	4	Bengali	0.262** (0.101)	1807	0.284*** (0.106)	1232	0.163 (0.105)	1232
2011	4	Science	0.124 (0.109)	1806	0.112 (0.127)	1231	0.005 (0.129)	1231
2012	3	Math	-0.087 (0.079)	2253	-0.050 (0.087)	1066	-0.211** (0.092)	1066
2012	3	English	0.317*** (0.091)	2253	0.339*** (0.101)	1066	0.196* (0.110)	1066
2012	3	Bengali	-0.006 (0.107)	2253	-0.010 (0.114)	1066	-0.126 (0.117)	1066
2012	3	Science	0.332*** (0.107)	2253	0.351*** (0.110)	1066	0.265** (0.117)	1066
2012	5	Math	0.420*** (0.129)	1870	0.386*** (0.142)	1178	0.417*** (0.142)	1178
2012	5	English	0.413*** (0.126)	1870	0.308** (0.144)	1178	0.327** (0.142)	1178
2012	5	Bengali	0.309*** (0.100)	1870	0.257** (0.111)	1178	0.257** (0.111)	1178
2012	5	Science	0.339*** (0.106)	1870	0.270** (0.119)	1178	0.235* (0.121)	1178

*Note:* Reported are standardized coefficient estimates ( $b$ ) and cluster robust standard errors ( $SE$ ; in parentheses). *Orig. Baselines (BL)* refers to **\*10b** and **12b1** variables. CC = Complete case subset with non-missing values on both baseline variable sets. *Alt. Baselines (BL)* refers to **\*\_b** variables. Estimates for grade 4, 2011, presented in column (1), differ slightly from the estimates in [Islam \(2019\)](#)'s Table 5 due to individuals lost when cleaning out those with duplicated or missing IDs in the 2011 files. \*  $p < 0.10$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .