Teacher Flexibility and School Productivity: Remedial Secondary Education in India

Sabrin Beg, Anne Fitzpatrick, Jason Kerwin, Adrienne M. Lucas,

and Khandker Wahedur Rahman*

January 31, 2023

Abstract

Public education in developing countries is often deficient, leading to increasing learning deficits as students age. A trade-off exists between ensuring uniformly high standards and introducing reforms that allow teachers flexibility. Through a 300 school RCT in Odisha, India, we compare the effects on Class 9 students of T1) rigidly defined remedial lessons that take time away from the curriculum, T2) teacher determined remedial lessons, or T3) control. Both interventions increased students test scores 0.11SD, about 60 percent of a year of learning, with gains throughout the learning distribution. The quality of implementation was high in both arms. Few teachers took advantage of the flexibility offered and defaulted into the regimented version.

JEL Codes: H40, I21, I25, J24, J28, M50, M53, M54, O15, O43.

Keywords: teacher autonomy, remedial education, secondary school, India

*Beg: University of Delaware. Fitzpatrick: University of Massachusetts Boston. Kerwin: University of Minnesota and J-PAL. Lucas: University of Delaware, J-PAL, and NBER. Rahman: University of Oxford and BIGD, BRAC University. This evaluation was funded by Kusuma Trust UK. All study activities were approved by the following Institutional Review Boards: IFMR, University of Minnesota, University of Delaware, and University of Massachusetts Boston. This evaluation would not be possible without our partners at Transform Schools, People for Action and the School and Mass Education Department of Odisha, including Pankaj Sharma, Bindiya Nagpal, Christine Oliver, and Kartik Sahu. We thank Sandra Sequiera and seminar participants at IFPRI, Lafayette College, Boston University, Universite Laval, and University of Nevada Reno for their comments. We also thank the JPAL-South Asia field operations staff, and all of our respondents for their time, cooperation, and insight. Soumya Das provided excellent research assistance.

1 Introduction

Public services in developing countries often have the simultaneous issues of a hierarchical bureaucracy where front line civil servants (e.g., teachers, health workers) are guided by rigid expectations about service delivery while also delivering insufficient services (Afridi, Iversen and Sharan, 2017; Bandiera, Best, Khan and Prat, 2020; Banerjee, Chattopadhyay, Raghabendra, Duflo, Esther, Keniston, Daniel and Singh, Nina, 2021; Muralidharan and Singh, 2020). Indian government schools are a classic example where both features are salient. Teachers (i.e., front line civil servants) in Indian government schools are directed to follow and complete the grade-level curriculum on a strict timetable. At the same time, many Indian secondary school students are multiple grade-levels behind and have on average among the lowest scores on international assessments (Das and Zajonc, 2010; Muralidharan, Singh and Ganimian, 2019; Muralidharan and Singh, 2020). Why are so many students behind grade level? The current system of standardized teaching approaches and uniform implementation could create high expectations of teachers, decrease monitoring and coordination costs, forestall issues of the moral hazard of teacher shirking, and be practical given concerns about teacher competency, especially at the secondary school level (Chaudhury, Hammer, Kremer, Muralidharan and Rogers, 2006; Duflo, Dupas and Kremer, 2015; Duflo, Hanna and Ryan, 2012)(Beteille, Ding, Molina, Pushparatnam and Wilichowski, 2020; Bold, Filmer, Martin, Molina, Stacy, Rockmore, Svensson and Wane, 2017a; Bold, Filmer, Martin, Molina, Rockmore, Stacy, Svensson and Wane, 2017b; Bold, Molina, Filmer and Svensson, 2018). This lack of autonomy for local teachers could also be stifiling creativity and intrinsic motivation (Pearson and Moomaw, 2005; Skaalvik and Skaalvik, 2014). Alternatively, teachers might have no interest in deviation if they are unaware that their students are behind grade level. Or teaching below grade level could require additional uncompensated effort that has no return on the govenment's high stakes, grade level tests. Finally, taking time out of the school day to teach remedial lessons could decrease grade level competencies. We use an randomized controlled trial in secondary schools in Odisha, India to test the effects

on time use in school and student learning of a program that lowered the effective costs of providing remedial lessons (through training, materials, and an explict time table) and allowed some teachers flexibility out of the program.

Specifically we partnered with the Odisha Department of School and Mass Education and the Indian non-governmental organization Transform Schools, People for Action to test how two policy choices compared to each other and the status quo: 1) training and materials for a structured remedial curriculum program or 2) the same training and materials with more flexiblity around adoption. We randomized 300 government secondary schools across these two arms with 100 in each arm and the final 100 in a control, business as usual arm. Math, English, Science, and Odia (local language) teachers in randomly selected treatment schools received training and materials to support the implementation of Utkarsh–excellence– a within school bookcamp program that set aside a few hours over a set number of days to focus on remedial skills. The second treatment, provided this same training but allowed teachers a more flexible teaching plan – for part of the intervention they could select on which modules to linger or which modules to skip–Flexible Utkarsh. The final arm was business as usual.

We have five main findings. First, students are substantially behind grade level – an average Class 9 student is 4 to 5 grades behind in English, in math, and Odia with substantial within school heterogeneity. In contrast to other findings, teachers are largely but imprecisely aware of this issue (see additional discussion below). Second, both treatment arms improved student learning by about 0.1 standard deviations (SDs) overall and did nor crowd-out grade level knowledge, nor did it bring students all the way up to grade level mastery. In cost effectiveness, this was 0.95 SDs per $100 at the 200 school scale. Third, implementation quality was high in both treatment arms – despite no additional incentives, teachers implemented the program and flexibility did not result in shirking. Fourth, when given the option of flexbility, teachers largely stuck to the prescribed plan. Finally, even though teachers in treatment schools believed that they and their students benefitted from

the program, they assessed their students as having a lower learning level.

In addition to its policy relevance as this program has been scaled within Odisha (population 40 million) and is under consideration in Karnataka (population 60 million) in India, this paper makes three major contributions to the economics literature. First, we show that remedial instruction with existing teachers at the secondary level can increase student learning. Remedial instruction as tracking, pull-out, or as a program outside the school day increases student learning in primary schools (Banerjee, Banerji, Berry, Duflo, Kannan, Mukerji, Shotland and Walton, 2017; Banerjee, Banerji, Duflo, Glennerster and Khemani, 2010; Banerjee, Cole, Duflo and Linden, 2007; Duflo, Kiessel and Lucas, 2020; Duflo, Dupas and Kremer, 2011; Lakshminarayana, Eble, Bhakta, Frost, Boone, Elbourne and Mann, 2013). In contrast, limited evidence exists for remedial instructional in at the secondary school level. Existing approaches are either out side of the school day (Muralidharan et al. (2019)) or focus on grade level material (Beg, Lucas, Halim and Saif, 2019). We show that improving learning with existing personnel is possible in resource-poor secondary schools.

Second, we provide new evidence on whether allowing point-of-service modifications improve the functioning in highly rigid bureaucracies in low-income settings. Much of the research in improving public sector service delivery focuses on providing incentives to service providers (Rasul and Rogger, 2018; Rasul, Rogger and Williams, 2018) (Barrera-Osorio and Raju, 2017; Brown and Andrabi, 2020; Duflo, Dupas and Kremer, 2011, 2015; Duflo, Hanna and Ryan, 2012; Glewwe, Ilias and Kremer, 2010; Muralidharan and Sundararaman, 2011; De Ree, Muralidharan, Pradhan and Rogers, 2018)) or empowering community members to register complaints (Bjorkman and Svensson, 2009; Duflo et al., 2012). A growing literature emphasizes that highly rigid bureaucracies may alternatively improve service delivery by reallocating tasks and specifically by increasing the autonomy among front-line civil servants (Bandiera, Best, Khan and Prat, 2020; Bloom, Lemos, Sadun and Reenen, 2015), though it is difficult to find optimal practices in a hierarchical bureaucracy (Banerjee et al., 2021). Whether these reforms would improve educational quality, however, is unclear:

(Piper, Sitabkhan, Mejia and Betts, 2018) find that guided lessons improve students' learning, but teacher modifications decrease lesson quality. We show that providing flexibility to the service providers does not differently facilitate or impede productivity improvements, nor result in increased shirking. Contrary to a widespread notion that teachers want full control over what and how they teach, we find that the demand for lesson flexibility is low.

Third, we use a conceptual framework to show that teachers are teaching grade-level content even when their students are well behind grade level for potentially three reasons that this intervention alleviated. Teachers might teach at an inappropriately high level because they feel that they have to complete the curriculum, they don't know their students' learning levels, or engaging with remedial teaching is costly. This program did not change any existing pressure on completing the curriculum but explicitly told teachers, with the support of their school principals, that setting aside particular hours each day and days per term to complete this alternative, remedial curriculum was important. In contrast to Sharnic Djaker (2022), teachers are aware that their students are below grade level but marginally over-estimate the extent of the deficit. One year after the program, teachers had lower and more correct beliefs about the learning levels of their students. Finally, this program provided materials and guides that lowered the effort expenditure to engage with remedial teaching. The second arm allowed more flexibility, but few teachers used it to revert back to the curriculum level material. Previous interventions that only provided materials or materials and training in primary schools did not increase learning (e.g., flipcharts, Banerjee et al. 2017, Duflo, Kiessel, and Lucas 2022). We show that setting aside specific time in the existing school day and days in the calendar could be a sufficient substitute to an additional, designated instructional hour as in Banerjee et al. (2017).

2 Background

Our study took place in 300 secondary schools in Odisha, a state in eastern India. In Odisha, like the rest of India, pre-tertiary education schooling is divided into four categories: primary

school (Class 1 to Class 5), middle school (Class 6 through 8), lower secondary (Classes 9 and 10), and higher secondary (Classes 11 and 12).[1] Students must pass standardized Board exams at the end of Class 10 and Class 12 to continue to additional education. The marks on the Board exam determine where and what field of study student can undertake. Our study follows students from the beginning of lower secondary school (Class 9) through their first year in higher secondary (Class 11). The school year starts in April with a month and a half summer vacation from early May to mid June and ends in March.

Odisha has a poverty rate of 32.6 percent, poorer than the national average, and a gross enrollment rate of 77.1 percent in secondary school, approximately the national average (NITI Aayog, 2021). Many secondary school students in Odisha are first-generation learners with about 40 percent of enrolled students being either scheduled caste or scheduled tribe students. About 50 percent of the enrolled students in Odisha fail to meet a basic benchmark of mathematical knowledge Das and Zajonc (2010).

As is typical in the Indian education system, schools in Odisha emphasize teacher-focused instruction and have many below grade level students and heterogeneous learning levels in the same classroom. The typical class period involves lecture-based pedagogy that strictly adheres to the official curriculum with limited pupil participation or deviation for students at learning levels below the official levels. At baseline, 95 percent of headmasters consider adhering to the curriculum to be an important component of their job (see Appendix Table 1). Our data confirms that student achievement in Odisha is low, and also reveals that it is highly variable. On average, students are about 4 grades levels behind in math and Odia, the local language.

[1] A "class" is the equivalent of a "grade" in the American school system; these classes are sometimes numbered with Roman Numerals, e.g. Class I or Class X.

3 The Utkarsh Program and Conceptual Framework

The *Utkarsh* program provides grade 3 through 8 content to Class 9 students. The program was a collaboration between the Odisha School Education Programme Authority (OSEPA), the authority within the Department of School and Mass Education responsible for education, and Transform Schools, People For Action (TSPFA), a large Indian NGO. It focused on Odia, English, math, and science and was designed to improve learning outcomes for students who are below grade level within the existing school day with existing teachers.

Our study covers two versions of Utkarsh – Standard Utkarsh and Flexible Utkarsh. We describe each below.

Standard Utkarsh

The program involved teacher training, teaching and learning materials, and a change in the allocation of time during the school day. Subject teachers and the school headmaster were invited to a one-week training session prior to the start of the 2019-2020 school year . At these sessions, participants learned how to use the teaching and learning materials to implement a more effective teaching practice. Appendix A contains additional details.

Once school started, the teachers were to test students to determine their learning levels, provide content to each learning level during prescribed days of the term, and then evaluate students learning again at the end of the academic year. Teachers assessed all of their Class 9 students with Utkarsh leveling exams, using the provided rubrices to categorize each student as inception (below Class 3), Class 3, Class 5, and Class 8 or above. The level of each student determined in which phase of the lessons the student would participate: Foundation Camp, Supported Learning Phase, or Consolidation Camp. All of the phases emphasized collaborative and student-centered active learning and included daily outlines of topics to cover and accompnaying worksheet pages in the student workbook. These were not scripted lessons. The program instructed teachers to provide alternative activities for students who were not engaged with a particular learning level.

Foundation Camp (FC): FC was for the students who initially tested at the Class 5 level

or below and was designed to support the learning of foundational concepts and skills. This phase was 4 hours per day for 18 days, for a total of 72 hours of instruction.

Supported Learning Phase (SLP): SLP targeted all students who tested at the start of the year below a Class 8 level. In this setting, this was about 90 percent of students. SLP further developed foundational concepts started in FC or from prior knowledge but at a higher level and with more advanced skills than FC. This phase was 3 hours of remedial teaching per day for 45 days, for a total of 135 hours of instruction.

Consolidation Camp (CC): The final phase, CC, included all students and focused on grade-level material in preparation for the Class 9 annual examinations. CC was 3 hours per day over 6 days, for a total of 18 hours of instruction.

At the end of the CC phase, approximately four months into the school year, teachers again assessed all students on their learning. These concluding assessments occurred over a period of two days and covered all four targeted subjects. In addition, towards the end of the school year, a review of student participation, concluding assessments and Class 9 exam results was conducted by the headmasters to assess student learning progress.

Flexible Utkarsh

Teachers and head masters received the same training and materials as the Standard Utkarsh version described above. In addition, teachers received instruction and materials on how to exercise their own discretion during SLP.

Foundation Camp (FC) as above.

Supported Learning Phase (SLP): During SLP, teachers could follow the standard Utkarsh plan as above or they could exercise their own discretion in planning the material and content during the 3 hours per day of SLP time that occurred over 45 days. They could spend more time on a particular SLP topic, skip SLP topics if their students did not need them, or use the time for standard curriculum instead of remedial SLP topics. They were encouarged to complete their Flexible Utkarsh SLP plan each week with the topics that they planned to cover for the week, whether copying from the standard plan or deviating. The training

instructed teachers to cover at least 50 percent of the material from the standard SLP curriculum, but could do so in any order and devoting less time to these lessons based on their local expertise about their students.

4 Empirical Strategy

The primary conceptual difficulty in assessing the effect of remedial learning on student outcomes is the typical correlation between remedial instruction and students' learning level or other school characteristics. To overcome this difficulty, we randomly assigned each of the schools in our 300 school sample into one of three groups: 1) Standard Utkarsh; 2) Flexible Utkarsh; and 3) control, i.e., business as usual.

We estimate the impact of the two variants of Utkarsh using the following equation:

$$y_{ist} = \alpha + \beta_1 StandardUtkarsh_s + \beta_2 FlexibleUtkarsh_s + \delta' X_{ist} + \epsilon_{isjt} \quad (1)$$

where y_{ist} is the outcome of interest for respondent i in school s at time t. $StandardUtkarsh_s$ and $FlexibleUtkarsh_s$ are dummy variables indicating the randomly assigned treatment status of the school. These indicators are mutually exclusive with the control group as the omitted category. X_{is} is vector of control characteristics, including the baseline value of the outcome variable (as appropriate), the wave of survey (if the outcome is measured at multiple waves), and strata, day of the week, and week of the year fixed effects. [2] Standard errors are clustered at the school level.

Our coefficients of interest are β_1, the effect of Standard Utkarsh relative to the conrol group, and β_2, the effect of Flexibile Utkarsh relative to the control group. The difference between $\beta_1$and β_2 is the difference in the effects of the two interventions.

Our primary outcomes of interest are student test scores. To understand the mechanisms behind test score changes, we also estimate the effect of the interventions on the program

[2]In all our specifications, if a control variable is missing, we dummy out that missing value by setting the missing values to zero and include as an additional control an indicator for the variable being missing.

implementation fiedelity and teching practices.

5 Sample Selection, Randomization, and Data

5.1 Sample Selection and Randomization

Our study occurred in Jajpur and Dhenkanal districts in Odisha State, India.[3] These districts had 711 secondary schools spread across 348 villages. We eliminated all schools that did not have positive reported enrollment in Class 9. To avoid contamination, we randomly selected only one secondary school from each village to be eligible for selection into the study. From these 348 schools, we randomly selected 300 to undergo additional screening to ensure the school used the the official state language (Odia), was governed by the SME and not the SC-ST Development Department, had enrolled Class 9 students, and was not a school for special needs students (i.e., deaf or blind). If a school from the list of first 300 schools did not clear the eligibility criteria, we replaced it with a randomly selected school from the list of the remaining 48 schools. We placed each of the 300 sample schools into one of 46 strata based on district, average pass rate on the prior year's Class 10 board exam, total Class 9 enrollment, teacher to student ratio, and distance to the district headquarters. Within each strata, we randomized equal number of schools into the three treatment conditions, resulting in 100 schools in each of the three treatment arms. The study design is pictured in Figure 1.

5.2 Data Collection

We conducted three waves of data collection during the school year (2019-2020) of implementation: a baseline survey, an unannounced monitoring visit during the school year, and a full follow-up at the conclusion of the intervention. The year after the intervention, we conducted an additional followup survey.

[3]The study was limited to two districts to more readily facilitate the training of teachers and head masters. These two districts were selected in consultation with the Odisha School and Mass Education Department.

Baseline

The baseline surveys took place in July and August of 2019, near the start of the school year but after the summer break and prior to the implementation of Utkarsh. We conducted demographic and background information from the school headmaster, teachers of the four Utkarsh subjects, and sample students; data about the school's infrastructure; and invigilated exams in Odia, math, and English. See Appendix B for additional test construction details. These exams were separate from the Utkarsh leveling exams, which teachers had not yet conducted at the time of our baseline exams.

Monitoring Visits

Between September and November 2019 we conducted one monitoring visit at each school. During these three months schools should have been implementing FC and SLP.[4] During each visit, enumerators arrived unannounced and recorded the attendance of the head master, teachers, and baseline students. Head masters and teachers responded to questions about program take-up and implementation. We also conducted classroom observations. [5]

Short Term Follow-up

We conducted the first follow-up from December 2019 to February 2020 at the end of the intervention.[6] Students responded to a short student survey that included a question about their board exam registration and registration number and completed subject exams in Odia, English, mathematics, and science. These tests were similar to those at baseline, but included additional more challenging questions and a science exam. We sought to interview and invigilate exams for all baseline students [7]

[4]We randomly assigned each school into receiving their visit during one of three monitoring visit phases: FC, early SLP, and late SLP. We block randomized assignment to monitoring visit phases by district and study arm. Monitoring visits in FC-phase schools took place when the FC phase of Utkarsh was going on. Similarly, in early-SLP-phase schools, monitoring visits took place immediately after SLP begun, while in late-SLP-phase schools we conducted monitoring visits late in the SLP phase of Utkarsh.

[5]Classroom observations occurred during the first period of the day. Enumerators sat in a classroom for one period and collected data on teacher behavior and presence, student behavior, and the use of teaching and learning materials. Monitoring visits occurred in 298 schools.

[6]Treatment schools should have still been still implementing the CC at the start of the fieldwork in December 2019. We randomly selected 9 strata to visit during December 2019, visiting all treatment and control schools in those strata. We visited 60 schools in December and the remaining 240 in January and February.

[7]Our analysis sample is all 5,448 students that completed both the surveys and assessments. Not all

During the endline survey visit, we also conducted surveys of teachers, and the school headmaster. The teacher survey asked teachers about their experience, autonomy, Utkarsh implementation, work load, and perception about Utkarsh. We also administered a competency test on English and math to teachers to test their knowledge on respective subject matters.[8] The headmaster survey covered a range of topics from information on their school and characteristics, to their personal background and school management practices. We confirmed the board exam registration information for each student in the sample with their head master.

To maximize response rate for this follow-up, we followed DiNardo, Matsudaira, McCrary and Sanbonmatsu (2021) and randomized the intensity of our mop-up visits at all schools to survey respondents (student, teacher, and headmaster) who were absent during the follow-up visit.[9] We conducted second mop-up visits in a random subset of schools where students remained absent during the first mop-up visit. We implement Lee bounds (Lee, 2009) to address potential biases of treatment effects due to non-random coverage of respondents in the follow-up.

Longer-Term Follow-Up Survey

We conducted a second follow-up via phone survey in December 2021, two years after the end of the program, to measure the impact of the program on longer-term enrollment rates and the transition to work.[10]

Administrative Board Marks

For all students in our sample we attempted to obtain the Class 10 Board Marks. Students in this sample should have taken their Board Exam in May 2021 at the end of the Class 10

students completed all exams.

[8]Creating a consistent and comparable sample of teachers across the study waves was challenging due to teacher vacancies and guest teachers (who may not work at the school each day of the week). Hence, we attempted to survey the same teachers over time, adding teachers as necessary and collecting demographics as they were added.

[9]Despite mop ups, in some cases we were unable to record student board exam registration for all students because the headmaster did not have time to access or could not access digital copies of the Class 10 board registration.

[10]This survey successfully reached 1,255 of the students from baseline (23 percent). This rate is similar across the treatment arms.

year. Due to the COVID-19 pandemic, the Board Exam for May 2021 was canceled. Instead, students received a Board Mark based on a weighted average of their in-class Class 9 scores (40 percent) and their in-class Class 10 scores (60 percent). The Class 9 and 10 scores were set by their teachers without knowing that they would be used as Board Marks. Because these Marks affect placement into fields of study and elite secondary schools, some students protested of this grading scheme. A formal Board Exam was eventually administered for students who wanted to try for a higher Board Mark. Approximately 5 percent of students took an actual Board Exam. The Board Marks in our data are the maximum of the original school-based exam marks and the Board Exam. Unfortunately, our data do not denote whether the student sat the formal Board Exam or whether the reported Board Marks are from the Exam or the from the school-based weighted average.

5.3 Summary Statistics and Baseline Balance

Appendix Table A1 shows that randomization successfully created three groups with balanced characteristics at baseline at the student and school-level; Appendix Table A2 shows that randomization successfully created three groups with balanced characteristics at baseline at the teacher level. Panel A of Table A1 describes student characteristics. Approximately half of the students in the sample are girls, and the average age is approximately 13 years. Nearly two-thirds (61 percent) of students belong to SC/ST/OBC, the disadvantged minority groups in India. About 15 percent of students in our sample are first generation learners (i.e., no parent can read and write). Eighty eight percent of our students participated in Utthan, another state-wide intervention focusing on Class 8 students. About 72 percent of the student sample reported currently paying for a private tutor ("private tuitions") in preparation for the Class X Board exams. Students in each study arm are similar to each other in terms of baseline test scores as well. We find a small difference in baseline English test that can be attributed to statistical noise since we are comparing several variables across the three groups to validate pre-intervention balance.

We find that on average our students have very low level of competency at baseline; see Appendix Figure A1. In all of the three baseline subjects, the mean student is over 4 grades behind. [11] Nearly half of all students are evaluated at the below Class 3 level. However, there is also substantial heterogeneity in the sample. Nearly 8 percent of students in Math and English, and nearly 18 percent of students in Odia are at grade level. This heterogeneity is not only across schools but within schools. For example, in math, the average interquartile ranking of competency (i.e., the mean difference within a school between the 75th and 25th percentiles) is 3.94 grades. Therefore, on average, although nearly half of Class 9 students are at a below 3 competency in a given classroom, teaching remedial instruction is not a clear choice as within each class there are generally some students who are at grade level. Teachers are also relatively accurate in their estimates of student proficieincy; for example, in math, teachers are only off by 2-6 percentage points as compared to the estimates of students at each level of competency (not shown).[12]

Panel B of Table A2 presents headmaster- and school-level characteristics from the survey of headmasters (or another informed deputy if no headmaster was available). About 93 percent of the baseline surveys were conducted by headmasters or headmaster in charge. Approximately 22 percent of these individuals were women. Our headmaster survey also indicates the dearth of autonomy in schools to adjust curriculum to meet the learning-level need of students. About 77 percent of headmasters in our baseline survey also share the view of teachers that official curriculum should be followed under such circumstances. In fact, 97 percent of headmasters consider ensuring adherence to the curriculum as an important part of their job and 80 percent of them think that they have influence over determining how the teachers deliver the curriculum lessons to students at school.On average, sample

[11] If a student has mastery of a specific grade, that the student is considered to have competency of that grade. If a student has grade 3 level mastery, then that students grade competency is 3. If a student has a mastery at a higher grade level, it is assumed that the student also has competency as all lower grades. For instance, if a student has grade 8 level mastery, then that students grade competency is 8 and it is assumed that the student has grade 3 and grade 5 level mastery. Students below grade 3 competency are assumed to have competency at grade 2.

[12] These results are similar to teacher-collected data from the program leveling test, which also find a high share of students have low-level of learning (see Appendix Table A3).

schools have 7.5 sanctioned positions for Class 9 teachers, but only 5 posts were filled, in spite of a relatively large student enrollment in 69 students in Class 9.Total enrollment is statistically different across the three arms. Class 9 enrollment in Standard Utkarsh schools is smaller than in the other two arms. However, some degree of imbalance is expected and is not expected to pose a challenge to our evaluation.

Appendix Table A2 describes teacher characteristics. Slightly less than half of the teachers are female (48 percent), and the average teacher is 42 years old. Approximately one-third (35 percent) of teachers have a teaching certificate, and the average teacher has 16.5 total years of experience. Almost all of our baseline teachers primarily teach an Utkarsh subject. Baseline data indicate that despite concerns of shirking, there is very little self-reported absence from work in the previous 6 days of the survey. Teachers report substantial amount of time spend about 21 hours each week in preparing lessons and grading. Teachers believe that on average approximately 60 percent of their students will pass board exams on their first try.

We measured teachers' reported autonomy over teaching and learning activities. About 95 percent of our teachers agree that they can select teaching materials, methods, and strategies. About 80 percent of teachers think that they are allowed to modify course timetable if students need more time to understand a topic. Only 34 percent of teachers say that others select evaluation and assessment activities and 28 percent of teachers feel pressure to complete curriculum during school year. There are no systematic differences in teacher burnout or autonomy across study arms. A majority of teachers (58 percent) believe that the official curriculum should be followed even if students learning level is below the grade level. Teachers also do not feel that it is their fault if students are lagging behind. Only 47 percent of teachers feel that it is their fault if students are not ready for the Class 10 board exam.

6 Results

We check for systematic attrition at endline across treatment arms and find that standard Utkarsh students are two percentage points less likely to not take an endline test. We constructed treatment bounds following (Lee, 2009) to mitigate concerns of differential attrition creating biases in our preliminary results. We find that in all cases the results from this exercise are similar to the primary results.

6.1 Student Outcomes: Achievement, Attrition, and Aspirations

The Utkarsh program improves student learning across all subjects. Table 1 reports treatment effects on test scores in standard deviations (SD), for overall scores as well as separately for each subject. Both versions of the Utkarsh program increased students' overall test scores by about 0.11 SD (column 1). Over this same period, overall test scores for the control group increased by 0.19 SD; thus, Utkarsh improved learning by 59% relative to the status quo. As measured in control-group SDs, Utkarsh's effects on subject-specific scores are similar to the overall effects: Utkarsh increased English and Math scores by 0.12 SD (columns 2 and 3)) and Odia scores by 0.09 SD (column 4), with these effects being about the same for both versions of the program. The magnitudes of the effects on science scores differ somewhat by program version: Standard Utkarsh increased science scores by 0.10 SD and the Flexible version increased scores by 0.14 SD, although we cannot reject the null hypothesis that the two effects are equal (column 5). Relative to the control group's rate of test score gains over the same period, the program increased learning in English by 57%, Math by 190%, and Odia by 43%.[13] In all subjects, the two treatments had statistically indistinguishable learning impacts. We find little evidence of heterogeneity by gender, caste/tribe, or first generation learner status (Tables A3-A5). These results are robust to multiple hypothesis testing corrections (not shown).

These impacts are not due to differential attrition by treatment status. Appendix Table

[13]We cannot show learning gains for science because there was no baseline science test.

A6 estimates differences across study arms in attrition, defined as a students not taking all four of the endline tests. We find that students in the Standard Utkarsh arm are two percentage points more likely to have an endline test scores. Because of this differential attrition, we constructed treatment bounds following Lee (2009). As shown in Appendix Table A7 in all cases both the magnitude and statistical significance are similar to the main effects. This is unsurprising as the attrition rate is very low, at just 6 percent in the control group.

One concern is that students who are at grade level may not benefit or may even be harmed by remedial education programs. We find no evidence of this concern. In fact we find compelling evidence that students throughout the test score distribution benefit from the program. In Figure 2 we plot non-parametric test score effects (i.e.,kernel-weighted locally smoothed regressions by baseline test score percentile) and find consistent learning gains throughout the entire distribution of baseline scores (see Appendix Figure A2). The results are similar when dividing students by tercile (Table A8); for overall test scores, students in the top, middle and bottom tercile of the baseline test score distribution have similar treatment effects. If anything, the effects are larger for the botton tercile, although we cannot reject the null hypothesis that the effects on all three terciles are equal. The distributional effects of the program vary by subject: for English, the treatment effects are significantly larger at the top of the distribution, while the opposite is true for science. One caveat to the science results is that we have no baseline science score, so for that exam we present heterogeneity in treatment effects by the overall baseline score instead. Overall, these results suggest that the program was beneficial for all students regardless of their baseline proficiency levels. Consistent with this theory, we find no treatment effects on measures of student aspirations including self-assessed rank, expected class 10 board marks, or aspirations for higher education (see Appendix Table A9).

Although Utkarsh improves students' learning overall, the program's significant effects on learning do not translate into improvements in mastery of grade-level material. Figure

3 reports treatment effects on mastery of grade 3, grade 5, and grade 8 material. Despite substantial improvements in grade 3 and grade 5 mastery in nearly all subjects, neither intervention improved grade-level (grade 8) mastery in any subject.One explanation for this is that students were significantly behind grade level at baseline; the Utkarsh program's remedial curriculum bridges some of the gap, but is not enough to bring them fully up to grade-level material.[14]

6.2 Implementation of Utkarsh

The Flexible version of Utkarsh led to very similar gains in learning to the Standard version. Given these similar results, one question is whether the program was implemented differently in the two treatment arms. Table 2 reports our results on program implementation, focusing on the program features that were intended to tbe the same in the two treatment arms. Despite concerns that allowing flexibility would lower program quality, fidelity to the program's intended design is very high in both versions of Utkarsh. Nearly all teachers reported conducting the leveling assessment and teachers in both treatment arms. The vast majority of teachers report teaching at least one Utkarsh lesson in the last 6 days, and almost all teachers reported doing an Utkarsh worksheet the last time they taught. Teacherswere conducting the correct phase of the program when we visited their classrooms. We also note that the correct groups of students were generally included in each stage of the program, and equally so in both versions of the program. There may be concerns that teacher-reported program fidelity measures are biased, and so inn Table 3, we consider outcomes that were observed by enumerators. Results consistently show that the program was implemented with a high degree of fidelity in both arms. First, nearly all students (90%) were using an

[14]We see some evidence that the program improves scores on PISA questions (Appendix Table A10, corroborating that the results are not merely due to teaching-to-the-test. Overall PISA scores (combining English and Math questions) improved by 0.07 SD for the Flexible arm and 0.05 SD for the Standard arm, with the former effect being significant at the 5 percent level. Looking at the subject-specific PISA scores, we see larger effects of Standard Utkarsh for English PISA questions, and larger effects of Flexible Utkarsh for Math PISA, although we can only reject the equality of the two effects for the latter, and only at the 10 percent level.

Utkarsh handbook during the unannounced spot check.As part of the program, schools were supposed to have a "Word Wall" to help students have additional resources for learning. Only 4% of control schools have a word wall, and we find that the likelihood of a word wall is 10 percentage points higher in the standard arm, and 14 percentage points higher in the Flexible Arm (although we cannot reject the equality of the two treatment effects). Students' desks are also significantly more likely to be arranged in small groups as directed by the program. Thus these additonal measures of classroom implementation indicate that the program was correctly adopted and implemented.

In both Utkarsh variants, what teachers taught largely matches the assigned curriculum during the FC stage, when both the Standard and Flexible versions of the program asked teachers to follow a set schedule of lessons. During spot check visits at the FC stage, roughly 90% of teachers report teaching the assigned lesson. We see just two statistically significant differences in program implementation out of the eleven comparisons in the table: Flexible Utkarsh teachers were three percentage points less likely to have done an Utkarsh worksheet on the most recent day they taught, and spent eight percentage points more of their classes doing Utkarsh lessons during the FC phase. Both of these differences are small relative to the overall mean, the latter difference is not evident during the SLP phase.. Thus implementation quality is high and largely identical in both versions of the program.

6.3 Flexibility Takeup

Did teachers actually use the additional autonomy that they were given in the Flexible version of Utkarsh? Our results, shown in Table 3, suggest that they generally did not. Self-reported measures do suggest that some teachers did follow the intended structure of the program, although most did not. Just 24 percent of teachers in the Flexible Utkarsh arm had the teaching plan that they were supposed to use to decide which lessons to teach during the SLP stage of the program, and even lower rates filled it out or followed it. While we can easily reject the equality of these rates with the Standard arm (and with the control

group), they indicate that the vast majority of teachers did not even attempt to deviate from the Utkarsh class schedule.

We see moderate effects on teacher reported autonomy. Flexible Utkarsh boosts the rate of teachers reporting some autonomy in using the lesson in class by 5 percentage points, but this is relative to a Standard Utakrsh mean of over 94 percent. There are also significant effects on teachers having control over the course timetable if students have trouble with a topic, which was one of the core goals of the Flexible version of the program. This increases from 77% in the Control group and 78% in the Standard arm to over 84% in the Flexible arm. Similarly, the rates of following the scheduled lesson during the SLP stage of the program (when flexibility was encouraged) are lower in the the Flexible arm than in the Standard arm, but still around 60 percent. Thus, a large share of teachers seem to not take up the flexibility they were given at all, and instead simply revert to the standard option.

6.4 Classroom Activities

To understand how Utkarsh led to changes in test scores, in Table 4 we present data on classroom activities practices. Utkarsh caused classrooms to become more active and engaging. We find that implementing the program improved teaching practices. Classroom observations allow to collect data on the following measures of teaching practices: at least one student had an opportunity to express their own idea, teacher asked a question to the class, teacher answered students' questions supportively, teacher answered students' questions without showing disrespect, teacher did not ignore students' questions, teacher seemed familier with content, teacher encouranges student, teacher respond to questions satisfactorily, and teacher teaching with student interaction. We observe significant treatment effects on several meeasures of teaching practices. Students have more opprotunity to express ideas in class—the likelihood of at least one student expressing an idea in class is 18 and 19 percentage points higher in Standard and Flexible classrooms, respectively. Teachers are significantly more likely to encourage students, respond to questions satisfactorily, and

teach with greater student interaction. There are no significant effects on the presence of teaching and learning materials in class, likelihood of teachers asking a question, ignoring student questions or answering questions supportively and without disrespect, or the observed teacher familiarity with the content. An index of these teaching practices improves by 0.39 and 0.35 SD in the Standard and Flexible treatment groups respectively.

In Appendix Table A11, we examine treatment effects on additional teacher-level outcomes. Treatment-group teachers are marginally more likely to be "burned out", stressed, and anxious (Columns 1-3), although these differences are not statistically significant.[15] The intervention did not change the total amount of time teachers spent doing lesson preparations or grading (Columns 4 and 5). However, 72-75 percent of teachers report that students benefitted from Utkarsh(Column 6), and 94-95 percent teachers themselves enjoyed teaching the lessons (Column 7); these treatment effects do not differ by treatment arm. Simlarly, about 90 percent of teachers reporeed that they also benefitted from Utkarsh (Column 8), with teachers in Flexible Utkarsh schools reporting statistically higher rates of satisfaction. We find some suggestive evidence that Flexible Utkarsh improved teacher competency in Math by approximately 3 percentage points; see Appendix Table A12. The effect of Standard Utkarsh was statistically insignificant, but we cannot reject the null hypothesis that the two effects are equal (Column 3).

6.5 Board Test Score Results and Teacher Perceptions of Students

Students in our sample were scheduled to take a high-stakes board exam one year after they the Utkarsh program ended, in June 2021. The board exam determines their admission to and track in upper secondary school (Classes 11-12) and is thus a crucial factor both for further educational success as well as eventual life outcomes. The Utkarsh intervention could

[15]Burnout index is an inverted covariance matrix weighted standardized index generated following Anderson (2008) from the following variables: teacher feeling mentally exhausted from work; feeling fatigued; feeling having a positive influence on people; feeling very energetic about job; feeling satisfied with job at school. Stress and anxiety are measured on the Depression Anxiety Stress Scale (DASS). To construct respective indices, response to relevant questions of the scale are summed and then standardized.

affect board exam outcomes in a number of ways. First, by improving learning outcomes subject competencies the program could lead to a better performance in the board exams. Alternately, the Utkarsh program may have subsituted teachers' and students' time away from the course syllabus, leading the students to perform worse on the board exams, which are solely focused on the content included in the official government syllabus, and thus may reward rote learning.

However, board exams scheduled for June 2021 were canceled due to the COVID-19 pandemic. Instead of the typical in-person exam, students were assigned "Board marks" based upon a weighted average of their school-based exams from Class 9 (40%) and Class 10 (60%), which were scored by the students' own teachers. Some students protested this approach, and thus an "offline exam" was offered; for the subset of students who protested, board marks were the higher of this offline exam or their teacher-assigned Board marks. Approximately 5 percent of students took the offline exam, and this rate does not differ by treatment arm.[16] Since the board marks were largely determined by the students' own teachers rather than an objective test, changes in teacher perceptions of students in treatment schools could also affect their eventual baord exam scores.

Given the potential role of teacher perceptions in driving the board exam results, Table 5, examines treatment effects on those perceptions. Despite improvements in student learning, teacher perceptions of student ability substantially fall as a result of the program. Teachers believe that students in their school are 4-5 percentage points less likely to pass the Board exam on the first try (Column 1). Similarly, students are less likely to be assessed as able to write a simple sentence in English (Column 2) or do a subtraction problem (Column 3); treatment teachers also report students as less likely to complete a Bachelor's degree (Column 4). None of these treatment effects differ significantly between the two treatment arms. One reason for these effects could be increases in accuracy: the Utkarsh program involves repeated assessments of students; since students are far below grade level, teachers might learn from those assessments that their students are weaker than they thought. We compare teachers'

[16]Our data do not allow us to distinguish between the offline exam scores and Board marks.

perceptions with the overall average performance of students from their school averages based upon group exams. We find that control-group teachers greatly overestimate their students' skills, and that the program made teachers more accurate (and thus more pessimistic) about their students' ability (Columns 5 and 8).

Next we pair our evaluation data with administrative data on the Class X Board Exam for the universe of all 18,551 students from the study schools to examine treatment effects on pass rates and exam scores. Column 1 of Table 6 shows that presence in the Board Exam score data is uncorrelated with treatment status. Column 1 of Table 6 shows that nearly all students (99%) passed the Board exam, and this average does not differ by treatment status. However, the Standard Utkarsh program leads to uniformly lower scores on all subjects; in total, across all Utkarsh subjects, they score approximately 2 percentage points lower than the Control group (from a base of 55.4%). Flexible Utkarsh students score approximately 1.2 percentage points lower, but we cannot distinguish this effect from the null of a zero difference. Students in the Standard Utkarsh program also score 2.9 percentage points lower even on Non-Utkarsh subjects (typically: a third language, such as Hindi). Similarly, Standard Utkarsh students are less likely to clear the bar to receive each letter grade from A-D (and thus are more likely to receive a grade below D). These results are consistent with teachers' diminished perceptions of students as a result of the program. The lower scores in non-Utkarsh subjecs suggest spillovers from perceptions of the directly targeted subjects (where teachers received information that their students where struggling) onto perceptions of student ability more generally . These lower scores did not lead to changes in actual pass rates on the exam.

We supplement the Board Exam scores with a phone survey of 1,255 study students conducted in November-December 2021. While the board exam scores cover the universe of students in the study schools, the coverage rate in the phone survey was just 23 percent In Appendix Table A14 we show that completion of the survey is not correlated with treatment status. Consistent with the lack of treatment effects on passing the Board Exam, there are

also no significant differences in whether the student is currently enrolled in school, currently enrolled in Class 11, or currently employed. Thus, although the treatment caused students to receive lower Board marks, we observe no changes in other long-term outcomes of interest.

7 Cost Effectiveness

As implemented at a 200-school scale, the intervention cost $11.64 per student. As the training and monitoring costs were identical in both arms, both arms were equally cost-effective. The observed cost per student translates into a 0.95 SD overall test score gain per $100 spent. This estimate is comparable to an after-school personalized tutoring intervention known as Mindspark for students primarily in grade 7 and 8 (0.93 SD per $100 at 50-school scale) (Muraldiharan et al., 2019). It is slightly less than the 1.4 SD per $100 at 200-school scale eLearn program which introduced school screens and videos in Pakistan (Beg et al., 2022).

8 Discussion

We find that at baseline, the mean Class 9 student in our sample is over 4 grade levels behind in math, English, and Odia. Moreover, there is substantial heterogeneity both within classrooms and across schools: the typical classroom also has a range of student competencies of 3.94 grades, and approximately 10 percent of students are at grade level. With these substantial learning gaps, as well as substantial heterogeneity, it is feared that introducing remedial education programs may crowd out grade-level skills and stall progress for students who are at grade level. Moreover, it is feared that lowering standards and allowing teachers more flexibility in their classroom may increase provide a disincentive for teachers to work hard, and may increase shirking. Our randomized evaluation empirically evaluates these concerns and finds that they are unfounded. We find that introducing a remedial education known as Utkarsh substantially improved student learning, by 58%, and did not crowd-out

grade level competencies (although they also did not change). .

Utkarsh also resulted in more effective classrooms; teachers report that both they and their students benefitted. Generally the two different implementation models have similarly high rates of fidelity to the program guidelines; we find that any differences observed are generally in favor of flexibility. The notable exceptions are that teachers in the Flexible arm are more likely to report being burned out but also significantly improve their competency at math. Part of this response is likely due to relatively low rates of take-up of the flexibility. Although there is anecdotal evidence that teachers want additional autonomy, the revealed preference of teachers in our sample is for the pedagogy as currently delivered.

In total, there are several lessons from this study. The first is that teachers were able to adapt and implement a remedial education program that was different from the curriculum. While offering teachers autonomy did not differentially improve learning, it also did no harm. These results bolster the interpretation that rigid bureaucracies could improve service delivery by modifying standards to adapt to local conditions. Despite concerns of coordination challenges of changing the status quo, our results suggest that teachers were able to effectively adapt to a new, more effective approach in the classroom that benefitted both students and teachers. Our results also suggest that assumptions that teachers demand autonomy is not always true. In our sample, our teachers have low demand for flexibility, and do not use flexibility when it is offered.

The second lesson from this study is whether or not introducing remedial instruction crowds out learning at the current curriculum at the secondary school level. There is scant evidence at what interventions are effective at improving learning at the secondary school level. Our results show that although nearly half of all students are substantially behind the curriculum, it is cost-effective to improve learning at the secondary school level and decrease the substantial heterogeneity observed within a classroom. The Utkarsh intervention also did not crowd out (nor change) grade-level mastery, a key concern. Teachers also become more accurate in their evaluation of student competencies. One caveat, however, is that teachers

may view a trade-off between the curriculum and remedial education, even when one does not exist. While we find that pass rates on the Board exam were unaffected, partially due to near universal pass rates, results are consistent with the interpretation that teachers rated students as less proficient at grade-level competencies. These results have important implications for ensuring that teachers are made aware of how far behind students are, and suggests that more emphasis should be made on disseminating information on how much students are learning so that teachers are not demotivated by remedial programs.

9 Conclusions

In spite of rigid bureaucratic rules to ensure uniform delivery of service and reinforce high standard of performance, public sectors in developing countries are often crippled with poor service deliver and low productivity. We partner with PFA—a non governmental organization—and the state government of Odisha to implement an RCT in 300 secondary schools to evaluate a remedial instructional program (Utkarsh) and answer whether allowing public servants (teachers, in this case) more flexibility to tailor their service (teaching remedial lessons) according to the need of their clients (students) improves the productivity (student learning) of the public service.

When students fall behind in school they are often unable to catch up. In Odisha, Class 9—the penultimate year of lower secondary school—is a critical year for students. PFA designed the Utkarsh program to enable students who were behind grade level in Class 9 to catch up to curriculum-level material on English, math, Odia, and science. The goal is to prepare them for entering Class 10, when students take board exams that determine whether they can continue on to upper secondary school.

We find that offering teachers autonomy did not differentially improve or harm school productivity—the program improved student learning the same irrespective of teacher having more flexibility or not. The positive effect of Utkarsh has important policy consequences as the state government is committed to scale-up Utkarsh to all 30 districts of Odisha, reaching

about 350,000 Class 9 students.

We also do not find any systematic difference in the fidelity of program implementation, assuaging any concern that providing flexibility to teachers might encourage them to shirk. However, we find that teachers generally did not use the flexibility offered, appearing to prefer teaching the lessons as-is.

However, one limitation of our finding is that the learning improvement is still inadequate to cover students' learning gap. After the program, students still suffer grade-level learning gap. This suggests that additional research is required to design interventions that would improve secondary students' learning in similar setting by larger amount.

References

Afridi, Farzana, Vegard Iversen, and M. R. Sharan, "Women Political Leaders, Corruption, and Learning: Evidence from a Large Public Program in India," *Economic Development and Cultural Change*, 2017, *66* (1), 1–30. Publisher: The University of Chicago Press.

Anderson, Michael L., "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American Statistical Association*, December 2008, *103* (484), 1481–1495.

Bandiera, Oriana, Michael Carlos Best, Adnan Qadir Khan, and Andrea Prat, "The Allocation of Authority in Organizations: A Field Experiment with Bureaucrats," NBER Working Paper 26733, National Bureau of Economic Research, Cambridge, MA 2020.

Banerjee, Abhijit, Chattopadhyay, Raghabendra, Duflo, Esther, Keniston, Daniel, and Singh, Nina, "Improving Police Performance in Rajasthan, India: Experimental Evidence on Incentives, Managerial Autonomy, and Training," *American Economic Journal: Economic Policy*, 2021, *13* (1), 31–66.

_ , Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland, and Michael Walton, "From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application," *Journal of Economic Perspectives*, 2017, *31* (4), 73–102.

Banerjee, Abhijit V, Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani, "Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India," *American Economic Journal: Economic Policy*, 2010, *2* (1), 1–30.

_ , Shawn Cole, Esther Duflo, and Leigh Linden, "Remedying Education: Evidence from Two Randomized Experiments in India," *Quarterly Journal of Economics*, 2007.

Barrera-Osorio, Felipe and Dhushyanth Raju, "Teacher Performance Pay: Experimental Evidence from Pakistan," *Journal of Public Economics*, 2017, *148*, 75–91.

Beg, Sabrin, Adrienne Lucas, Waqas Halim, and Umar Saif, "Engaging Teachers with Technology Increased Achievement, Bypassing Teachers Did Not," Technical Report w25704, National Bureau of Economic Research, Cambridge, MA 2019.

Beteille, Tara, Elaine Ding, Ezequiel Molina, Adelle Pushparatnam, and Tracy Wilichowski, *Three Principles to Support Teacher Effectiveness During COVID-19*, World Bank, Washington, DC, 2020.

Bjorkman, Martina and Jakob Svensson, "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda," *The Quarterly Journal of Economics*, May 2009, *124* (2), 735–769.

Bloom, Nicholas, Renata Lemos, Raffaella Sadun, and John Van Reenen, "Does Management Matter in Schools?," *The Economic Journal*, 2015, *125* (584), 647–674. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecoj.12267.

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson, and Waly Wane, "Enrollment without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa," *Journal of Economic Perspectives*, 2017, *31* (4), 185–204.

_ , _ , _ , _ , **Christophe Rockmore, Brian Stacy, Jakob Svensson, and Waly Wane**, "What Do Teachers Know and Do? Does It Matter?," Working Paper 7956, World Bank, Washington, DC 2017.

_ , **Ezequiel Molina, Deon Filmer, and Jakob Svensson**, "The Lost Human Capital: Teacher Knowledge and Student Learning in Africa," 2018. Publisher: CEPR Discussion Paper No. DP12956.

Brown, Christina and Tahir Andrabi, "Inducing Positive Sorting Through Performance Pay: Experimental Evidence from Pakistani Schools," *University of California at Berkeley Working Paper*, 2020.

Chaudhury, Nazmul, Jeffrey Hammer, Michael Kremer, Karthik Muralidharan, and F. Halsey Rogers, "Missing in Action: Teacher and Health Worker Absence in Developing Countries," *Journal of Economic Perspectives*, 2006, *20* (1), 91–116.

Das, Jishnu and Tristan Zajonc, "India Shining and Bharat Drowning: Comparing Two Indian States to the Worldwide Distribution in Mathematics Achievement," *Journal of Development Economics*, 2010, *92* (2), 175–187.

DiNardo, John, Jordan Matsudaira, Justin McCrary, and Lisa Sanbonmatsu, "A Practical Proactive Proposal for Dealing with Attrition: Alternative Approaches and an Empirical Example," *Journal of Labor Economics*, 2021, *39* (S2), S507–S541. Publisher: The University of Chicago Press.

Duflo, Annie, Jessica Kiessel, and Adrienne Lucas, "Experimental Evidence on Alternative Policies to Increase Learning at Scale," Technical Report w27298, National Bureau of Economic Research, Cambridge, MA 2020.

Duflo, Esther, Pascaline Dupas, and Michael Kremer, "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, 2011, *101* (5), 1739–1774.

_ , _ , **and** _ , "School Governance, Teacher Incentives, and Pupil–Teacher Ratios: Experimental Evidence from Kenyan Primary Schools," *Journal of Public Economics*, 2015, *123*, 92–110.

_ , **Rema Hanna, and Stephen P. Ryan**, "Incentives Work: Getting Teachers to Come to School," *The American Economic Review*, 2012, *102* (4), 1241–1278. Publisher: American Economic Association.

Ganimian, Shwetlena Sabarwal Sharnic Djaker Alejandro, "Primary- and middle-school teachers in South Asia overestimate the performance of their students," *New York University Working Paper*, 2022.

Glewwe, Paul, Nauman Ilias, and Michael Kremer, "Teacher Incentives," *American Economic Journal: Applied Economics*, 2010, *2* (3), 205–227.

Lakshminarayana, Rashmi, Alex Eble, Preetha Bhakta, Chris Frost, Peter Boone, Diana Elbourne, and Vera Mann, "The Support to Rural India's Public Education System (STRIPES) Trial: A Cluster Randomised Controlled Trial of Supplementary Teaching, Learning Material and Material Support," *PLoS ONE*, 2013, *8* (7), e65775.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 2009, *76* (3), 1071–1102.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian, "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India," *American Economic Review*, April 2019, *109* (4), 1426–1460.

_ **and** _ , "Improving Public Sector Management at Scale? Experimental Evidence on School Governance India," Technical Report w28129, National Bureau of Economic Research, Cambridge, MA 2020.

_ **and Venkatesh Sundararaman**, "Teacher Performance Pay: Experimental Evidence from India," *Journal of Political Economy*, 2011, *119* (1), 39–77.

NITI Aayog, "SDG India Index & Dashboard 2020-21," Technical Report, National Institute of Transforming India, New Dehli 2021.

Pearson, L Carolyn and William Moomaw, "The Relationship Between Teacher Autonomy and Stress, Work Satisfaction, Empowerment, and Professionalism.," *Educational Research Quarterly*, 2005, *29* (1), 38–54. Publisher: ERIC.

Piper, Benjamin, Yasmin Sitabkhan, Jessica Mejia, and Kellie Betts, "Effectiveness of Teachers' Guides in the Global South: Scripting, Learning Outcomes, and Classroom Utilization," Technical Report, RTI Press 2018.

Rasul, Imran and Daniel Rogger, "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service," *The Economic Journal*, 2018, *128* (608), 413–446. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecoj.12418.

_ **,** _ **, and Martin J. Williams**, "Management and Bureaucratic Effectiveness: Evidence from the Ghanaian Civil Service," Technical Report 8595 2018.

Ree, Joppe De, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers, "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia," *The Quarterly Journal of Economics*, 2018, *133* (2), 993–1039. Publisher: Oxford University Press.

Skaalvik, Einar M. and Sidsel Skaalvik, "Teacher Self-Efficacy and Perceived Autonomy: Relations with Teacher Engagement, Job Satisfaction, and Emotional Exhaustion," *Psychological Reports*, February 2014, *114* (1), 68–77.

A Additional Implementation Details

Teachers and school The program begins by holding a one-week training session for all schools in the program immediately before the beginning of the school year. School headmasters and a one teacher for each of the four targeted subjects are invited, and the training centers around how to use handbooks that explain how to implement the program Utkarsh subject-specific handbooks into an effective teaching practice. All program schools receive teaching and learning materials developed by PFA, which include the teacher handbooks as well as student handbooks and workbooks; the workbooks have worksheets for the students to complete for each day of the program. PFA helps to run the training sessions and collaborates with the government to monitor implementation and maintain quality. In Odisha, the partner government department is the Department of School and Mass Education (SME). For the version of the program we study in this paper, PFA conducted the training sessions themselves. The program was also scaled up in the rest of Odisha starting in the 2019-20 school year; for this broader scale-up, PFA used a cascade (train-the-trainers) model to run the trainings, teaching SME staff how to do the actual training of teachers .

B Additional Test Construction Details

We used bespoke exams to test student learning. We constructed the tests specifically for evaluation and did not share them with PFA, SME, or any other entity involved in the implementation of Utkarsh during the period of evaluation. Our test questions were based upon learning objectives from the official curriculum and covered material from Class 3 through Class 9. For English and math tests, we included questions from PISA that may not necessarily map to the official curriculum.

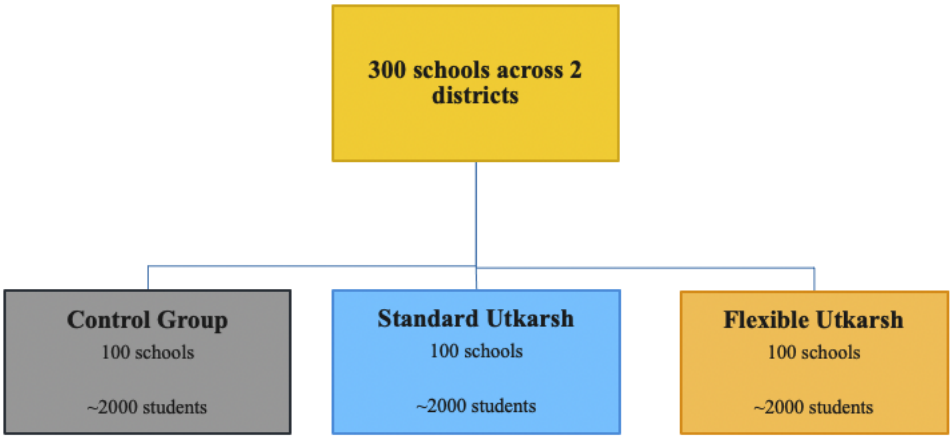
We used item response theory (IRT) to construct students' test scores, pooling all questions from baseline and follow-up. The IRT scores for math, English, and Odia are standardized by the baseline mean and standard deviation. Since we administered the science test

only at the follow-up, we use the control group mean and standard deviation to standardized the science test score. We also construct an overall score using IRT by pooling all baseline and follow-up questions of all subjects. We then standardize with respect to baseline mean and standard deviation.

We calculated the students' grade-level mastery, competency level, and number of grades behind in English, math, and Odia based on their ability to respond to specific grade-level questions correctly.[17]

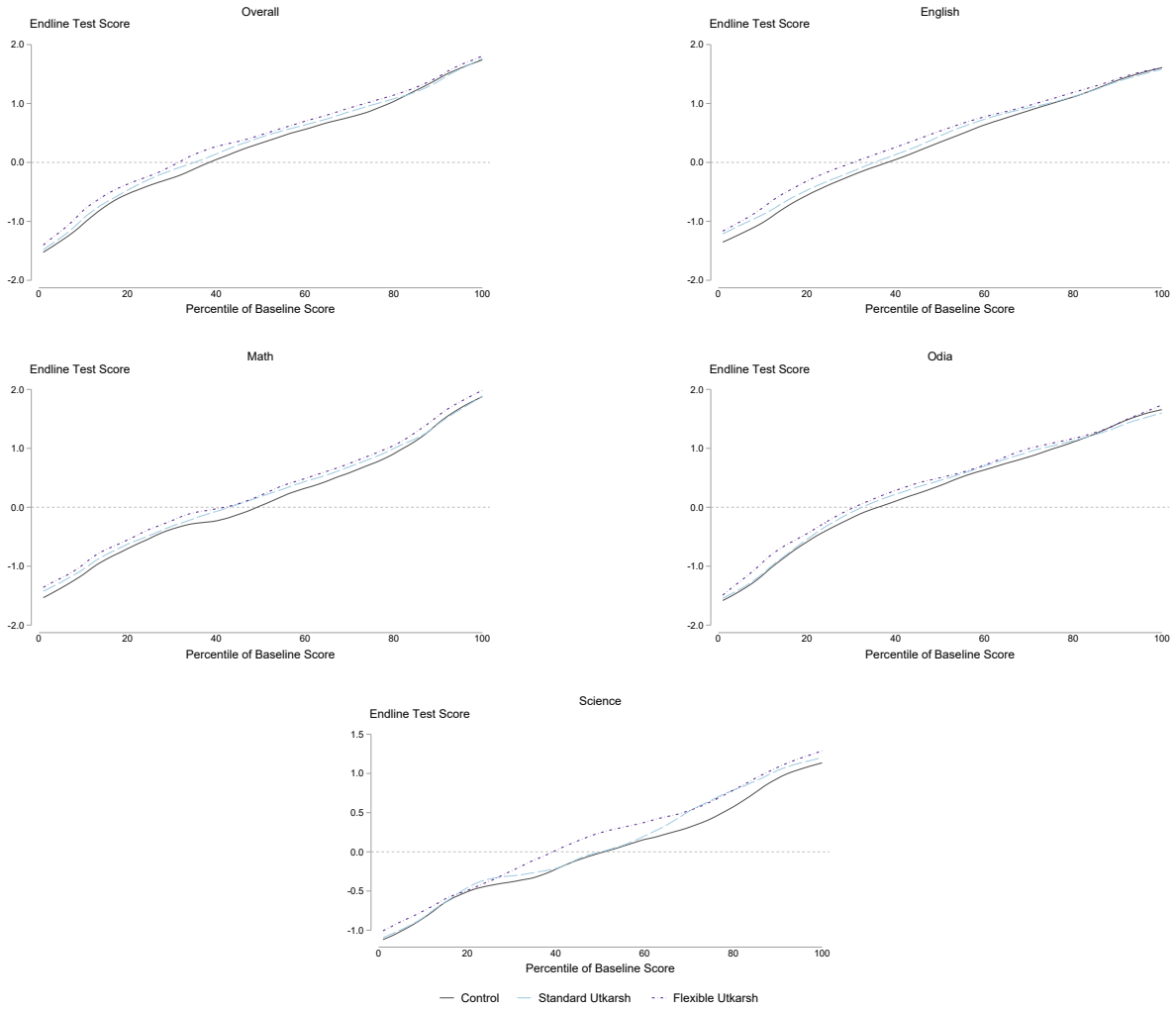
Figures

Figure 1: Study Design



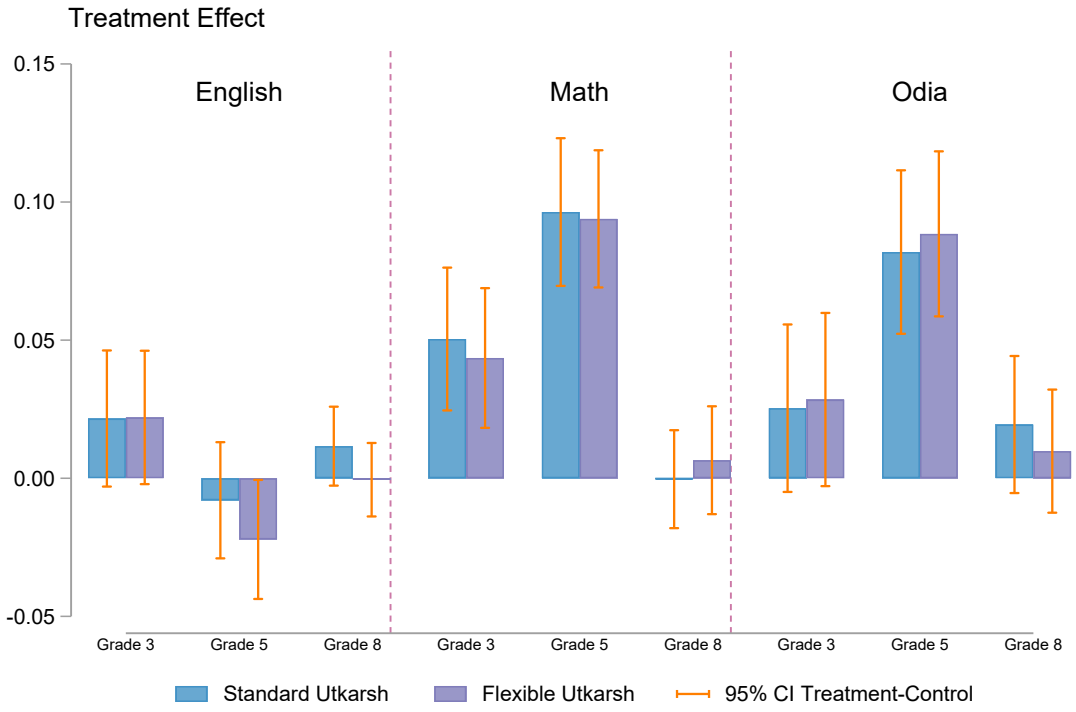
[17]Students have Class 8 mastery if they correctly answered at least 75 percent of Class 8 level questions. Students had Class 5 level mastery if they answered at least 75 percent of Class 5 level questions correctly but less than 75 percent of Class 8 level questions. Similarly, a student had grade 3 level mastery if they correctly answered at least 75 percent of Class 3 level questions but less than 75 percent of Class 5 questions. Students who answered fewer than 75 percent of class 3 questions correctly are considered to have class 2 mastery. We construct grades behind by subtracting grade level competency from 8. For instance, a student with competency at grade level 8 is 0 grades behind, while a student with below grade 3 level competency is 6 grades behind.

Figure 2: Non-parametric Distribution of Test Scores by Study Arms



Note: The figure is showing kernel-weighted local mean smoothed distributions of endline test scores over baseline test score percentiles by study arms. Since there was no science test at baseline, we use the average of all subjects at baseline as proxy for science's baseline. Test scores of standard Utkarsh students at low- and mid-baseline-score-percentiles are higher than the control group in all subjects across. Flexible Utkarsh students' endline scores in all subjects are higher than the control group across the entire distribution of the baseline test scores.

Figure 3: Treatment Effect on Grade Level Mastery



Note: This figure is showing the treatment effects on English, math, and Odia grade level mastery. A student is considered to have a specific grade level mastery in a subject if the student correctly answered at least 75 percent questions that relate to that specific grade's learning level.

Figure 4: Math Competency by Mean Baseline Competency

School-Level Impacts of Utkarsh on Math Competency by Mean Baseline Competency

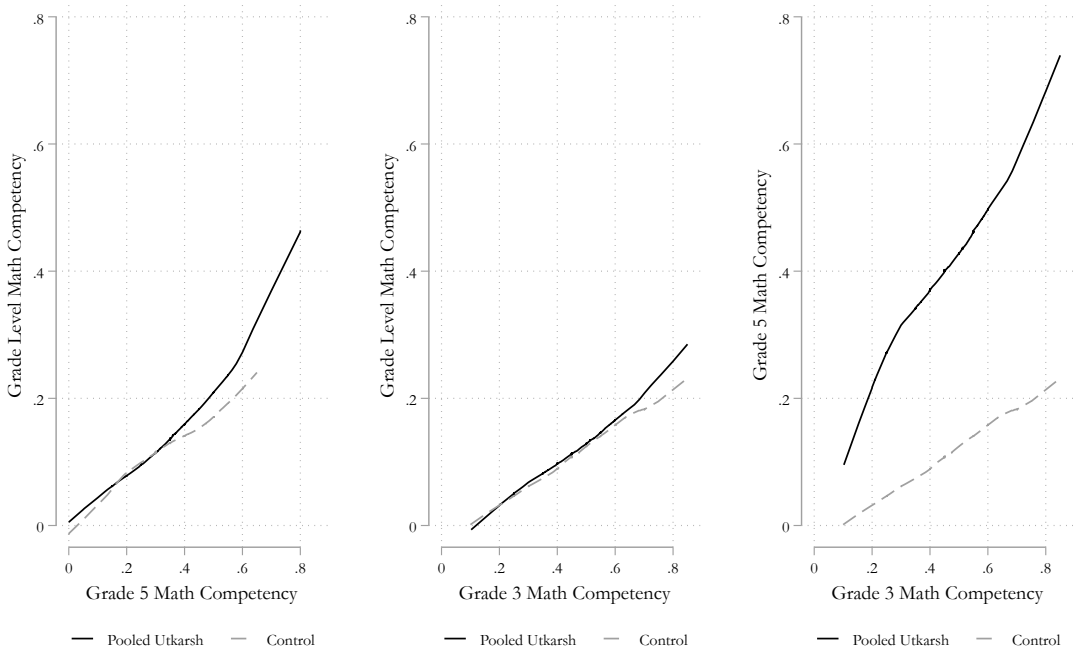
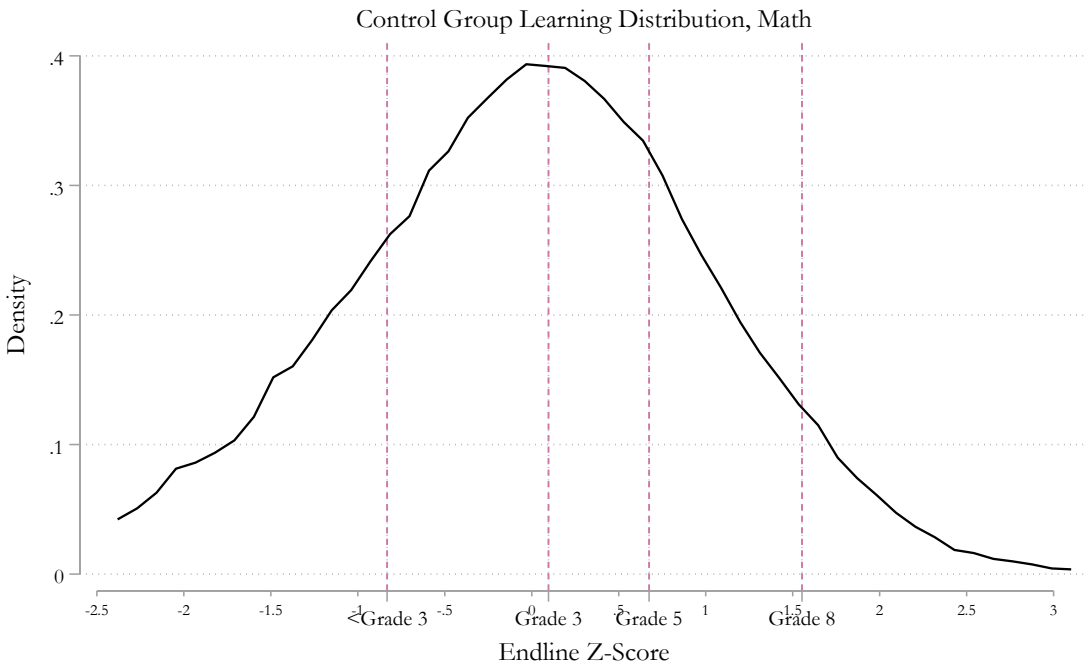


Figure 5: Control Group Math Learning



kernel = epanechnikov, bandwidth = 0.2010

Table 1: Treatment Effects on Test Score

	Overall	English	Math	Odia	Science
	(1)	(2)	(3)	(4)	(5)
Standard Utkarsh	0.107***	0.118***	0.119***	0.089***	0.104***
	(0.015)	(0.020)	(0.020)	(0.017)	(0.031)
Flexible Utkarsh	0.110***	0.116***	0.123***	0.086***	0.143***
	(0.015)	(0.018)	(0.021)	(0.017)	(0.032)
Observations	5,448	5,448	5,448	5,448	5,448
Control-group change (baseline to endline)	0.19	0.21	0.06	0.21	N/A
Standard Utkarsh=Flexible Utkarsh (*p*- value)	0.81	0.21	0.85	0.89	0.21

Notes: All regressions include strata, week, and day-of-week fixed effects; standardized IRT scores from baseline English, Math, and Odia tests, a dummy for being female, age of the pupil, and indicator variables for endline interview phase. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. *** p<0.01 ** p<0.05 * p<0.1

Table 2: Implementation of Utkarsh

Outcomes	Standard Utkarsh	Flexible Utkarsh	N	Standard Utkarsh=Flexible Utkarsh (*p* -value)
	(1)	(2)	(3)	(4)
Panel A: Outcomes measured in all or SLP phase of spot visit, or during endline				
Conducted levelling assessment	0.945***	0.956***	299	0.60
	(0.015)	(0.015)		
Number of days taught Utkarsh in the previous six days	4.958***	5.062***	298	0.28
	(0.080)	(0.079)		
Did Utkarsh worksheet on the day of the survey or on the recent most day the teacher taught	0.977***	0.943***	298	0.04
	(0.012)	(0.014)		
Implementing the correct phase	0.940***	0.965***	299	0.17
	(0.016)	(0.013)		
Students who attended FC met the inclusion criteria (teacher reported)	0.833***	0.853***	298	0.57
	(0.028)	(0.028)		
Students who attended SLP met the inclusion criteria (teacher reported)	0.804***	0.796***	233	0.87
	(0.039)	(0.036)		
Students who attended CC met the inclusion criteria (teacher reported)	0.776***	0.761***	289	0.74
	(0.033)	(0.033)		
At least 75 percent of lesson matches Utkarsh curriculum during FC	0.795***	0.812***	298	0.68
	(0.033)	(0.035)		
Followed Utkarsh lesson exactly as instructed in the lesson guides	0.977***	0.985***	295	0.60
	(0.011)	(0.010)		
Students currently using handbooks in class	0.980***	0.913***	244	0.08
	(0.025)	(0.032)		
Classroom has a word wall	0.093**	0.123***	299	0.55
	(0.043)	(0.042)		
Student desks arranged in small group	0.798***	0.781***	249	0.77
	(0.051)	(0.051)		
Panel B: Outcomes measured only at schools where spot visit took place during FC phase				
Teaching planned lesson during FC	0.862***	0.900***	84	0.50
	(0.058)	(0.047)		
Number of days taught Utkarsh in the previous six days that teacher followed planned Utkarsh during FC	4.284***	4.775***	84	0.00
	(0.250)	(0.172)		

Note : Table is showing school level implementation fidelity of the Utkarsh program. Outcomes in Panel A are measured at either a) all or SLP phase of spot visit, or b) during endline. Outcomes in Panel B are measured only at schools where spot visit took place during FC phase. Teacher reported measure of whether students attending each phase met the inclusion criteria does not consider any inclusion error. This outcomes are set to zero for all control group schools. This outcome was not measured for SLP phase if the spot visit took place during the FC phase, hence the sample size is less than other variables. Each regression includes average teacher age and its square, average teacher experience and its sqaure and share of female teachers. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. *** p<0.01 ** p<0.05 * p<0.1

Table 3: Flexibility Take-up

Measured at:	*Endline*		*Midline*				
	Teacher had some autonomy in using the Utkarsh lesson in class	Teacher has control over course timetable if students have difficulty in understanding topic	Teacher has the flexible teaching plan	Filled out flexible Utkarsh teaching plan	Followed flexible Utkarsh teaching plan	At least 75 percent of lesson matches Utkarsh curriculum during SLP	Followed standard Utkarsh lesson plan during SLP
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Standard Utkarsh	0.944***	0.007	0.010	0.009	0.009	0.597***	0.659***
	(0.013)	(0.030)	(0.019)	(0.019)	(0.015)	(0.042)	(0.041)
Flexible Utkarsh	0.989***	0.071**	0.237***	0.217***	0.161***	0.646***	0.567***
	(0.009)	(0.030)	(0.040)	(0.038)	(0.032)	(0.050)	(0.036)
Observations	834	834	565	565	565	565	565
Control group mean	0.00	0.77	0.00	0.00	0.00	0.00	0.00
Standard Utkarsh=Flexible Utkarsh (*p*- value)	0.00	0.03	0.00	0.00	0.00	0.40	0.05

Notes : All regressions include strata, week, and day-of-week fixed effects, teacher's age in years and age squared, teacher's years of experience and experience squared, a dummy for teacher being female, and a vector of dummy variables for main subject taught by teacher in the baseline. Columns 1-2, measured at endline, includes an indicator variable for early endline visit and Columns 3-7, measured at midline, include indicator variables for monitoring visit phase. Columns 1-2 are teacher reported measures of having autonomy in using Utkarsh lessons and having control over course timetable, respectively. Column 3 measures whether the teacher has shown the flexible teaching plan, and Column 4 measures whether the teaching plan is filled out. Column 5 measures whether teachers followed the teaching plan that they filled out. Column 6 is a self-reported measure of how much tachers followed the Utkarsh curriculum. Column 7 measures whether teacher followed the teaching plan of standard Utkarsh or not during SLP. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. *** p<0.01 ** p<0.05 * p<0.1

Table 4: Impact Teaching Practices

	Teaching and learning material visible	At least one student had an opportunity to express their own idea	Teacher asked a question to the class	Teacher answered students' questions supportively	Teacher answered students' questions without showing disrespect	Teacher did not ignore students' questions	Teacher seemed familiar with content	Teacher encourages student	Teacher responds to questions satisfactorily	Teaching with student interaction	Teaching Practices Index
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Standard Utkarsh	0.012	0.181**	-0.032	0.075	0.008	0.006	-0.009	0.162**	0.130*	0.146**	0.385**
	(0.012)	(0.073)	(0.049)	(0.076)	(0.027)	(0.028)	(0.029)	(0.075)	(0.078)	(0.071)	(0.159)
Flexible Utkarsh	0.013	0.186**	-0.043	0.051	0.004	-0.005	-0.037	0.149**	0.159**	0.134*	0.346**
	(0.013)	(0.078)	(0.045)	(0.077)	(0.028)	(0.026)	(0.031)	(0.075)	(0.076)	(0.073)	(0.157)
Observations	290	290	290	290	290	290	290	290	290	290	290
Control mean	0.989	0.495	0.895	0.526	0.958	0.968	0.979	0.295	0.463	0.568	0.000
Standard Utkarsh=Flexible Utkarsh (*p*- value)	0.910	0.938	0.810	0.764	0.860	0.608	0.404	0.867	0.698	0.854	0.801

All outcomes are observed by enumerators while teacher is teaching in the classroom. All regressions include strata, week, and day-of-week fixed effects, teacher's age in years and age squared, teacher's years of experience and experience squared, a dummy for teacher being female, and a vector of dummy variables for main subject taught by teacher in the baseline. Column 11 is constructed by first calculating the proportion of varibles in Columns 1-10 that is true for each observation and then standardize that with respect to the control group. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. *** p<0.01 ** p<0.05 * p<0.1

Table 5: Teachers' Perception of Student

	All Teachers		*English Teachers Only*			*Math Teachers Only*		
Measure of Students' Ability	Will pass the board exam	Will eventually complete a bachelor's degree	Can write a simple English sentence			Can do a three digit minus two digit substraction		
Outcome Variable	Teacher Forecasts that ….Percent of Student	Teacher Forecasts that ….Percent of Student	Teacher Forecasts that ….Percent of Student	Actual School-Average Percent of Student …	Teacher Forecast-Actual School Average of Percent of Student…	Teacher Forecasts that ….Percent of Student	Actual School-Average Percent of Student …	Teacher Forecast-Actual School Average of Percent of Student…
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Standard Utkarsh	-4.111**	-5.975***	-9.513***	3.218	-12.731***	-5.233**	0.913	-6.146**
	(1.712)	(1.980)	(3.373)	(1.954)	(3.461)	(2.639)	(2.251)	(3.094)
Flexible Utkarsh	-5.453***	-7.491***	-7.676**	5.824***	-13.500***	-7.474***	3.222	-10.696***
	(1.626)	(1.811)	(3.164)	(1.959)	(3.413)	(2.652)	(2.429)	(3.327)
Observations	836	836	302	302	302	323	323	323
Control group mean	60.600	49.560	59.29	17.10	42.19	76.29	55.55	20.74
Standard Utkarsh=Flexible Utkarsh (*p*- value)	0.435	0.428	0.55	0.19	0.82	0.40	0.33	0.16

Notes: Sample is restricted to teachers who primiarliy teach Utkarsh subjects. All regressions include strata, week, and day-of-week fixed effects. All regressions include teacher's age in years and age squared, years of experience and experience squared, a dummy for being female, a vector of dummy variables for main subject taught in the baseline, and indicator variables for early endline visit. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. *** p<0.01 ** p<0.05 * p<0.1

Table 6: Board Exam Results

		Conditional on the board exam scores' presence										
		Grades				*Test Score in Percentage*						
	Passed	A	B or above	C or above	D or above	English	Math	Odia	Science	Utkarsh Total	Non-Utkarsh Total	Total
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Standard Utkarsh	-0.003	-0.011**	-0.056***	-0.069***	-0.046**	-0.018**	-0.020**	-0.019***	-0.021***	-0.019***	-0.029***	-0.023***
	(0.003)	(0.005)	(0.019)	(0.022)	(0.018)	(0.008)	(0.009)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)
Flexible Utkarsh	-0.004	-0.002	-0.021	-0.046**	-0.028	-0.014*	-0.007	-0.012*	-0.013	-0.012*	-0.011	-0.012
	(0.003)	(0.006)	(0.020)	(0.022)	(0.018)	(0.008)	(0.009)	(0.007)	(0.008)	(0.007)	(0.008)	(0.007)
Observations	18,551	18,551	18,551	18,551	18,551	18,551	18,551	18,551	18,551	18,551	18,551	18,551
Control group mean	0.99	0.05	0.39	0.65	0.85	0.53	0.58	0.56	0.55	0.55	0.57	0.56
Standard Utkarsh=Flexible Utkarsh (*p*- value)	0.88	0.15	0.11	0.33	0.35	0.60	0.17	0.36	0.37	0.30	0.04	0.14

Notes: All regressions include strata fixed effects; standardized IRT scores from baseline English, Math, and Odia tests, a dummy for being female, and age of the pupil. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. Colun 14 is restricted to our baseline study sample. *** p<0.01 ** p<0.05 * p<0.1

Appendix Table 1: Balance Table (Student and School)

	Control	Standard Utkarsh	Flexible Utkarsh	*p* -value from test of equality (1)=(2)=(3)
Panel A: Student-Level Variables	(1)	(2)	(3)	(4)
Female (=1)	0.495	0.490	0.511	0.61
Age (in years)	13.163	13.148	13.152	0.88
Scheduled caste, scheduled tribe, or other backward caste (=1)	0.595	0.588	0.635	0.06*
No parent can read and write (=1)	0.142	0.157	0.154	0.64
Participated in Utthan (=1)	0.878	0.892	0.879	0.40
Takes private tuition (=1)	0.733	0.714	0.731	0.71
Baseline English test	0.007	-0.055	0.046	0.09*
Baseline Math test	-0.011	-0.016	0.027	0.62
Baseline Odia test	-0.006	-0.022	0.027	0.52
Baseline English competency	2.651	2.532	2.582	0.04**
Baseline Math competency	3.449	3.365	3.452	0.49
Baseline Odia competency	3.931	3.845	3.933	0.54
Observations	1,949	1,876	1,931	
Panel B: Headmaster- and School-Level Variables	(1)	(2)	(3)	(4)
Female headmaster (=1)	0.26	0.20	0.19	0.46
Age of the headmaster (in years)	52.46	52.44	52.44	1.00
Experience in current position (years)	5.53	5.83	4.70	0.43
Headmaster thinks teacher should follow curriculum (=1)	0.74	0.76	0.80	0.56
Headmaster considers ensuring adherence to curriculum as important job component (=1)	0.95	0.98	0.96	0.45
Headmaster thinks they have influence on determining delivering curriculum lessons (=1)	0.84	0.79	0.77	0.41
Head of school's designation: Headmaster or in charge (=1)	0.95	0.92	0.92	0.59
Sanctioned class 9 teacher posts in the school	7.55	7.83	7.38	0.31
Number of teacher posts filled	5.19	4.98	5.16	0.57
Total enrollment in class 9	72.38	62.51	72.87	0.02**
Observations	99	100	100	

Notes: Table is showing reported characteristics of the respondents during the baseline survey. Standard errors are clustered at school. ***, **, and * indicate significance at the 1, 5, and 10 percent levels.

Appendix Table 2: Balance Table (Teacher)

	Control	Standard Utkarsh	Flexible Utkarsh	*p* -value from test of equality
				(1)=(2)=(3)
Panel B: Teacher-Level Variables	(1)	(2)	(3)	(4)
Female (=1)	0.49	0.47	0.50	0.85
Age of the teacher (in years)	41.83	43.08	41.22	0.23
Have a teaching certificate (=1)	0.34	0.36	0.35	0.81
Teaching experience (years)	16.29	17.77	15.63	0.12
Teacher utkarsh subject (=1)	0.99	1.00	1.00	0.36
Days absent from work	0.37	0.40	0.36	0.82
Works in another school(=1)	0.03	0.02	0.02	0.87
Time spent in preparing for lesson (hours/week)	12.05	12.61	12.78	0.76
Time spent grading (hours/week)	8.32	7.89	8.75	0.15
According to the teacher, percent of student who				
Will pass board exam in first try	63.23	61.63	64.14	0.51
Can write a simple English sentence	55.06	53.02	56.96	0.15
Can do a three digits sum	70.45	72.76	72.99	0.17
Select teaching materials, methods, strategies (=1)	0.95	0.96	0.94	0.64
Allowed to modify course timetable (=1)	0.83	0.78	0.78	0.15
Others select evaluation activities (=1)	0.65	0.67	0.67	0.74
Feel pressure to complete curriculum during school year (=1)	0.27	0.26	0.30	0.64
Burnout index	0.00	-0.10	-0.09	0.53
Autonomy index	0.00	-0.10	-0.16	0.13
Teacher feels (=1)				
Curriculum should be followed even if students have lower learning level	0.58	0.56	0.62	0.38
That if students' not being ready for board exam would be teacher's own fault	0.46	0.48	0.46	0.96
Valued and appreciated	0.69	0.70	0.76	0.08*
Satisfied with job	0.85	0.83	0.85	0.72
That their opinion seems to count	0.90	0.91	0.91	0.81
That they have the materials and equipment to teach effectively	0.62	0.61	0.68	0.26
Similarly or more effective compared to colleagues	0.95	0.94	0.95	0.82
Observations	276	277	278	

Notes: Table is showing reported characteristics of the respondents during the baseline survey. Standard errors are clustered at school. ***, **, and * indicate significance at the 1, 5, and 10 percent levels.

Appendix Table 3: Treatment Effect Heterogeneity by Student Gender

	Overall	English	Math	Odia	Science
	(1)	(2)	(3)	(4)	(5)
Standard Utkarsh*Female	0.123***	0.146***	0.106***	0.114***	0.113***
	(0.018)	(0.024)	(0.024)	(0.024)	(0.040)
Standard Utkarsh*Male	0.091***	0.091***	0.132***	0.065***	0.094**
	(0.019)	(0.024)	(0.028)	(0.022)	(0.039)
Flexible Utkarsh*Female	0.115***	0.119***	0.102***	0.111***	0.159***
	(0.018)	(0.024)	(0.026)	(0.022)	(0.040)
Flexible Utkarsh*Male	0.105***	0.114***	0.145***	0.062***	0.125***
	(0.019)	(0.023)	(0.028)	(0.021)	(0.042)
Observations	5,448	5,448	5,448	5,448	5,448
Control-group change (baseline to endline)	0.19	0.21	0.06	0.21	N/A
Standard Utkarsh: Female=Male (*p*- value)	0.12	0.21	0.44	0.10	0.70
Flexible Utkarsh: Female=Male (*p*- value)	0.64	0.86	0.22	0.07	0.51

Notes: All regressions include strata, week, and day-of-week fixed effects; standardized IRT scores from baseline English, Math, and Odia tests, a dummy for being female, age of the pupil, and indicator variables for endline interview phase. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. *** p<0.01 ** p<0.05 * p<0.1

Appnedix Table 4 : Treatment Effect Heterogeneity by Student Caste

	Overall	English	Math	Odia	Science
	(1)	(2)	(3)	(4)	(5)
Standard Utkarsh*SC/ST/OBC	0.112***	0.129***	0.117***	0.100***	0.081**
	(0.018)	(0.024)	(0.024)	(0.021)	(0.039)
Standard Utkarsh*Other Castes	0.097***	0.101***	0.112***	0.070***	0.140***
	(0.020)	(0.027)	(0.027)	(0.024)	(0.044)
Flexible Utkarsh*SC/ST/OBC	0.117***	0.140***	0.111***	0.093***	0.155***
	(0.017)	(0.022)	(0.025)	(0.021)	(0.039)
Flexible Utkarsh*Other Castes	0.106***	0.082***	0.154***	0.080***	0.128***
	(0.019)	(0.025)	(0.029)	(0.024)	(0.046)
Observations	5,448	5,448	5,448	5,448	5,448
Control-group change (baseline to endline)	0.19	0.21	0.06	0.21	N/A
Standard Utkarsh: SC/ST/OBC=Other Castes (*p*- value)	0.49	0.37	0.89	0.31	0.26
Flexible Utkarsh: SC/ST/OBC=Other Castes (*p*- value)	0.60	0.05	0.20	0.65	0.63

Notes: All regressions include strata, week, and day-of-week fixed effects; standardized IRT scores from baseline English, Math, and Odia tests, a dummy for belonging to ST/SC/OBC, being female, age of the pupil, and indicator variables for endline interview phase. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. *** p<0.01 ** p<0.05 * p<0.1

Appendix Table 5: Treatment Effect Heterogeneity by Student's First Generation Learner Status

	Overall	English	Math	Odia	Science
	(1)	(2)	(3)	(4)	(5)
Standard Utkarsh*First Generation	0.096**	0.092*	0.136***	0.114***	0.028
	(0.037)	(0.048)	(0.048)	(0.042)	(0.070)
Standard Utkarsh*Not First Generation	0.109***	0.123***	0.118***	0.086***	0.114***
	(0.015)	(0.020)	(0.020)	(0.018)	(0.033)
Flexible Utkarsh*First Generation	0.111***	0.104**	0.133***	0.153***	0.017
	(0.036)	(0.045)	(0.051)	(0.041)	(0.073)
Flexible Utkarsh*Not First Generation	0.111***	0.119***	0.124***	0.076***	0.163***
	(0.015)	(0.018)	(0.021)	(0.017)	(0.034)
Observations	5,448	5,448	5,448	5,448	5,448
Control-group change (baseline to endline)	0.19	0.21	0.06	0.21	N/A
Standard Utkarsh: First Generation=Not First Generation (*p*- value)	0.19	0.52	0.70	0.50	0.25
Flexible Utkarsh: First Generation=Not First Generation (*p*- value)	0.19	0.73	0.86	0.06	0.06

Notes: All regressions include strata, week, and day-of-week fixed effects; standardized IRT scores from baseline English, Math, and Odia tests, a dummy for being a first generation learner, a dummy for being female, age of the pupil, and indicator variables for endline interview phase. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. *** p<0.01 ** p<0.05 * p<0.1

Appendix Table 6: Attrition by Treatment Arms

	(1)	(2)	(3)	(4)	(5)
	Missing Endline Score	Missing Endline Score	Missing Endline Score	Missing Endline Score After First Endline	Missing Endline Score after Second Endline
Standard Utkarsh	-0.023**	-0.023**	-0.023**	-0.005	-0.017
	(0.011)	(0.011)	(0.011)	(0.022)	(0.012)
Flexible Utkarsh	-0.013	-0.013	-0.013	-0.011	-0.006
	(0.011)	(0.011)	(0.011)	(0.020)	(0.012)
Standard Utkarsh*Overall baseline test score			0.019*		
			(0.011)		
Flexible Utkarsh*Overall baseline test score			0.003		
			(0.012)		
Overall baseline test score			-0.069		
			(0.060)		
Endline Visit = Early		-0.012			
		(0.030)			
Observations	5,756	5,756	5,756	5,756	5,756
Control Mean	0.06	0.06	0.06	0.23	0.08
Standard Utkarsh=Flexible Utkarsh (*p*- value)	0.27	0.27	0.28	0.78	0.30

Notes: Regression includes strata fixed effects, standardized IRT scores from baseline English, Math, and Odia tests, a dummy for being female, age of the pupil, dummies for missing control variables. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. *** p<0.01 ** p<0.05 * p<0.1

Appendix Table 7: Lee Bounds

	Overall Test Score		English Test Score		Math Test Score		Odia Test Score		Science Test Score	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound	Lower Bound	Upper Bound
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Standard Utkarsh	0.103***	0.124***	0.110***	0.142***	0.106***	0.133***	0.079***	0.105***	0.075**	0.134***
	(0.015)	(0.015)	(0.020)	(0.019)	(0.020)	(0.019)	(0.017)	(0.017)	(0.030)	(0.031)
Flexible Utkarsh	0.106***	0.123***	0.110***	0.131***	0.113***	0.130***	0.081***	0.098***	0.125***	0.168***
	(0.015)	(0.014)	(0.018)	(0.018)	(0.021)	(0.021)	(0.017)	(0.017)	(0.031)	(0.032)
Observations	5,382	5,383	5,382	5,383	5,382	5,383	5,382	5,383	5,382	5,383
Adjusted R-squared	0.888	0.893	0.807	0.809	0.800	0.803	0.839	0.836	0.445	0.435
Control Mean	0.24	0.24	0.27	0.27	0.10	0.10	0.26	0.26	0.00	0.00
Standard Utkarsh=Flexible Utkarsh (*p*- value)	0.87	0.98	1.00	0.49	0.72	0.91	0.86	0.62	0.09	0.28

Regression includes strata fixed effects. All regressions include standardized IRT weights from baseline English, Math, and Odia tests, a dummy for being female, age of the pupil, dummies for missing control variables.
Heteroskedasticity-robust standard errors, clustered by school, in parentheses. *** p<0.01 ** p<0.05 * p<0.1

Append Table 8: Treatment Effect Heterogeneity by Baseline Test Score

	Overall	English	Math	Odia	Science
	(1)	(2)	(3)	(4)	(5)
Standard Utkarsh*BL Test Score Bottom Third	0.123***	0.165***	0.106***	0.102***	0.086
	(0.028)	(0.032)	(0.030)	(0.032)	(0.054)
Standard Utkarsh*BL Test Score Middle Third	0.104***	0.124***	0.166***	0.100***	0.042
	(0.018)	(0.030)	(0.027)	(0.023)	(0.046)
Standard Utkarsh*BL Test Score Top Third	0.088***	0.058**	0.079**	0.056**	0.192***
	(0.017)	(0.023)	(0.032)	(0.025)	(0.046)
Flexible Utkarsh*BL Test Score Bottom Third	0.126***	0.175***	0.124***	0.122***	0.045
	(0.027)	(0.031)	(0.035)	(0.031)	(0.052)
Flexible Utkarsh*BL Test Score Middle Third	0.099***	0.115***	0.110***	0.070***	0.202***
	(0.019)	(0.028)	(0.027)	(0.023)	(0.045)
Flexible Utkarsh*BL Test Score Top Third	0.101***	0.056***	0.142***	0.066**	0.173***
	(0.017)	(0.020)	(0.031)	(0.026)	(0.045)
Observations	5,448	5,448	5,448	5,448	5,448
Control-group change (baseline to endline)	0.19	0.21	0.06	0.21	N/A
Standard Utkarsh: Bottom Third=Middle Third=Top Third (*p*- value)	0.51	0.01	0.06	0.33	0.06
Flexible Utkarsh: Bottom Third=Middle Third=Top Third (*p* -value)	0.61	0.00	0.69	0.29	0.04

Notes: In each regression, treatment indicators are interacted with indicatores for tercile of baseline test score for the respective test. Since there is no science baseline, overall baseline test score is used to estimate the heterogeneity of treatment effect on science. All regressions include strata, week, and day-of-week fixed effects; standardized IRT scores from baseline English, Math, and Odia tests, a dummy for being female, age of the pupil, and indicator variables for endline interview phase. Heteroskedasticity-robust standard errors, clustered by school, in parentheses.
*** p<0.01 ** p<0.05 * p<0.1

Appendix Table 9: Student's Self-Assessment and Expected Educational Outcomes

		Expected Score in Class 10				
	Self-assessed Rank among Peers (1-10)	English	Math	Odia	Science	Highest Level of Education Hope to Achieve: Bachelor's degre or above
	(1)	(2)	(3)	(4)	(5)	(6)
Standard Utkarsh	-0.067	0.093	0.314	0.576	0.957*	0.030*
	(0.047)	(0.567)	(0.592)	(0.553)	(0.537)	(0.017)
Flexible Utkarsh	-0.050	0.525	0.757	0.710	0.806	0.007
	(0.055)	(0.582)	(0.576)	(0.521)	(0.522)	(0.017)
Observations	5,397	5,397	5,397	5,397	5,397	5,397
Control mean	4.06	62.67	66.77	70.48	64.86	0.51
Standard Utkarsh=Flexible Utkarsh (*p*- value)	0.76	0.47	0.46	0.82	0.80	0.17

Notes: All regressions include strata, week, and day-of-week fixed effects. All regressions include standardized IRT weights from baseline English, Math, and Odia tests, age of the pupil, and indicator variables for endline interview phase. In adition, Column 1 includes self-assessed rank among peers at baseline, Columns 2-5 include expected score in class 10 board exam at baseline, and Column 6 includes an indicator variable for baseline level aspiration for highest level of education. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. *** p<0.01 ** p<0.05 * p<0.1

Appendix Table 10: Treatment Effects on PISA and Foundation Camp Test Scores

	PISA questions		
	English+Math	English	Math
	(1)	(2)	(3)
Standard Utkarsh	0.047	0.064*	0.016
	(0.032)	(0.035)	(0.034)
Flexible Utkarsh	0.071**	0.029	0.073**
	(0.032)	(0.034)	(0.035)
Observations	5,448	5,448	5,448
Control-group change (baseline to endline)	0.22	0.15	0.19
Standard Utkarsh=Flexible Utkarsh (*p*- value)	0.50	0.33	0.09

Notes: All regressions include strata, week, and day-of-week fixed effects; standardized IRT scores from baseline English, Math, and Odia tests, a dummy for being female, age of the pupil, and indicator variables for endline interview phase. Heteroskedasticity-robust standard errors, clustered by school, in parentheses.
*** p<0.01 ** p<0.05 * p<0.1

Appendix Table 11: Additional Teacher-level Outcomes

	Burnout index	Stress index	Anxiety index	Lesson preparation time (hours/week)	Grading time (hours/week)	Students benefitted from Utkarsh	Teacher enjoyed Utkarsh	Teacher benefitted from Utkarsh	Self-Assessed Effectiveness		
									Atleast As Effective as Other Teachers	More Effective Than Other Teachers	Much More Effective Than Other Teachers
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Standard Utkarsh	0.105	0.059	0.025	-0.229	0.410	0.749***	0.953***	0.868***	0.012	-0.005	-0.007
	(0.089)	(0.082)	(0.090)	(0.625)	(0.558)	(0.031)	(0.014)	(0.019)	(0.012)	(0.041)	(0.021)
Flexible Utkarsh	0.135	0.082	0.049	-0.842	0.259	0.717***	0.940***	0.919***	0.003	0.020	-0.021
	(0.089)	(0.090)	(0.096)	(0.649)	(0.551)	(0.030)	(0.016)	(0.020)	(0.013)	(0.040)	(0.020)
Observations	834	834	834	834	834	834	834	834	834	834	834
Control group mean	0.000	0.000	0.000	13.65	8.94	0.00	0.00	0.00	0.971	0.412	0.076
Standard Utkarsh=Flexible Utkarsh (*p*- value)	0.751	0.786	0.794	0.34	0.78	0.43	0.54	0.04	0.388	0.517	0.505

Notes: All regressions include strata, week, and day-of-week fixed effects. All regressions include teacher's age in years and age squared, years of experience and experience squared, a dummy for being female, a vector of dummy variables for main subject taught in the baseline, and indicator variables for early endline visit. Burnout index is an inverted covariance matrix weighted standardized indix generated following Anderson (2008) from the following variables: teacher feeling mentally exhausted from work; feeling fatugied; feeling having a positive influence on people; feeling very energetic about job; feeling satisfied with job at school. Stress and anxiety are measured on the Depression Anxiety Stress Scale (DASS). To construct respectived indices, response to relevant questions of the scale are summed and then standardized. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. *** p<0.01 ** p<0.05 * p<0.1

Appendix Table 12: Teacher Competency

		Teacher Surveyed at Baseline		*Test Score (Percent)*	
	Test Administered in School (=1)	Subject Taught			
		English	Math	English	Math
	(1)	(2)	(3)	(4)	(5)
Standard Utkarsh	0.009	0.069	-0.008	-0.476	1.067
	(0.019)	(0.072)	(0.056)	(2.399)	(1.591)
Flexible Utkarsh	-0.015	0.050	-0.078	-0.943	2.909**
	(0.023)	(0.076)	(0.058)	(2.655)	(1.413)
Observations	300	226	303	226	303
Control mean	0.800	0.653	0.636	57.826	89.212
Standard Utkarsh=Flexible Utkarsh (*p*- value)	0.242	0.763	0.235	0.833	0.244

Notes: Outcome variable in Column 1 indicates that the school participated in the teacher competency examination. Outcome variables in Columns 2 and 3 are indicator variables whether the teacher was surveyed at baseline. Outcome variables in Columns 4 and 5 are test scores in respective subjects by respective subject teachers on the teacher competency examination. All regressions include strata fixed effects. Column 1 includes a dummy for headmaster being female, headmaster's age and age squared, headmaster's yeras of experience and years of experience squared, school having multiple class 9 sections, and total enrollment. Columns 2-5 include week-of-survey, and day-of-week fixed effects, a dummy for teacher being female, age of the teacher and age squared, teacher's years and experience and experience squared. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. *** p<0.01 ** p<0.05 * p<0.1

Appendix Table 13: Follow-up Survey Details

	Board exam scores of our study sample collected	Analytical Sample Students Surveyed in the Follow-up Survey	Student Took offline exam
	(1)	(2)	(3)
Standard Utkarsh	-0.010	0.019	0.018
	(0.020)	(0.013)	(0.014)
Flexible Utkarsh	0.021	0.008	0.011
	(0.019)	(0.014)	(0.015)
Observations	5,756	5,457	1,253
Control group mean	0.917	0.221	0.055
Standard Utkarsh=Flexible Utkarsh (*p*- value)	0.128	0.406	0.642

Notes: All regressions include strata fixed effects; standardized IRT scores from baseline English, Math, and Odia tests, a dummy for being female, and age of the pupil. Heteroskedasticity-robust standard errors, clustered by school, in parentheses. *** p<0.01 ** p<0.05 * p<0.1