When Given Discretion, Teachers Did Not Shirk: Evidence from Remedial Education in Secondary Schools

Sabrin Beg, Anne Fitzpatrick, Jason T. Kerwin,

Adrienne M. Lucas, and Khandker Wahedur Rahman*

July 21, 2025

Abstract

How rigid should bureaucracies be? More standardization means less scope for shirking—but also less flexibility to solve problems. We examine this tradeoff in secondary schools in Odisha, India, where students test far below benchmarks and teachers are pressured to stick to the grade-level curriculum. Schools were randomized to a rigid remedial learning intervention, a variation that encouraged teacher discretion over lesson selection, or control. Both interventions increased learning by 0.11SD (58 percent), with gains throughout the distribution. Remedial education neither displaced grade-level mastery nor reduced progression to upper secondary. Greater discretion did not lower implementation quality or induce shirking.

JEL Codes: H40, I25, I28, J24, M54, O15.
Keywords: Randomized trial, public service delivery, flexibility, remedial education, secondary school, India

*Beg: University of Delaware and J-PAL. Fitzpatrick: The Ohio State University. Kerwin: University of Washington, IZA, and J-PAL. Lucas: University of Delaware, J-PAL, and NBER. Rahman: University of Oxford and BIGD, BRAC University. 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 Rebecca Thornton, Sandra Sequiera and seminar participants at Bentley University, Boston University, IFPRI, Lafayette College, Universite Laval, the University of Delaware, University of Illinois Urbana-Champaign, University of Nevada Reno, Wellesley College, the University of Maryland, Miami University, as well as our discussants at the following conferences: AEFP, AFE, CSAE, NEUDC, and PAA 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. Funding for this project was provided by the UK Foreign, Commonwealth & Development Office, awarded through J-PAL's Post-Primary Education Initiative, and the Kusuma Trust UK. AEA RCT #AEARCTR-0004138. None of the authors have any potential conflicts of interest.

Introduction

Government bureaucracies are famously bound up in red tape. Rigid standards prescribe what civil servants can and cannot do—which sometimes prohibits the correct solution to problems. This bureaucratic inertia has been noted to be a problem in the public sector since the 1940s (Jackson and Morgan, 1982), and is even more true in the developing world (Rauch and Evans, 2000). One motivation for this standardization is to make monitoring easier, thereby minimizing shirking and ensuring that the public receives at least a minimal level of public services (Afridi et al., 2017; Bandiera et al., 2020; Banerjee et al., 2021; Muralidharan and Singh, 2020). Yet these rigid rules often ignore local conditions and may limit the ability of front-line civil servants to make the best decisions. This means there is a trade-off between reducing shirking and allowing public servants the flexibility to solve problems.

This trade-off between flexibility and preventing shirking is particularly stark in education, and governments tend to come down on the side of standardization. The typical approach to managing public education is to have bureaucrats set a standardized curriculum and enforce its delivery. Indian government schools are often cited as an example of a highly regimented public-sector organizational structure with poor service delivery outcomes (Muralidharan et al., 2019; Muralidharan and Singh, 2020). Teachers (who are the front-line civil servants) are directed to follow and complete the grade-level curriculum on a strict timetable decided by the state education office, enforced by high-stakes testing (National Steering Committee for National Curriculum Frameworks, 2023). This system of standardized teaching with (supposedly) uniform implementation is designed to create high expectations of teachers and students, forestall teacher shirking, and decrease monitoring costs. Nevertheless, approximately half of secondary school students fail to meet standardized benchmarks, and many are well behind grade level (Das and Zajonc, 2010).

An alternative approach is to let teachers choose the level at which to teach to maximize overall student learning based upon their classroom's needs. In India, this would mean

allowing teachers to teach remedial instruction if they choose. However, there are clear challenges to this approach. First, while teaching students at their learning level instead of their grade level increases achievement in primary schools (Banerjee et al., 2007; Duflo et al., 2024), whether remedial instruction will work in secondary schools is unclear. The range of learning levels in secondary school classrooms can be vast, which could make remedial lessons more difficult to target and less effective than they are in primary schools. Second, it may not be optimal to invest substantial teaching effort in students who are substantially behind—it may just be too late. Third, some students *are* at the level of the curriculum; introducing remedial lessons during class time may harm them by crowding out grade-level material. Finally, whether this would work without direct institutional support is unclear. Teaching remedial lessons requires unrewarded teacher effort: in the status quo, teachers do not have the training, materials, or incentives to teach at anything other than grade level. Variations in teaching practice would also be more difficult to monitor, and shirking would potentially look similar to teaching a lesson that is "too easy." There may also be concerns with teacher competency, particularly at the secondary school level. Overall, allowing for teacher discretion on what classroom topics to teach in a setting with weak institutional capacity and no support may create confusion, lower achievement for students, or exacerbate already high levels of absenteeism and shirking (Muralidharan and Sundararaman, 2011; Chaudhury et al., 2006).

In this paper, we address fundamental questions relevant to teaching worldwide: what should be taught in a classroom, and who should make that decision? To answer this question, we implemented a three-arm randomized controlled trial in grade 9 across 300 secondary schools in Odisha, India. The RCT tested the effect of a new standardized curriculum that included remedial content, relative to a control group that kept the *status quo* curriculum. We also tested an alternative implementation of this new curriculum that allowed teachers more flexibility. This study was carried out in partnership with existing school personnel, the Odisha Department of School and Mass Education, and the Indian non-governmental

organization Transform Schools.

Our randomized experiment had two treatment groups, randomized at the school level, that were variations on the Utkarsh ("excellence") remedial program. In the Utkarsh program, teachers test students to determine their learning levels, then set aside prescribed hours during the school day to implement remedial lessons. In the first treatment group, existing teachers were given Utkarsh training, materials, and an adjusted timetable with specific days and hours that teachers were to engage in remedial teaching instead of the usual curriculum in a bootcamp style: multiple hours for a limited number of days were set aside for the lessons. These were not additional schooling hours; lessons directly displaced curriculum-level lessons, and only existing teachers were used to deliver the lessons. Teachers in the second treatment group received the same training, materials, and adjusted timetable with additional encouragement and flexibility to adjust the program based upon their own assessment of student needs. This Flexible Utkarsh arm gave teachers explicit permission to use part of the Utkarsh time for material they felt was a priority, whether remedial or grade-level. Teachers in this study arm received a scheduling handout that they used to note whether to follow the scheduled Utkarsh lesson for each day, or replace it with other content.

We report four main findings based on one school year of implementation, using immediate test scores and longer-term administrative data on student achievement. First, bureaucrats have standardized the *status quo* curriculum at the wrong level, and teachers are aware of this mismatch between the curriculum and their students. Teachers were largely aware that many of their students were behind grade level, in contrast to the prevailing understanding (based on primary schools) that teachers are unaware of how much students are lagging behind (Djaker et al., 2022). Our study students were on average 4-5 grades behind in English, math, and Odia (the local language). Teachers know their students are far behind, but do not realize just how far: the average control-group teacher thinks that barely half of their students can write a simple English sentence, while the true rate is a dismal 17 percent.

Second, both versions of Utkarsh improved student test scores by about 0.1 standard deviations (SDs) more than the status quo, equivalent to a 58 percent increase in growth compared to the control group. The likelihood of achieving Grade 3 and Grade 5 competency increased by between 3 to 8 percentage points, depending upon subject. The intervention did not crowd out grade-level knowledge, a common concern about introducing remedial lessons (Zhao, 2018; Figlio and Page, 2002). It also did not bring students to Grade 9-level mastery of the material. There was no effect on students' chances of qualifying to attend upper secondary school, consistent with no change to grade-level mastery. Overall, the program was also cost-effective, raising test scores by 0.95 SDs per $100 spent when implemented at a 200-school scale. Therefore, intervening with remedial instruction in secondary school is not too late. Instead, remedial instruction in secondary school allows many more students to be reached who might have fallen behind, and does not harm those at grade-level.

Third, implementation quality was high in both treatment arms despite teachers receiving no additional incentives to implement the program. The high degree of implementation is noteworthy. For teachers, knowing students are behind is not sufficient for them to teach non-grade level content. Instead, they need the tools, resources, and permissions that make teaching at an appropriate level for students an equal- or lower-effort task to teaching the curriculum. The flexibility offered in the Flexible Utkarsh intervention neither greatly altered how the program was implemented nor increased shirking, which addresses a key concern about allowing more decision-making at the point of service. Teachers usually taught Utkarsh during the designated times, but not always the exact lesson prescribed for that day—teachers in both treatment arms modified the program, but stuck with remedial lessons instead of reverting to the standard curriculum. Indeed, our classroom observations reveal that the majority of *Standard* Utkarsh teachers were covering a different Utkarsh topic than the official schedule. These findings imply that teachers have a revealed preference for both remedial instruction and structure—once they had additional support (such as teaching and learning materials) they modified classroom activities to better fit student learning abilities,

and were approximately equally likely to stay within the rigid confines of this new curriculum even when flexibility was encouraged.

Finally, at the end of the treatment, teachers had more accurate beliefs about their students' learning levels. Even though treatment teachers believed their students benefited from the program, the program reduced their beliefs about their students' ability levels. This made their perceptions more accurate: since students are far below grade-level benchmarks, these more pessimistic assessments were closer to the truth. This is reflected not just in survey responses but also in the marks that teachers assigned to students, which were lower in the treatment arms than in the control group.

We interpret these results through a conceptual framework closely related to the Dessein (2002) model from organizational economics, which outlines the conditions that make delegation of authority to local decision-makers optimal. Delegation is beneficial when local decision-makers (in our setting, teachers) have actionable information that the central authority lacks. Teachers were aware of their students' learning levels but were unable to take action without the additional tools provided in the Utkarsh program. We find limited benefit or cost from allowing added flexibility in the Utkarsh program. The model suggests that this is because the Utkarsh program's instructional support made it low-cost for teachers to provide optimal teaching, and also because the program content already correctly reflected the teachers' private information about student performance. The latter is possible because instruction was targeted to student ability via the leveling exams. Indeed, the program's effects on teacher perceptions suggest that they accumulated much of their private information through the leveling exams. Therefore, standardization does not inherently cause the poor provision of government services; instead, it is a mismatch between the standardization and the front-line reality that reduces productivity. This mismatch can be mitigated by creating government services that address the specific details of a problem (as done by the Utkarsh leveling exams).

This paper makes two major contributions to the economics literature. First, we provide

new evidence that remedial education provided by existing personnel, during the school day, is an efficient use of resources at the secondary level. Directing teachers to run the remedial program and providing them with materials to do so increased student test score growth by 0.11 SD, or 58 percent. Previous studies outside of the U.S. focusing on secondary school students have either studied interventions that take place outside of the school day (Lavy and Schlosser, 2005, Muralidharan et al., 2019, Chiplunkar et al., 2023) or that occur within school, but focus on grade level material (Beg, Lucas, Halim and Saif, 2019). Given that our intervention operated within the school day and used existing teachers, the most comparable studies are those evaluating within-school "'targeted instruction" at the primary school level. Angrist and Meager (2023) report an average effect of 0.07 SDs on average for within-school targeted instruction using existing teachers, which is smaller than our estimated effect size. The impact of Utkarsh on learning is comparable to the median across all types of education interventions (Evans and Yuan, 2022).

Second, we show that allowing more flexibility at the point of service need not lead to shirking: with the right support and training, governments can make good performance the optimal choice for civil servants. Much of the research in improving public sector service delivery focuses on providing incentives to service providers (Glewwe, Ilias and Kremer, 2010; Duflo, Dupas and Kremer, 2011; Muralidharan and Sundararaman, 2011; Duflo, Hanna and Ryan, 2012; Duflo, Dupas and Kremer, 2015; Barrera-Osorio and Raju, 2017; De Ree, Muralidharan, Pradhan and Rogers, 2018; Rasul and Rogger, 2018; Rasul, Rogger and Williams, 2018; Brown and Andrabi, 2020) or empowering community members to register complaints (Bjorkman and Svensson, 2009; Duflo et al., 2012). A growing literature emphasizes that making highly rigid bureaucracies more flexible may improve service delivery by increasing autonomy among front-line civil servants (Bandiera, Best, Khan and Prat, 2021; 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. Scripted or guided lessons can increase student learning,

and teacher modifications tend to decrease lesson quality (Gray-Lobe et al., 2022, Piper, Sitabkhan, Mejia and Betts, 2018). We find that treatment-group teachers replaced curriculum lessons with remedial lessons as instructed, and when given the opportunity to deviate back to the grade level curriculum, stuck with the remedial lessons, which was at a more appropriate learning level for their students. Allowing teachers additional flexibility was not worse than standardizing the intervention's implementation.

Our findings show that switching to remedial education (in lieu of the standard curriculum) is a cost-effective intervention for improving learning in secondary schools, and that adding flexibility to remedial education programs has a limited impact on their effectiveness. They also suggest that the Indian education sector may be less rigid than commonly thought: differential uptake of additional flexibility was limited in part because many teachers already adapted the lessons to student needs. Furthermore, well-designed remedial education programs can address the key information problem in the school system by measuring students' ability levels and using them to determine the correct level of the program for each student. The Utkarsh program has been scaled up both across the original state (Odisha, population 40 million) and also a new states (Karnataka, population 60 million; Chhattisgarh, 30 million; Haryana, 25 million; and Uttar Pradesh, 236 million). It has already reached more than 9.1 million students, and there are plans to expand the program further.

This paper is structured as follows. First, we provide background on the context and setting in Section 2; in Section 3, we describe the intervention in detail. We outline our empirical strategy in Section 4 and describe our sample creation in Section 5. We present our results in Section 6. In Section 7 we show how these results can be interpreted through the lens of our conceptual framework. In Section 8, we present cost-effectiveness. In Section 9 we discuss our results and conclude. We present additional figures in Appendix A, and additional tables in Appendix B. Appendix C we provide additional details on the intervention, and Appendix D we provide additional details on test construction.

Background

Our study takes place in Odisha, a relatively poor state in eastern India where about a third of the population is below the national poverty line (NITI Aayog, 2021). Many secondary school students are first-generation learners, and about 40 percent of enrolled students are from either a scheduled caste or a scheduled tribe. In India, pre-tertiary education schooling is 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). Teachers in lower secondary, the focus of this study, are subject teachers, teaching the same subject to Class 9 and 10 students in separate sections, and sometimes teaching multiple subjects across grades depending on the school size. As is typical in the Indian education system, schools at all levels in Odisha emphasize teacher-focused instruction, have many below grade-level students, and are characterized by wide variation in student ability levels within the same classroom. The typical class period involves lecture-based pedagogy, with limited pupil participation or deviation for students at ability levels below the expectations for their grade. The expectation is that the pace and content of the lessons strictly adhere to the official curriculum. At baseline, 95 percent of headmasters considered adhering to the curriculum to be an important component of their job.

Students must pass standardized, district-run Board Exams at the end of Classes 10 and 12 to continue their education.[1] The marks on the Board Exam determine which school students can attend and what field of future study students can undertake, and are intended to ensure adequate student grade-level competency.

The school year in Odisha starts in April, has a break from early May to mid-June, and ends in March. Our study follows students from the beginning of lower secondary school (Class 9) through their first year in higher secondary (Class 11). The intervention occurred only during Class 9.

[1] At other grades there are school-based exams that determine whether students are allowed to progress to the next grade. Compared to other low-income countries, India has one of the world's lowest grade repetition rates at 1 percent (Hares et al., 2020).

The Utkarsh Program

The Utkarsh program provided Class 9 teachers with training, teaching and learning materials, and a designated schedule to follow to deliver this content. The program was a collaboration between the Odisha School Education Programme Authority (the authority within the Department of School and Mass Education, or SMED, responsible for education) and Transform Schools, a large Indian NGO.[2] It focused on Odia, English, math, and science and was designed to improve learning outcomes for students who were below grade level. All instruction took place within the existing school day with existing teachers.

Subject teachers for English, Math, Odia, and Science as well as the school headmaster, were invited to a one-week training session in August of the 2019-2020 school year. At these sessions, participants learned how to use Utkarsh teaching and learning materials to implement a more effective teaching practice. The materials were remedial and covered the content that should have been covered in Classes 3 through 8, divided into specific Class-specific phases. The training also emphasized collaborative and student-centered active learning, including students sitting and working in groups. Daily outlines of topics to cover and accompanying student worksheets were provided, but lessons were not scripted. Appendix C contains additional details about the program.

Upon returning to school, teachers were to test all students in Class 9 to determine their ability levels based on provided rubrics. These "leveling exams" categorized each student's initial ability level as being either Inception (below Class 3), Class 3, Class 5, or 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. Each phase took place during the school day and was designed to displace regular lessons for a specific amount of time over a set number of days. The program instructed teachers to provide alternative activities for students who did not need remediation. Although the program was designed so that a student's level would determine participation in each phase,

[2]Transform Schools is a collaboration between People For Action, The Transform Trust, and Transform Schools UK.

in practice, teachers administered the full program to all students, whether or not they need remediation participated in all of the same lessons, in the same classroom, for all parts of the intervention.

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, about 90 percent of students in our sample. SLP further developed the FC concepts at a higher level and with more advanced skills, moving students from Class 5 to Class 7 level material. Teachers were to adhere to these lessons for 3 hours 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 after the start of the program, teachers again assessed all students on their ability levels in each subject.

Our study covers two versions of Utkarsh*:* Standard Utkarsh and Flexible Utkarsh. Standard Utkarsh is the original version of the Utkarsh program as described above. Flexible Utkarsh modifies the original version to add flexibility. Specifically, teachers and headmasters received all of the same training and materials as in Standard Utkarsh version described above. They were also instructed to implement both FC and CC as above. However, they were explicitly told that during SLP they could either follow the official Utkarsh lessons, or exercise their own discretion in planning the material and content. During the 3 hours per day of Utkarsh lessons that occurred over the 45 days of SLP, they could spend more time on a particular SLP topic, repeat previous topics from FC or SLP, or use the time for the standard curriculum instead of covering remedial content. This meant they could also skip SLP topics if their students did not need them. To facilitate and encourage the use of

flexibility, teachers were provided with a Flexible Utkarsh Plan, which was a worksheet on which teachers had to list the topics that they planned to cover for the week. These topic choices could be either following the standard Utkarsh schedule or the alternative topics the teacher chose. The training instructed teachers to cover at least 50 percent of the material from the standard SLP curriculum; they had flexibility to select the rest of the content. After the training there was no additional monitoring.

Empirical Strategy

The primary conceptual difficulty in assessing the effect of remedial education on student outcomes is the typical correlation between remedial instruction and student ability levels: weaker students are more likely to receive remedial education, leading to reverse causality and thus to biased estimates. Other student, teacher, or school characteristics may also be correlated with the likelihood of receiving remedial education; stronger students may have better teachers who are more used to adapting curriculum, for example. To overcome this difficulty, we conducted a randomized trial, randomly assigning 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:

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

where y_{ist} is the outcome of interest for respondent i in school s at time t. $StandardUtkarsh_s$ and $FlexibleUtkarsh_s$ are indicator variables for the treatment status of school s. These indicators are mutually exclusive with the control group as the omitted category. X_{ist} is a 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.[3] Standard errors are clustered at the school level.

Our coefficients of interest are β_1, the effect of Standard Utkarsh relative to the control group, and β_2, the effect of Flexible 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 at the conclusion of the intervention. To understand the mechanisms behind test score changes, we also estimate the effect of the interventions on teacher classroom behavior, practices, and perceptions of students.[4] We also analyze the effects on longer-term student outcomes.[5]

Sample Selection, Randomization, and Data

Sample Selection and Randomization

To arrive at our 300 school sample we started with an administrative list of all 711 secondary schools in Jajpur and Dhenkanal districts in Odisha State, India. Schools that did not report any students enrolled in Class 9 were eliminated, leaving 348 villages with at least one

[3]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. Our strata are 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.

[4]Because we have only one wave of follow-up data, t is constant in most of our specifications. Our classroom observation data includes several waves and so t takes multiple values for those analyses.

[5]We filed a pre-analysis plan (PAP) prior to collecting the endline data for the study, which is available at https://www.socialscienceregistry.org/versions/59999/docs/version/document. We adhere to the PAP exactly for the use of test score outcomes as our primary outcome of interest and for our choice of regression specification. We also study the same three primary hypotheses, which are about the effects of the two versions of Utkarsh and the difference between them, and similarly examine test scores, implementation, attendance, and other outcomes. However, we deviate from the PAP in several key ways. First, we changed how we applied IRT to construct the test scores. In the PAP, we said that we will apply IRT only to the endline scores. In the paper, we apply IRT to both waves jointly to allow us to calculate the growth in test scores between the two test score rounds. This small modification does not change the key findings. We still conduct multiple hypothesis testing, but limit the groupings to the family of test score outcomes. Third, for non-test score outcomes such as implementation, we changed the coding, changing the exact list of items in the family to be more intuitive and improve the interpretability of our results. Per the original PAP, these other analyses cover separate families of outcomes from test scores, and thus there would not have been joint multiple testing adjustment across these two categories. While key results are not sensitive to these changes, we caution that these analysis choices were made post-hoc.

secondary school. To minimize contamination, we randomly selected one secondary school from each of these villages. We randomly ordered these schools and directly confirmed with each one that it used the official state language (Odia), was governed by the SMED and not the Scheduled Caste-Scheduled Tribe Development Department, had students enrolled in Class 9, and was not a school for students with special needs (e.g., deaf or blind students). We proceeded down the randomly-ordered list of schools until we reached 300 schools that passed the screening criteria. We placed each of the 300 schools that passed the screening test 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.[6] Within each stratum, we randomized an equal number of schools into the three treatment conditions, resulting in 100 schools in each of the three treatment arms. Figure 1 shows this study design.

Data Collection

We conducted four waves of data. Three were collected during the the year of implementation (academic year 2019-2020): a baseline survey, an unannounced monitoring visit when treatment schools should have been engaged in Utkarsh, and a full follow-up at the conclusion of the intervention. The year after the intervention, we conducted an additional follow-up survey via phone. We augment these data with administrative data from 2021. Figure 2 shows the study timeline.

Baseline

The baseline surveys took place in July and August 2019, near the start of the school year but after the summer break and prior to the implementation of Utkarsh. We collected 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 D for additional test construction details.

[6] Four of the strata had 12 schools while the other 42 had six schools apiece.

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, treatment schools should have been implementing FC and SLP. We randomly assigned each school to receive 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. During each visit, enumerators arrived unannounced and recorded the attendance of the headmaster, teachers, and baseline students. Headmasters and teachers responded to questions about program take-up and implementation. We also conducted classroom observations.[7]

Endline

We conducted the endline data collection from December 2019 to February 2020, slightly overlapping with the conclusion of the intervention in December 2019.[8] Students responded to a short student survey that included a question about their Board Exam 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 test all students from the baseline sample. Our analysis sample is all 5,448 students who completed both the baseline and endline surveys and assessments. During this follow-up visit, we also conducted surveys of teachers and the school headmaster. The teacher survey asked teachers about their experience, autonomy, Utkarsh implementation, workload, and perceptions about Utkarsh. We also administered a competency test in English and math to teachers to test their knowledge of these two

[7]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; 2 schools did not consent.

[8]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. We started the endline in December to complete data collection prior to schools beginning their preparations for end-of-term exams. The treatment effect for test scores are the same for early versus late endline.

subjects.[9] The headmaster survey included information on their school and characteristics as well as their personal background and school management practices. We confirmed the Board Exam registration information for each student in the sample with their headmaster.

To maximize the response rate for this follow-up, we followed DiNardo et al. (2021) and randomized the intensity of our mop-up visits to survey respondents who were absent during the follow-up visit.[10] Specifically, we conducted second mop-up visits in a random subset of schools where students remained absent during the first mop-up visit.[11] We implement Lee bounds (Lee, 2009) to address potential biases of treatment effects due to non-random coverage of respondents in the follow-up, although this procedure does not change our key findings.

Follow-Up Survey

We conducted an additional follow-up survey via phone in December 2021, after the COVID school closures and two years after the end of the program, to measure the impact of the program on longer-term school enrollment and the transition to additional schooling or work.[12]

Administrative Board Marks

We test for the effect of the program on students' longer term outcomes using their Class 10 Board Marks. We planned to acquire Class 10 Board Exam results, but due to the Covid-19 pandemic, the May 2021 Board Exams were canceled. Instead, students received Board Marks based on a weighted average of their teacher-assigned Class 9 (40 percent) and Class 10 (60 percent) marks. Teachers did not know in advance that their scores would be used for Board Marks. As a response to students who objected to this grading scheme, an optional Board Exam was eventually administered, which approximately 5 percent of

[9]We attempted to survey the same teachers over time, adding teachers as necessary and collecting demographics as they were added.

[10]Despite conducting additional mop-up visits, in some cases we were unable to confirm student Board Exam registration and corresponding numbers for all students because the headmaster did not have time to or could not access digital copies of the Class 10 Board Exam registration.

[11]The success rate at the first interview attempt was 78%. The success rate at the first mop up (second interview attempt) was 69%. The success rate at the second mop up (third attempt) was 29%

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

Class 10 students completed. The Board Marks recorded in the administrative data are the maximum of the school-based weighted average and the Board Exam. Unfortunately, the administrative data do not denote whether the student sat for the formal Board Exam, or whether the final Board Marks are from the Board Exam or the school-based weighted average.

Summary Statistics and Baseline Balance

Appendix Tables B1 and B2 show that randomization successfully created three groups with balanced characteristics at baseline at the student, school, and teacher level. 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 either a scheduled caste, a scheduled tribe, or other backwards caste, the disadvantaged minority groups in India. About 15 percent of students in our sample have illiterate parents. 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. Teachers report very little absenteeism from work. Teachers report spending about 21 hours each week preparing lessons and grading. Teachers believe that approximately 60 percent of their students will pass their Board Exams on their first try.

Our headmaster survey indicates the dearth of autonomy in schools to adjust the curriculum to meet the varying ability levels and needs of students. About 77 percent of headmasters in our baseline survey also share the view of teachers that the 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. About 22 percent of headmasters are women. 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, a randomization inference-based F-test of joint balance (as recommended in Kerwin et al. 2024) across all variables in Panel B yields a p-value of 0.59, indicating no overall balance issues.

Results

Student Outcomes

Baseline Achievement We begin by documenting the existing low student performance levels to provide context for our study. In all of the three baseline subjects, the mean student is over four grade levels below Class 9 (Appendix Figure A1).[13] Nearly half of all students are evaluated below Class 3, but there is substantial heterogeneity in grade-level mastery: nearly 8 percent of students in Math and English, and 18 percent of students in Odia, are at grade level.[14] This heterogeneity exists both across schools and within schools—in math, the average interquartile range of competency (i.e., the mean within-school difference between the 75th and 25th percentiles) is 3.94 grade levels. Therefore, on average, although nearly half of Class 9 students are below Class 3 competency in a given classroom, our test scores suggest that for up to a fifth of students remedial instruction is not needed, and may even be harmful. Teachers are relatively accurate in their estimates of student proficiency—they only overestimate the percent of their students who are at or above Class 5 by about 6 percentage points.

Follow-Up Achievement We now show the main effects of the program on test score growth. The Utkarsh program improved student growth across all four target subjects. Each version of Utkarsh increased students' overall test scores by about 0.11 SD (Table 1, column 1). Over this same period, overall test scores for the control group increased

[13]Grade-level mastery is based on the percent correct of baseline questions from that grade level's curriculum. The exams contained no grade 1 or 2 material. Students below Class 3 competency are given a competency of Class 2 even thought their actual performance level might have been Class 1.

[14]These results are similar to teacher-collected data from the program leveling test, which also reveal a high share of students with very low performance (see Appendix Table B3).

by 0.19 SD. Therefore, Utkarsh improved test score growth by 59 percent relative to the status quo, regardless of implementation approach. Utkarsh's effects on subject-specific scores are similar to its overall effects: the program increased English and Math scores by 0.12 SD (columns 2 and 3), Odia scores by 0.09 SD (column 4), and Science scores by 0.10 SD (Standard Utkarsh) and 0.14 SD (Flexible Utkarsh, column 5). Relative to the control group's rate of test score gains over the same period, the program increased growth in English by 57 percent, Math by 190 percent, and Odia by 43 percent.[15] These results are robust to multiple hypothesis testing corrections (Appendix Table B5).[16] In Appendix Tables B6-B8, we test for heterogeneity by gender, scheduled caste and scheduled tribe, or first generation learner status and find some evidence that Utkarsh language instruction was especially beneficial to female students.

Utkarsh increased students' effective grade levels, but did not bring them up to Class 9-level mastery. Recall that at baseline students were on average over 4 grade levels behind. Figure 3 shows the effect of the two versions of Utkarsh on achieving different levels of mastery for English, Math, and Odia. The bars indicate treatment effects on the probability of achieving each grade level for a given subject. The program increased the likelihood that students were at least grade 3 or grade 5 in Math and at least grade 5 in Odia. Neither intervention improved the likelihood of Class 8 mastery in any subject. However, the interventions also did not decrease the likelihood of Class 8 mastery, a common concern about remedial programs. One reason that Utkarsh may have not brought students all the way up to grade level is the relatively short time period that the program ran; another is the wide

[15] We cannot compare science increases to control-group growth since student did not take a baseline science test. In Appendix Table B4 we show that Utkarsh also increased test scores on a subset of 5 questions selected from the PISA in Math and 4 questions selected from the PISA in English, showing that the results are not merely due to teaching to the test. Recall that PISA is a test that evaluates the performance of 15 year-olds worldwide in reading, mathematics, and science. 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 questions, although we can only reject the equality of the two effects for the latter, and only at the 10 percent level.

[16] We report adjusted q-values using the Haushofer and Shapiro (2016) implementation of the Anderson (2008) Family-Wise Error Rate (FWER) adjustment.

distribution of initial student ability levels in relation to grade-level competency. Figure A2 shows the distribution of endline math test scores for the control group. In math, students at Grade 3 competency in Math are close to the overall mean, while students who are at Grade 8 competency are approximately 1.5 standard deviations above the mean. Therefore, to bring students up to grade level would require a 1.5-SD improvement in test scores, approximately 15 times the observed Utkarsh treatment effect, and larger than the effects of virtually all education programs (Evans and Yuan, 2022).

Even though they did not necessarily reach grade level competency, students throughout the baseline test score distribution benefited from the program. In Appendix Figure A3, we plot non-parametric test score effects and find gains in student growth throughout the baseline test score distribution. Appendix Table B9 divides students into terciles based on their baseline test scores. For English, we reject equality across the terciles, finding the largest effects for the lowest tercile (0.17 SD) and smaller effects for the highest tercile (0.06 SD). For Science, we also reject equality across the three terciles, only finding statistically significant effects for the top tercile for Standard Utkarsh (0.19SD) and the top two terciles for Flexible Utkarsh (effect sizes of 0.20 SD for the middle third and 0.17 SD for the top third). There is no evidence of treatment effect heterogeneity for overall scores, Odia, or Math.

Attrition As with any RCT, one concern is attrition at the follow-up generating differential selection into the test. To limit attrition, we attempted to follow-up with all students from the baseline, even those who were not present in school the day of the follow-up visit. As a result, overall attrition was very low: only 6 percent in the control group. Appendix Table B10 estimates differences across study arms in the likelihood that students completed the achievement follow-up. Students in the Standard Utkarsh arm were two percentage points more likely to complete the achievement follow-up (column 1), and this is differential by baseline test score with lower scoring baseline students more likely to complete the achievement follow-up in the standard arm (column 2). Because of this differential attrition,

we constructed treatment bounds following Lee (2009). As shown in Appendix Table B11 in all cases both the magnitude and statistical significance are similar to the main effects.

Non-Cognitive Outcomes The Utkarsh program could have been encouraging to students because they were being taught at their ability level or discouraging because they were told that they needed remedial attention. We find no evidence of discouragement and limited evidence of encouragement (Appendix Table B12). The treatment does not affect students' self-reported ranking among peers or their estimated Board Exam scores in English, Math, or Odia. Students in the Standard Utkarsh arm report marginally higher expected Science scores (1 percentage point over a control group mean of 65 out of 100) and are 3 percentage points more likely to desire a bachelor's degree, relative to a control-group mean of 51 percent.

Classroom Practices

Teachers in the treatment arms improved their teaching practices and implemented the Utkarsh program (Table 2). The effects on test scores in the previous sub-section were not because teachers were more likely to be present at the start of the school day or teaching during a classroom observation (columns 1 and 2). Instead, the treatments increased the quality of classroom teaching by 0.35 (Flexible) and 0.39 percentage points (Standard)—classrooms became more active and engaging and more likely to involve interactions between teachers and students.[17] Teachers also implemented specific aspects of Utkarsh beyond active pedagogy—on average each school implemented about 81 percent of the fourteen different components (column 4).[18] The interventions did not significantly change the likelihood that headmasters were present, although the point estimates are positive (see Appendix Table B15). The changes in teaching practices are also not driven by differential teacher attrition

[17]Results for each of the specific items of the teaching practices index are in Appendix Table B13.

[18]These components of Utkarsh were common to the two interventions. Results for each of the specific items of the implementation index are in Appendix Table B14. This includes both teacher-reported and enumerator-observed aspects. Results are similar when limited only to enumerator-observed components.

from the unannounced monitoring visit (see Appendix Table B16).

Curriculum

The previous two subsections showed that student test score growth and teaching quality increased equally across the two arms. In this section we address whether the curriculum delivered across the two arms differed—when teachers were given the flexibility to diverge from the standard Utkarsh timetable, did they? Table 3 contains these outcomes.

Only about 20 percent of teachers in the Flexible arm completed the Flexible Utkarsh teaching plan and 15 percent reported that they followed a Flexible plan during the week of observation (columns 1 and 2). Teachers in the Standard arm were supposed to follow the prescribed lessons while those in the Flexible arm were instructed to either complete a Flexible plan and follow that or continue to follow the Standard plan. Completing a Flexible plan is neither a necessary nor sufficient condition for engaging in flexibility. Even though almost all teachers in both arms felt like they had autonomy over using Utkarsh, a statistically significant 4 percent more of Flexible teachers agreed with this statement (column 3). Teachers in the Flexible arm were also 7 percentage points more likely to say that they could adjust topics or pace if students were struggling with a concept (column 4). Somewhat surprisingly, given the system's emphasis on completing the curriculum, a majority of control teachers also reported that they could do this.[19] Therefore, the majority of teachers already feel like they can adjust the course content, but they largely do not, instead relying on teaching a curriculum that is multiple grade levels above their students' ability levels. Thus, the teachers are enacting the wishes of the bureaucrats, even while acknowledging there is limited oversight to enforce that directive.[20]

[19]The exact question phrasing is "If students need more time to understand a topic, I am allowed to modify the course timetable." It is possible that control-group teachers interpret this as meaning they can make adjustments within the grade-level curriculum, rather than across grade levels.

[20]In Appendix Table B17 we examine the characteristics of teachers who "take up" the flexibility. Younger, less experienced teachers are less likely to take up flexibility. There are also some differences by the subject taught and responses to the baseline survey, although these patterns are not particularly consistent across definitions of teacher flexibility. In Appendix Table B18 we show that teachers in the Flexible arm with a wider range of student abilities in the classroom are more likely to respond that they have discretion in the

We further test for the implementation of flexibility by comparing what teachers said they were covering relative to the Standard Utkarsh schedule. Table 4 compares teachers in the Flexible arm to those in the Standard arm. About 95 percent of teachers reported doing an Utkarsh lesson during the week (column 1). Teachers in both treatments made Utkarsh their own—only 43 percent were doing the prescribed Standard lesson for that week (column 2). Relative to the prescribed schedule, Flexible teachers were more likely to deviate widely: 82 percent of Standard teachers were within a week of the Standard schedule and Flexible teachers were 9 percentage points less likely to be covering a lesson on that same approximately on time schedule.

Overall, teachers in both arms embraced some level of flexibility. The additional flexibility exerted by teachers in the Flexible arm did not lead to more shirking or differentially change student test scores. Flexibility and discretion are touted as ways to improve worker motivation and decrease burnout (Skaalvik and Skaalvik, 2014). Conversely, this new program could have increased teacher stress and anxiety. We find similar levels of teacher burnout, stress, and anxiety across all three arms. Teachers further did not alter their self reported lesson preparation time or grading time—any Utkarsh preparation time was offset by a decrease in curriculum-level preparation time (Appendix Table B19).

Teachers' Perceptions of the Program, Students, and Themselves

We asked teachers their opinions about the program, their students, and themselves. Teachers overwhelmingly believed that they and their students benefited from Utkarsh (Table 5, columns 1 and 2). Yet they lowered their estimates of the percentage of their students who could perform a basic literacy (writing a simple English sentence) or numeracy (three-digit minus two-digit) operation by 5 percentage points, which made their beliefs more accurate. In addition to lowering their expectations about specific tasks, teachers reduced their expectations about the percent of students who would pass the Class 10 Board Exam by about 5

classroom at endline.

percentage points from a control group mean of 61 percent (column 6).[21] They also reduced their estimates of how many students would earn a bachelor's degree by about 4 percentage points from the control group estimate of 50 percent (column 7).

Despite believing that they and their students benefited from Utkarsh, it did not lead teachers to self-assess themselves as being any more effective than teachers at schools similar to theirs, perhaps because they assessed students more accurately and realized how little their students knew (Appendix Table B20). As another measure of teaching effectiveness, we also implemented tests of teacher competency.[22] The program at most marginally objectively improved teacher content knowledge, improving the percent correct on a math test designed to be at the 4th grade level (Appendix Table B20).

Board Marks and Longer Run Outcomes

Students in our sample were scheduled to take the high-stakes Class 10 Board Exam one year after the Utkarsh program ended, in June 2021. Board Exams scheduled for June 2021 were canceled due to the Covid-19 pandemic. Instead, 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 tests written and scored by the students' own teachers during the school year.[23] Therefore, the Board Marks were determined by teachers rather than a state-wide standardized test. Thus, any effects on Board Marks incorporate both changes to teachers' perceptions as well as differences in underlying student performance. The Board Mark measures we use are both the continuous score as well as the binary pass/no-pass.

The interventions did not change the likelihood that a student received passing Board Marks (Table 6, column 1).[24] In our study schools, the Board pass rate was over 99 percent,

[21] Statewide the average annual pass rate from 2012 and 2020 (pre-pandemic) was between 71 and 85 percent; the control group was likely underestimating the pass rate.

[22] This test of teacher competency was developed by the World Bank. Teachers were given a fictitious homework assignment to grade and were evaluated based upon how many mistakes they caught.

[23] Five percent of students chose to take an actual Board Exam. The "Board Marks" we use are the maximum of the school-based and test-based scores. We do not know which students sat for a Board Exam.

[24] To maximize power we use the entire sample from our study schools, not only those students who are part of the analysis sample used above.

similar to the statewide average of 98 percent, which was 13 percentage points higher than a previous historical average. The Standard Utkarsh treatment lowered the average Board Marks of students by 0.16 standard deviations. These results are consistent with teachers' diminished perceptions of students as a result of the program persisting well after the program ended, although they could also reflect lower grade-level competency. The lower scores in non-Utkarsh subjects suggest spillovers from perceptions of the directly targeted subjects (where teachers received information that their students were struggling) onto perceptions of student ability more generally. While these lower Board Marks did not decrease pass rates, they suggest that Utkarsh increased teachers' knowledge of how far behind their students are. The lack of an effect on pass rates parallels Chiplunkar et al. (2023), who also find that a remedial education program in Indian secondary schools did not affect Board Exam pass rates.

Appendix Figure A5 shows non-parametric plots of the distribution of Board Marks by study arm. If Utkarsh reduced the test score growth of students who were initially at grade level, one would expect to see lower Board Marks specifically for students at high levels of baseline competency, who were closer to grade-level mastery at baseline. We do not observe such effects. Instead, this plot suggests that—similar to the results at the end of the intervention—there is no crowd-out for students near grade level at baseline in the Flexible Utkarsh arm. Consistent with the average treatment effects in Appendix Table B21, there is no noticeable gap in Board Marks between the control group and the Flexible Utkarsh arm anywhere in the distribution. However, students throughout the Standard Utkarsh distribution receive lower marks, lower than students in either the control group or the Flexible Utkarsh group. While this could indicate crowd-out of grade-level competencies, it is also consistent with teachers lowering their perceptions of their student's performance. Regardless, these potentially negative impacts are only observed in the Standard Utkarsh arm, suggesting that allowing teachers flexibility may be preferable to more rigid service delivery.

As an additional test of longer run outcomes, we conducted a phone survey in November and December 2021.[25] Despite lower average Board Marks in the two treatment arms, the interventions did not affect whether the student was enrolled in school, enrolled in Class 11, or employed—consistent with the lack of a treatment effect on receiving passing Board Marks (see Table 6, Columns 6-8). Thus, although the treatment caused students to receive lower Board Marks, we observe no changes in other long-term outcomes of interest.

Conceptual Framework

To interpret the results of the RCT, we build on the seminal model of Dessein (2002). This model outlines the conditions under which it is optimal for a principal (for example, company owner, or, in this case, a bureaucrat) to delegate decision-making authority to a subordinate (in this case, a teacher). We apply this model to the question of who should make the decision of what to teach in a particular classroom. While a large number of models in political science outline the trade-offs of decision-making authority versus delegation in bureaucracies—such as Gailmard (2002) and Jo and Rothenberg (2014)—several aspects of the Dessein (2002) model are particularly relevant to modeling the choice of whether to prescribe a centrally-set curriculum or instead allow for point-of-service modifications to what is taught. First, we assume that bureaucrats have a preference for maximizing student test score growth, and have the authority to decree instructional content, such as teaching a given curriculum or instead remedial education. However, they lack key information regarding local conditions: specifically, they do not observe the distribution of ability of the students at a particular school in the same detail as teachers do. Thus, they must make a decision about the curriculum without knowing which specific choice would maximize growth. Teachers (the "agents") have this information, but may have different preferences from the bureaucrats. For example, they may wish to shirk, have a preference for a certain pedagogy, or fear that

[25]The coverage rate in the phone survey was 23 percent, uncorrelated with treatment status (Appendix Table B22).

revealing the truth of how behind their students are will get them into trouble. Furthermore, the bureaucrats cannot design a contract with teachers to elicit the private information regarding the actual abilities of their students and act upon that information. Instead, bureaucrats may either: 1) take the teachers' reports on the local conditions and make the decision regarding what content the teacher should teach; or, 2) allocate decision-making power on what content to teach to the teacher. In other words, the bureaucrat must decide whether the bureaucracy should make the decision regarding instructional content, after consulting with civil servants (who may give biased reports) or, instead, to delegate the decision to teachers.

The bureaucrat is considering various policies over classroom teaching; these different curriculum guidelines are captured by $y \in \mathscr{R}$. Since bureaucrats have a preference for maximizing student growth, their payoff is given by student test score levels

$$U_B(y, m) = L(y, m)$$

where student ability, m, is a random variable with density $f(m)$, bounded over some range and y is the content of classroom instruction. Following Dessein (2002) we define the bureaucrat's utility function as:

$U_B(y, m) = U_B(m, m) - \lambda(|y - m|)$

where λ is a function with a positive second derivative and $\lambda(0) = 0$. Student test scores, and thus the bureaucrat's utility, are maximized when $y = m$, i.e. when the content is matched to the ability level of the students. However, the bureaucrat does not observe m,while the the teacher does. The teacher's utility is $U_T(y, m; b)$ where b is an additional parameter that captures the extent to which the bureaucrat's preferences differ from those of the teacher or could capture an effort cost to implementing a specific y. Thus their utility is maximized when $y = m + b$:

$U_T(y, m; b) = U_T(m + b, m) - \lambda(|y - (m + b)|)$

Because only the teacher knows the observed value of m, the bureaucrat has two choices:

either delegate authority (i.e., allow the teacher to choose y), or fix the curriculum in a centralized manner, setting $y = E[m]$.[26]

Proposition 1: Delegation is optimal if teachers' preferences are sufficiently close to those of the bureaucrat.

Dessein (2002) shows that, together with general assumptions on the structure of the private information, delegation (i.e., allowing the teacher to choose y) is optimal if and only if the difference term b is smaller than some cutoff value. Thus, delegation is optimal (leading to $y = m$, and thus to high-fidelity implementation of Utkarsh) for a range of possible differences in preferences, and does not require that the preferences of the teacher and the bureaucrat are in alignment; the teacher's preferences just cannot be overly different from the bureaucrat's. Note that this holds even though the bureaucrat cannot directly control what action the teacher ultimately chooses in the classroom.

The degree to which teacher preferences agree with the preferences of the bureaucrat is ultimately an empirical question. Shirking is a concern, both on the extensive and intensive margin. For example, Chaudhury et al. (2006) find that as many as 19 percent of teachers are absent at any given time; other studies have similarly found high rates of both absenteeism and low levels of effort on the job (Duflo et al., 2012). Our data shows a similar pattern, with 16 percent of the control-group teachers being absent from the classroom at the start of the observation (Table 2, Column 1), although many teachers showed up once they realized they were being observed (Column 2). Furthermore, teachers' preferences over what to teach in the classroom may be different than that of policymakers and other stakeholders. For example, if policymakers dictate that teachers begin implementing remedial instruction, they may or may not do so. One indication of the degree of disagreement between teachers and bureaucrats is whether the school curriculum is taught under the *status quo*: secondary schools in Odisha impose a standard curriculum, rather than tailoring content to student ability levels, even though students are frequently far behind grade level and teachers are

[26]The bureaucrat could also ask the teacher what the value of m, but the teacher's incentive is to strategically misreport to achieve their own preferred value of y, and so this is equivalent to delegating authority to the teacher.

aware of it. One potential measure of the degree of disagreement between bureaucrats and teachers is the dispersion of test scores within a geographic region or school. The model predicts that the program should work better where this dispersion is lower. We explored this possibility by running a number of exploratory tests of treatment effect heterogeneity by variation in within-school test scores, and found no evidence of this pattern (results available upon request).

We now consider how to extend the model to explain both the Standard Utkarsh and the Flexible version. Utkarsh changes the decision problem in two key ways. First, Utkarsh directly measures student ability in the classroom, allowing the bureaucrat to set y in a more granular way. Specifically, the leveling exams act as an informative signal s of each student's ability, and bureaucrats can now choose between delegation and setting $y = E[m|s]$. Since test scores are maximized at $y = m$,we have the following result:

Proposition 2: Utkarsh increases student growth if there is no delegation.

This is consistent with our empirical estimates of the effect of the program on test score growth, if we assume that there is no delegation under either the status quo or in the treatment group.

The second way Utkarsh changes the decision problem is that it provides instructional support: training and teaching and learning materials to facilitate remedial instruction. This lowers effort costs for teachers to implement the program, effectively bringing their preferences more in line with those that maximize student test scores. We model this as modifying the difference term from b to $b_S = (1-k)b$, where $k \in (0,1)$ represents the instructional support that teachers receive. This modification makes teacher utility into the following:

$U_T(y, m; b) = U_T(m + b_S, m) - \lambda(|y - (m + b_S)|)$

This yields the following result:

Proposition 3: Flexible Utkarsh will increase student growth relative to Standard Utkarsh unless the leveling exams are highly informative or the instructional support is

ineffective.

We can see that this is true by considering the limiting cases. Suppose that the instructional support is completely effective, so $k = 1$. Then delegation (i.e. the Flexible version of the program) maximizes student growth. Less effective support will lead to lower growth. Alternatively, suppose that the leveling exams are completely informative, so $E[m|s] = m$. In that case, student growth is maximized when y is dictated centrally and allowed to vary based on s, as in Standard Utkarsh, and Flexibility cannot increase student growth.

How do our results relate to this prediction? Unlike the bureaucrat in the model, our data allows us to directly observe teacher behavior. We see very high adherence to the curriculum absent the intervention, and the program makes teachers more aware that many students are behind. We also see high implementation fidelity in the Flexible Utkarsh study arm, which is consistent with the model's predictions if instructional support is highly effective and teacher preferences are effectively matched with those of the bureaucrat. However, we also see very little difference in outcomes for the Flexible arm, which suggests that the signal from the leveling exams is highly informative for the bureaucrat. This implies that the leveling exams overcome any informational barriers that would prevent student growth from being maximized. The effects of the program on teachers' beliefs about student performance suggest that much of teachers' private information about student ability is actually derived from the assessments that are conducted as part of the program, most importantly the beginning-of-year leveling exams. More broadly, other remedial education programs like TaRL that also use pre-program tests to measure student ability levels and change classroom content accordingly may have limited benefits from additional teacher flexibility for the same reason. Our results support this interpretation: teachers prefer to change classroom content to be more effective (when it is at the level of the student) and have preferences in line with the bureaucrat. However, they rely upon support from the bureaucrat to make such large changes such as changing typical classroom practice.

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. We are not aware of any previous estimates of the cost-effectiveness of secondary-school interventions in developing countries. However, this cost-effectiveness estimate is comparable to that for two different middle-school programs that were also evaluated in South Asia. An after-school personalized tutoring intervention known as Mindspark for students primarily in grades 7 and 8 increased test scores by 0.93 SD per $100 at 50-school scale (Muralidharan et al., 2019), and the eLearn program which introduced school screens and videos in Pakistan for students in middle school increased test scores by 1.4 SD per $100 at 200-school scale (Beg et al., 2019).

Discussion and Conclusion

Public sector services in developing countries typically have poor service delivery outcomes, some of which may partly be due to highly regimented service delivery. In this paper we analyze the introduction of a remedial instruction program to help better understand the causes and challenges of improving educational productivity. Our evaluation leads to two important empirical findings. First, 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, much of this variation is within classrooms: the typical classroom has a range of student competencies of 3.94 grades, although approximately 10 percent of students are at grade level. With these substantial learning gaps, as well as substantial heterogeneity, one fear about introducing remedial education programs is that they may crowd out grade-level skills and stall progress for students who are at grade level. Another concern is that allowing teachers to teach remedial education will be equivalent to lowering standards, and parents and policymakers

alike worry that allowing teachers to teach at the level of the student in their classroom may provide a disincentive for teachers to work hard, and will not help students pass high-stakes tests. Our randomized evaluation empirically evaluates these concerns and finds that they are unfounded. We find that introducing a remedial education program known as Utkarsh substantially improved student growth, increasing progress by 58 percent relative to the status quo—and did not crowd-out grade-level competencies (although they also did not improve). Thus, one key finding is that in contexts where many students are behind, introducing remedial instruction benefits students without the feared consequences.

Our second empirical finding is that allowing more flexibility in service delivery—and specifically, allowing headmasters and teachers agency in what is taught in the classroom—did not meaningfully change its quality. Both approaches to rolling out Utkarsh were highly effective at delivering the new remedial instruction. Generally, the two different implementation models have similarly high rates of fidelity to the program guidelines, with minimal differences. Part of this result is likely due to the fact that teachers generally adjust lesson timetables and content even without being instructed to do so. While few teachers in the Flexible arm filled out the teaching plan that indicated that they intended to deviate from the recommended Utkarsh schedule, during enumerator observations 43 percent of teachers in both versions of the program were modifying the timetable. These results echo anecdotal evidence that teachers do not just want additional autonomy, but will actively seize it: many teachers will do what they think is best for students even if not directly told to do so, and even if they are officially supposed to be doing something else. Modifications relative to the official schedule are higher when teachers are explicitly given flexibility, with no end difference in student achievement.

There are several lessons from this study that help provide new evidence on how to improve secondary-school education in developing countries. First, many Indian secondary school students are substantially behind grade level. Moreover, our results suggest that teachers are generally aware of these gaps, although there is substantial measurement error

in teacher perceptions of student ability. Second, despite a traditional emphasis on rigid delivery of the curriculum, as well as concerns that teachers may lack adequate skills needed to deviate from prescribed lessons, teachers were able to adapt to offering remedial instruction successfully. Moreover, many teachers adapt their lesson plans and timetable regardless of what is advised, suggesting that teachers generally try and deliver content to students as they see fit. Thus, our results suggest that one reason why service delivery is poor in the public sector is because teachers lack the appropriate materials and direction to allow them to teach at the level of the student.

Third, our study provides guidance on the optimal allocation of authority in the public sector. While offering teachers flexibility did not improve student growth relative to the standard version of Utkarsh, it also did no harm. These results bolster the interpretation that rigid bureaucracies could improve service delivery by modifying rules and potentially explicitly giving teachers increased ability to adapt to local conditions. Despite concerns about coordination challenges in changing the status quo, our results suggest that teachers were able to effectively adapt to a new, more effective approach in the classroom that benefited both students and teachers.

Finally, we shed light on whether remedial instruction is a wise policy choice at the secondary school level and more generally build upon the scant policy base of what interventions are effective at improving student growth at the secondary school level. Despite concerns about crowd-out, or that secondary school level may be too late to introduce effective learning interventions, we find substantial increases in student achievement: students in the Utkarsh program learned 60 percent more than the status quo. Thus, our results suggest that remedial education is a good use of class time for secondary school students; it is a cost-effective way to improve student growth at the secondary school level and decrease the substantial heterogeneity in learning outcomes.

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.

Angrist, Noam and Rachael Meager, "Implementation Matters: Generalizing Treatment Effects in Education. EdWorkingPaper No. 23-802.," *Annenberg Institute for School Reform at Brown University*, 2023.

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.

_ , _ , _ , and _ , "The allocation of authority in organizations: A field experiment with bureaucrats," *The Quarterly Journal of Economics*, 2021, *136* (4), 2195–2242.

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.

Banerjee, Abhijit V, 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.

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.

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.

Chiplunkar, Gaurav, Diva Dhar, and Radhika Nagesh, "Not too Little, but too Late: Improving Post-Primary Learning Outcomes in India," 2023.

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.

Dessein, Wouter, "Authority and communication in organizations," *The Review of Economic Studies*, 2002, *69* (4), 811–838.

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.

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

Duflo, Annie, Jessica Kiessel, and Adrienne M Lucas, "Experimental Evidence on Four Policies to Increase Learning at Scale," *The Economic Journal*, 02 2024, *134* (661), 1985–2008.

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.

Evans, David K and Fei Yuan, "How big are effect sizes in international education studies?," *Educational Evaluation and Policy Analysis*, 2022, *44* (3), 532–540.

Figlio, David N and Marianne E Page, "School choice and the distributional effects of ability tracking: does separation increase inequality?," *Journal of Urban Economics*, 2002, *51* (3), 497–514.

Gailmard, Sean, "Expertise, subversion, and bureaucratic discretion," *Journal of Law, Economics, and Organization*, 2002, *18* (2), 536–555.

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

Gray-Lobe, Guthrie, Anthony Keats, Michael Kremer, Isaac Mbiti, and Owen W Ozier, "Can education be standardized? Evidence from Kenya," *Evidence from Kenya (September 16, 2022). University of Chicago, Becker Friedman Institute for Economics Working Paper*, 2022, (2022-68).

Hares, S, AL Minardi, and J Rossiter, "Grade Repetition in Developing Countries: Repeat to Fail or Second Times a Charm," 2020.

Haushofer, Johannes and Jeremy Shapiro, "The Short-term Impact of Unconditional Cash Transfers to the Poor: ExperimentalEvidence from Kenya*," *The Quarterly Journal of Economics*, 2016, *131* (4), 1973–2042.

Jackson, JohnÂ H. and CyrilÂ P. Morgan, *Organization Theory: A Macro Perspective for Management*, 2nd ed., Englewood Cliffs, NJ: Prentice Hall, 1982.

Jo, Jinhee and Lawrence S Rothenberg, "The importance of bureaucratic hierarchy: conflicting preferences, incomplete control, and policy outcomes," *Economics & Politics*, 2014, *26* (1), 157–183.

Kerwin, Jason, Nada Rostom, and Olivier Sterck, "Striking the Right Balance: Why Standard Balance Tests Over-Reject the Null, and How to Fix it," Technical Report IZA DP No. 17217 August 2024.

Lavy, Victor and Analia Schlosser, "Targeted remedial education for underperforming teenagers: Costs and benefits," *Journal of Labor Economics*, 2005, *23* (4), 839–874.

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.

National Steering Committee for National Curriculum Frameworks, "National Curriculum Framework for School Education," Technical Report, Ministry of Education, Government of India 2023.

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

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.

Rauch, James E and Peter B Evans, "Bureaucratic structure and bureaucratic performance in less developed countries," *Journal of public economics*, 2000, *75* (1), 49–71.

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.

Zhao, Yong, *What works may hurt: Side effects in education*, Teachers College Press, 2018.

Figures

Figure 1: Study Design

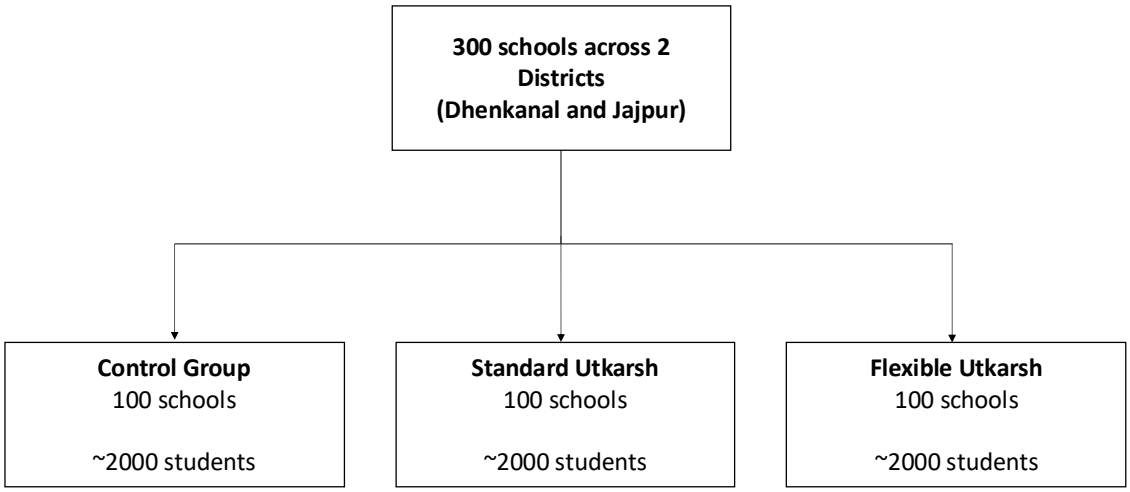


Figure 2: Timeline

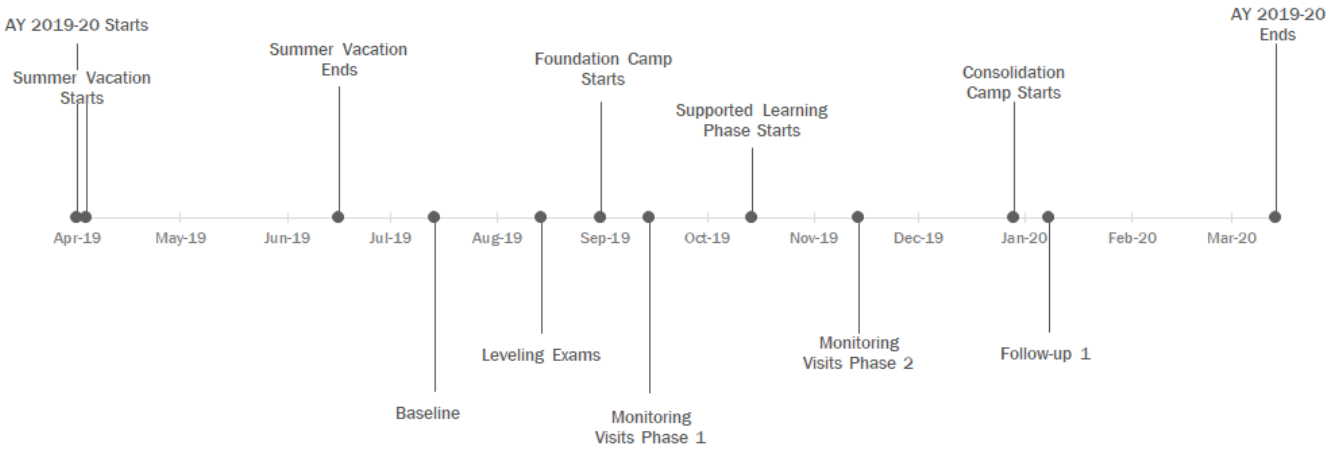
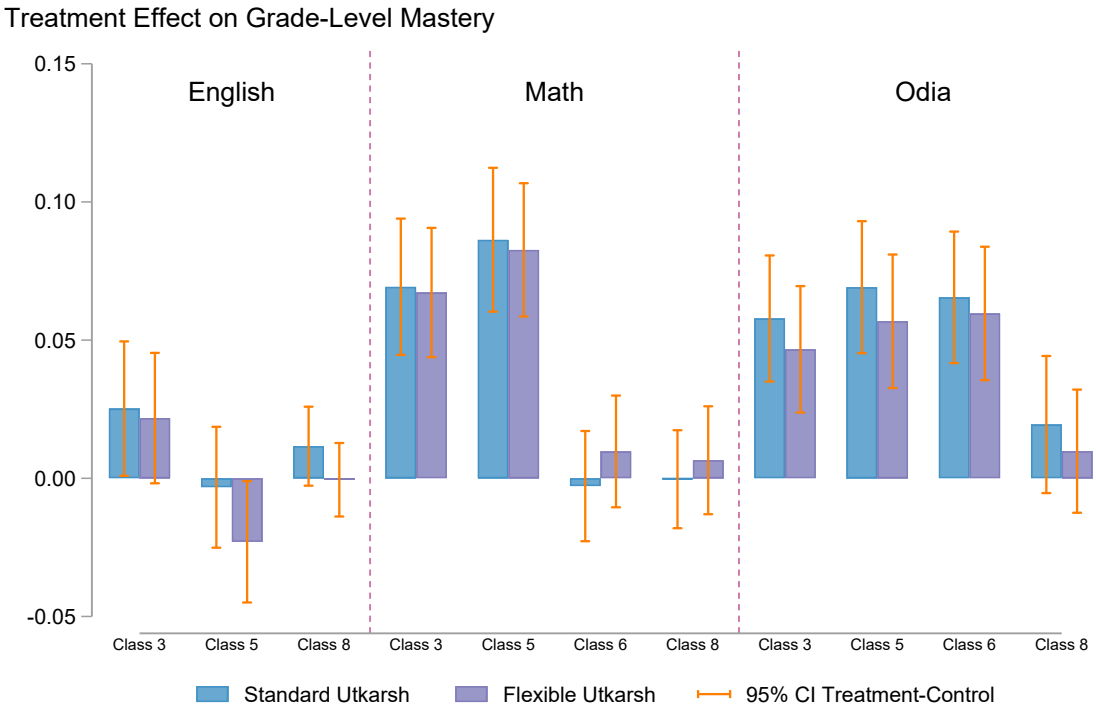
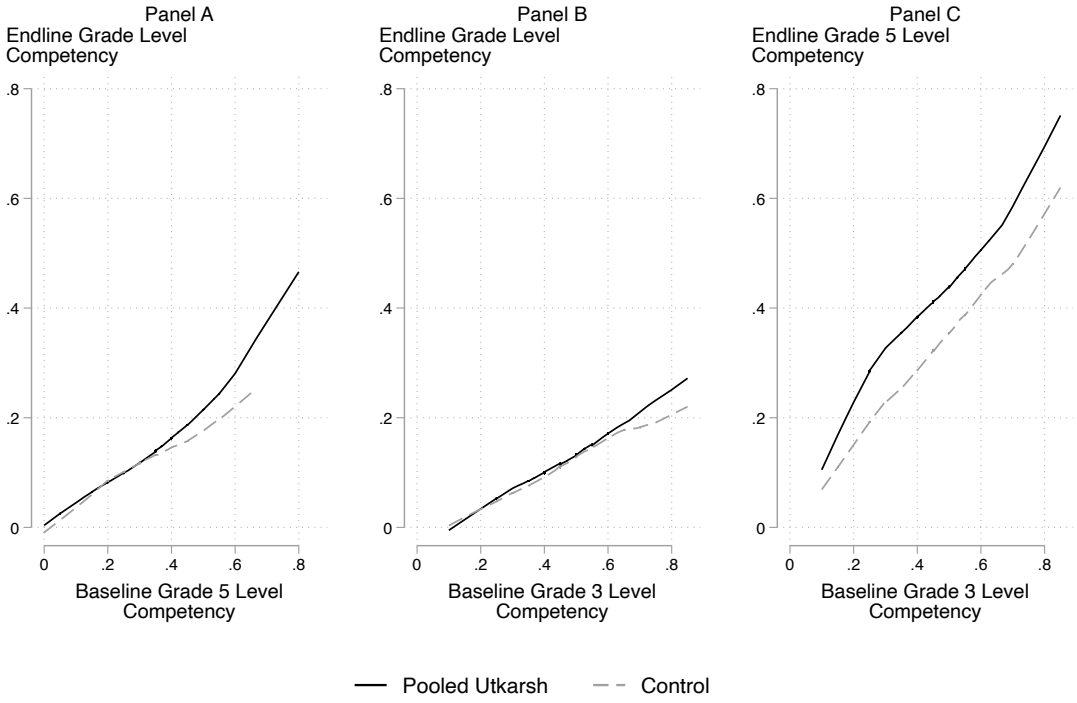


Figure 3: Treatment Effect on Grade-Level Mastery



Notes: This figure shows the treatment effects on achieving at least the respective grade-level mastery in English, math, and Odia relative to the status quo. A student is considered to have achieved a specific grade-level mastery in a subject if they correctly answered at least 75 percent of the questions related to that grade's learning level.

Figure 4: Math Competency by Mean Baseline Competency



Notes: This figure shows the school-level distribution of math competency by mean baseline math competency, separately for the pooled Utkarsh group and the control group. In Panel A, the x-axis represents the school-level share of students who had grade 5 level competency in math at baseline, while the y-axis represents the school-level share of students who achieved grade-appropriate competency in math at endline. In Panel B, the x-axis represents the school-level share of students with grade 3 level competency in math at baseline, while the y-axis remains the same as in Panel A. In Panel C, the x-axis is the same as in Panel B, while the y-axis represents the school-level share of students who had grade 5 level competency in math at endline.

Tables

Table 1: Treatment Effects on Students' Test Scores

	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)	0.80	0.21	0.85	0.89	0.21

Notes: This table reports the treatment effect on students' standardized IRT scores from endline tests. Column 1: overall score based on all subject. Columns 2-5: scores in respective subjects. All regressions include strata, week, and day-of-week fixed effects; student's standardized IRT scores from baseline English, math, and Odia tests, a dummy for student being female, age of student, and indicator variables for endline interview phase. Heteroskedasticity-robust standard errors, clustered at the school level, in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 2: Impacts on Teacher Behavior

	Teacher in the classroom upon arrival	Teaching	Teaching Practices Index	Implementation Portion
	(1)	(2)	(3)	(4)
Standard Utkarsh	0.032	-0.021	0.385**	0.814***
	(0.052)	(0.030)	(0.159)	(0.010)
Flexible Utkarsh	-0.054	-0.046	0.346**	0.809***
	(0.056)	(0.031)	(0.157)	(0.010)
Observations	299	290	290	299
Control mean	0.84	0.98	0.000	0.00
Standard Utkarsh=Flexible Utkarsh (p-value)	0.11	0.45	0.801	0.73

Notes: Outcomes in Columns 1 and 2 are teacher-level indicator variables measured by enumerators while the teacher was supposed to be teaching in the classroom (one teacher was observed in each school). Column 3, also a teacher-level observation, is constructed from the following list of indicator variables observed by enumerators while the teacher was supposed to be teaching in the classroom: teaching and learning materials visible; at least one student had an opportunity to express their own idea; the teacher asked a question to the class; the teacher answered students' questions supportively; the teacher answered students' questions without showing disrespect; the teacher did not ignore students' questions; the teacher seemed familiar with the content; the teacher encouraged students; the teacher responded to questions satisfactorily; and the teacher engaged in student interaction. The index is constructed by calculating the proportion of these variables that are true for each observation. Column 4 presents school-level implementation fidelity of the Utkarsh program, constructed from teacher- and school-level variables. The measure of implementation portion is based on the following list of teacher-reported variables: whether a leveling assessment was conducted; the share of the previous six days that the teacher taught Utkarsh; whether the teacher completed an Utkarsh worksheet on the day of the survey or the most recent day they taught; whether the teacher implemented the correct phase; whether students attending each phase of Utkarsh met the inclusion criteria of the respective phase; the share of all FC lessons that are Utkarsh lessons; whether the teacher followed the Utkarsh lesson exactly as instructed in the lesson guides; whether the teacher taught the planned lesson during FC; the percentage of the previous six days that the teacher followed the planned Utkarsh lessons during FC; and the following three enumerator-observed variables: whether students used handbooks in class, whether the classroom had a word wall, and whether student desks were arranged in small groups. Each underlying variable ranges from 0 to 1 and is set to zero for all control group schools. We first calculate the school-level average of each variable and then take the average across these variables for each school to construct the implementation measure. Columns 1–3 include the teacher's age in years and age squared, the teacher's years of experience and experience squared, a dummy variable for whether the teacher is female, a vector of dummy variables for the main subject taught by the teacher at baseline and indicator variables for the monitoring visit phase. Column 4 includes the average teacher age and its square, the average teacher experience and its square, and the share of female teachers. All regressions include strata, week, and day-of-week fixed effects. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table 3: Teacher Flexibility and Discretion

	Filled out flexible Utkarsh teaching plan	Followed flexible Utkarsh teaching plan for this week	Had autonomy in using Utkarsh	Teacher can adjust content if students have difficulty
	(1)	(2)	(3)	(4)
Standard Utkarsh	0.016	0.013	0.944***	0.007
	(0.018)	(0.015)	(0.013)	(0.030)
Flexible Utkarsh	0.201***	0.150***	0.989***	0.071**
	(0.037)	(0.031)	(0.009)	(0.030)
Observations	569	569	834	834
Control group mean	0.00	0.00	0.00	0.77
Standard Utkarsh=Flexible Utkarsh (p-value)	0.00	0.00	0.00	0.03
Source	Directly Observed	Self-reported	Self-reported	Self-reported

Notes: Column 1 measures whether the teaching plan was filled out. Column 2 is a self-reported measure of whether teachers followed the teaching plan that they filled out for that week. Columns 3 is a self-reported measure of teachers' autonomy in using Utkarsh lessons. Column 4 is a self-reported measure of whether teachers can adjust the content of the lesson. Columns 1 and 2, measured at midline, include indicator variables for the monitoring visit phase. Columns 3 and 4, measured at endline, include an indicator variable for the early endline visit. All columns report indicator variables. All regressions include strata, week, and day-of-week fixed effects; the teacher's age in years and age squared, the teacher's years of experience and experience squared, a dummy for whether the teacher is teacher being female, and a vector of dummy variables for the main subject taught by the teacher at baseline. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table 4: Lesson Topic Relative to Standard Utkarsh

Topic was:	Any Standard Utkarsh lesson	This week's scheduled Standard Utkarsh	Scheduled Standard Utkarsh lesson for this week or an adjacent week
	(1)	(2)	(3)
Flexible Utkarsh	-0.044	0.001	-0.086*
	(0.028)	(0.054)	(0.051)
Observations	289	289	289
Standard Utkarsh mean	0.95	0.43	0.82

Notes: This table presents results for teachers surveyed during the SLP-phase of midline and is restricted to those who taught English, math, and science Utkarsh lessons. Odia teachers are not included in this table because Odia lessons are reported at a level that is too generalized. Column 1 provides a directly observed snapshot of whether or not the teacher followed any scheduled Standard Utkarsh curriculum during that week of SLP. Column 2 measures whether or not the teacher followed that week's Standard Utkarsh curriculum during SLP. Column 3 reports whether the teacher followed that week's or an adjacent week's Standard Utkarsh curriculum. All columns report indicator variables. All regressions include strata, week, and day-of-week fixed effects, the teacher's age in years and age squared, the teacher's years of experience and experience squared, a dummy variable for whether the teacher is female; a vector of dummy variables for the main subject taught by the teacher at baseline, and indicator variables for the monitoring visit phase. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table 5: Teachers' Perceptions of Students and Utkarsh

	Students benefitted from Utkarsh	Teacher benefitted from Utkarsh	*Percent of students who can write a simple English sentence*			*Percent of students who can do a three digit minus two digit substraction*			Teacher Forecasts that ....Percent of Student	
			Teacher estimate	Actual	Teacher estimate-Actual	Teacher estimate	Actual	Teacher estimate-Actual	Will pass the board exam	Will eventually complete a bachelor's degree
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Standard Utkarsh	0.749***	0.868***	-5.679***	3.510**	-9.190***	-5.303***	0.719	-6.023***	-4.463**	-6.297***
	(0.031)	(0.019)	(2.152)	(1.598)	(2.416)	(1.974)	(1.906)	(2.273)	(1.727)	(2.001)
Flexible Utkarsh	0.717***	0.919***	-5.078***	6.346***	-11.424***	-6.259***	1.438	-7.697***	-5.766***	-7.787***
	(0.030)	(0.020)	(1.912)	(1.617)	(2.283)	(1.706)	(2.016)	(2.386)	(1.633)	(1.823)
Observations	834	834	823	823	823	823	823	823	823	823
Control group mean	0.00	0.00	56.76	16.97	39.79	73.35	56.39	16.96	60.64	49.63
Standard Utkarsh=Flexible Utkarsh (p-value)	0.43	0.04	0.76	0.08	0.34	0.61	0.71	0.48	0.46	0.44

Notes: This table presents the teacher's perceptions of their students' abilities and the benefits of Utkarsh. Columns 1-3, 5-6, and 8-10 are reported by teachers at endline, while Columns 4 and 7 represent actual school-level averages based on endline assessments. Columns 1-2 are indicator variables. The possible values for Columns 3–10 range from 0 to 100. All regressions include the teacher's age in years and age squared, years of experience and experience squared, a dummy variable for whether the teacher is female, a vector of dummy variables for the main subject taught at baseline, and indicator variables for the early endline visit. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

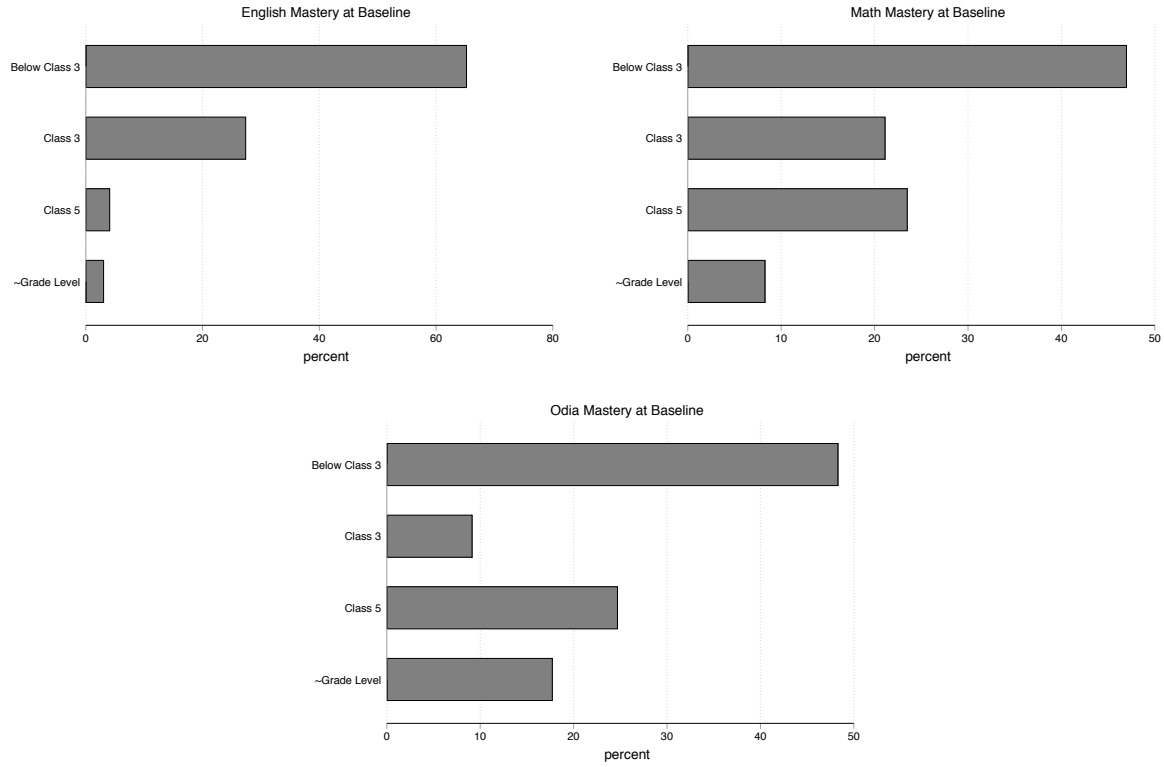
Table 6: Effects on Board Exam Marks and Status After Board Exam

	Passed	*Test Score*			*Grade*	*Status After Board Exam*		
		Total	Utkarsh total	Non-Utkarsh total	B or above	Ernolled in school	Enrolled in class 11	Employed
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Standard Utkarsh	-0.003	-0.157***	-0.134***	-0.192***	-0.056***	0.018	0.013	0.000
	(0.003)	(0.048)	(0.047)	(0.054)	(0.019)	(0.020)	(0.020)	(0.012)
Flexible Utkarsh	-0.004	-0.081	-0.081*	-0.076	-0.021	-0.009	-0.020	-0.004
	(0.003)	(0.049)	(0.048)	(0.056)	(0.020)	(0.020)	(0.021)	(0.011)
Observations	18,551	18,551	18,551	18,551	18,551	1,255	1,255	1,255
Control group mean	0.99	0.00	0.00	0.00	0.39	0.89	0.88	0.04
Standard Utkarsh=Flexible Utkarsh (p-value)	0.88	0.14	0.30	0.04	0.11	0.15	0.11	0.66
Raw control mean		0.56	0.55	0.57				
Raw control SD		0.14	0.15	0.15				

Notes: This table presents the effects of Utkarsh on board exam marks and students' enrollment status after the board exam. Columns 1-5 report outcomes based on administrative data from board exam results. Column 1 indicates whether the student passed the exam. Columns 2-4 provide standardized total scores in all subjects, Utkarsh subjects, and non-Utkarsh subjects, respectively. Colum 5 indicates whether the student achieved a grade of B or above. Columns 6-8 report data from a follow-up survey conducted after the board exam results were published. Column 6 measures whether the student is enrolled in school, while Column 7 measures whether the student is enrolled in class 11. Column 8 measures whether the student is employed. All regressions include strata fixed effects; standardized IRT scores from baseline English, math, and Odia tests, a dummy variable for whether the student is female, and the age of the student. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

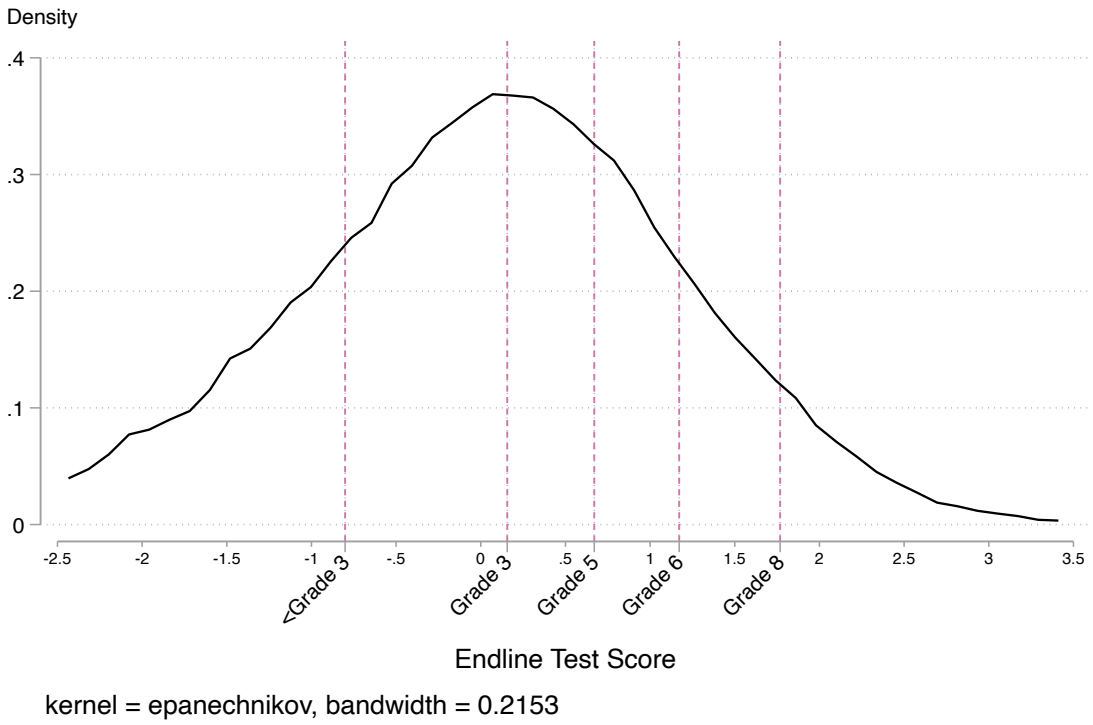
A Additional Figures

Figure A1: Baseline measures of Student Ability, by Grade Level Competency



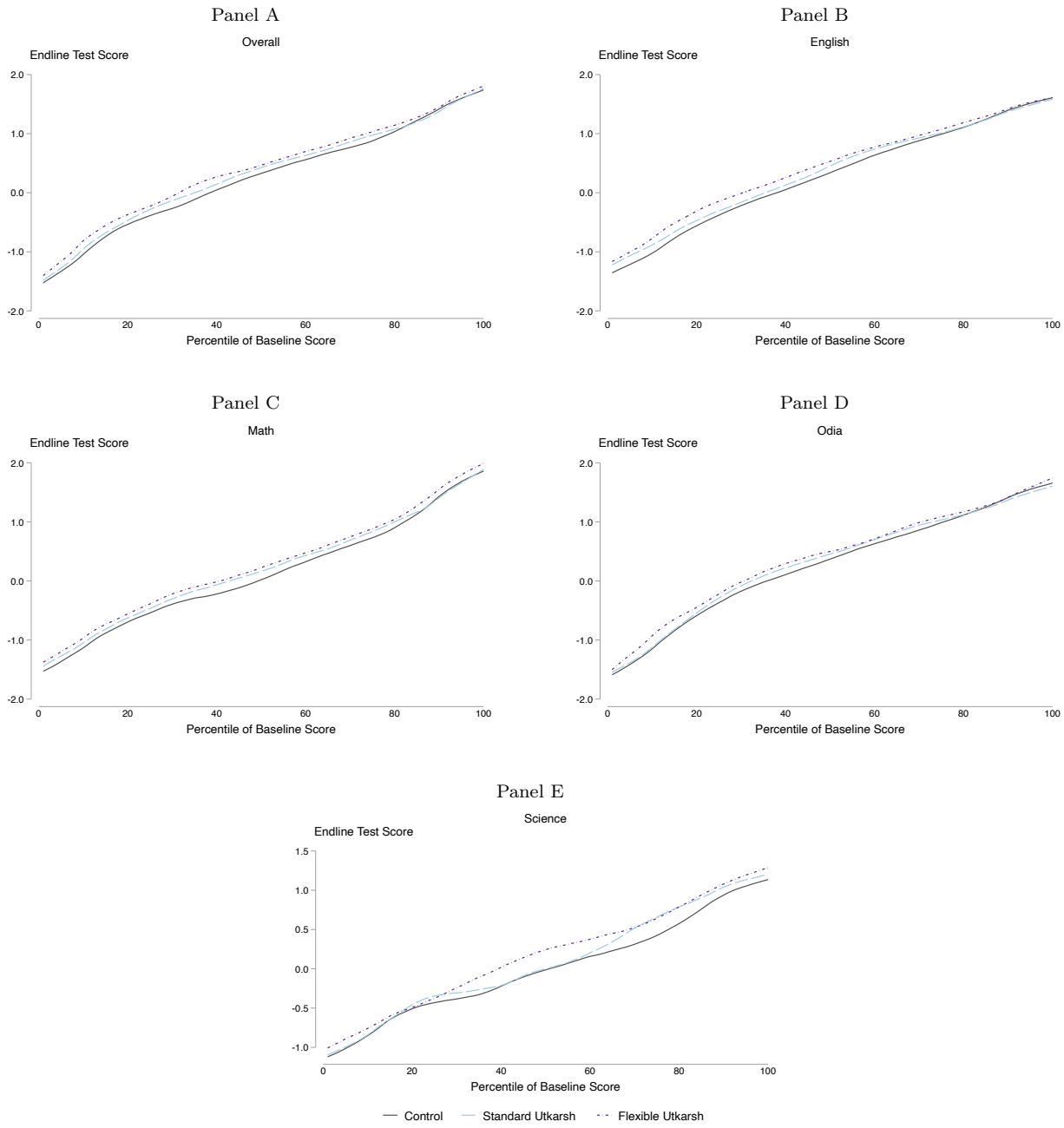
Notes: The figure shows the unweighted local mean results of the students, categorized by their learning level as described in the text. Results are presented by English, math, and Odia (local language). The science test was not administered at baseline.

Figure A2: Control Group Math Learning



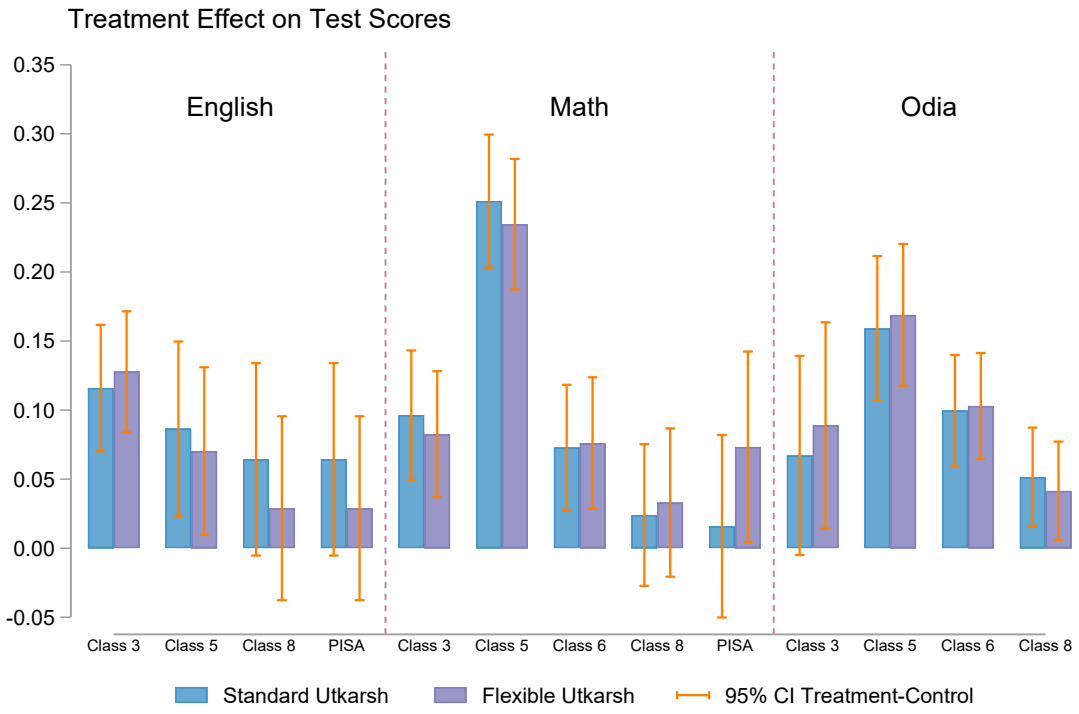
Notes: This figure shows the distribution of math learning among control group students. The vertical lines represent learning levels corresponding to their respective grade-level learning.

Figure A3: Non-parametric Distribution of Test Scores by Study Arms



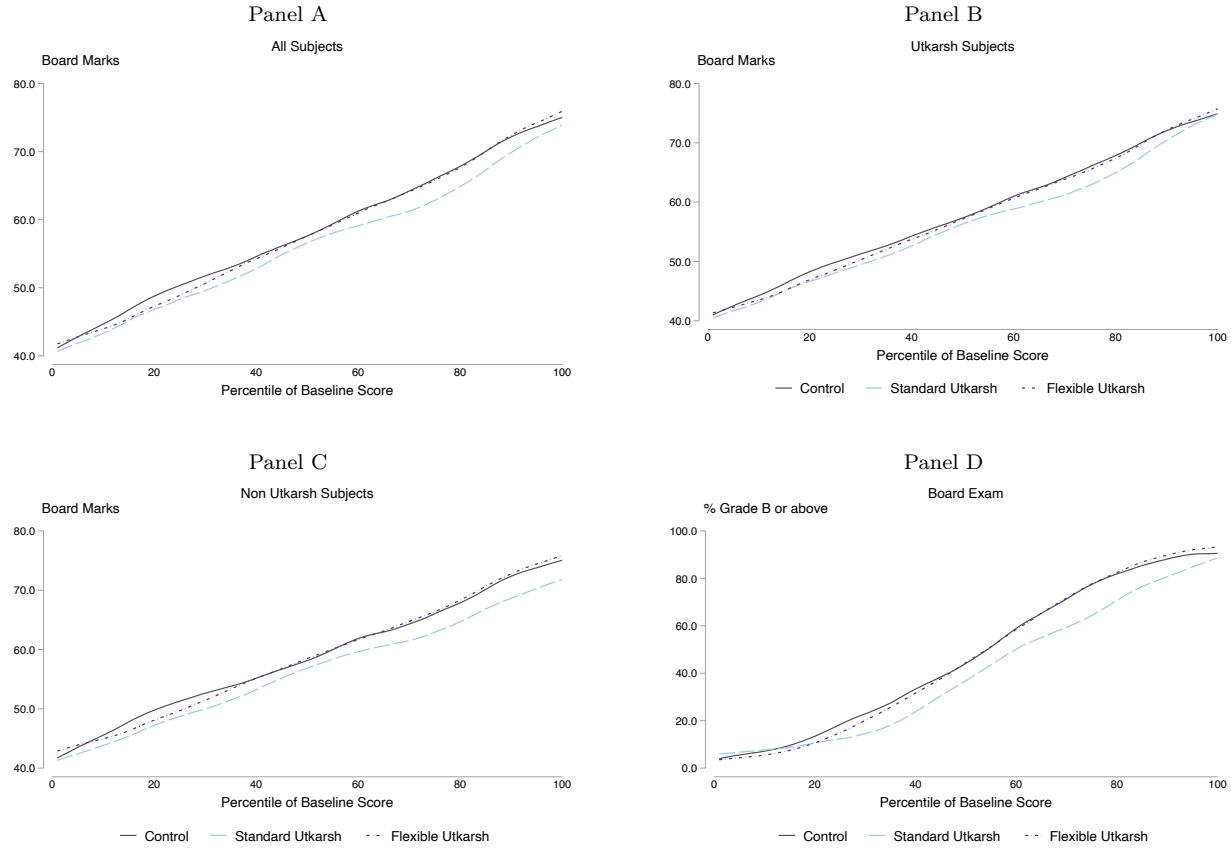
Notes: The figure shows the kernel-weighted local mean smoothed distributions of endline test scores over baseline test score percentiles by study arms. In Panel A, we use the overall baseline test score, while in Panels B-D, we use the baseline test scores for the respective subjects. Since there was no science test at baseline, we use the average of all subjects at baseline as a proxy for the science's baseline in Panel E. Test scores of standard Utkarsh students at low- and mid-baseline-score-percentiles are higher than those of the control group in all subjects. Flexible Utkarsh students' endline scores in all subjects are higher than those of the control group across the entire distribution of baseline test scores.

Figure A4: Subject-wise Test Scores in Grade-Level Questions



Notes: This figure shows the treatment effects on test scores for grade-level questions in English, math, and Odia.

Figure A5: Non-parametric Distribution of Board Marks by Study Arms



Notes: The figure shows kernel-weighted local mean smoothed distributions of board exam scores over baseline test score percentiles by study arm. We show board marks for all subjects, Utkarsh subjects, and non-Utkarsh subjects in Panels A, B, and C, respectively. Panel D shows the probability of obtaining an overall grade of B or higher in the board exam.

B Additional Tables

Table B1: 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)	0.50	0.49	0.51	0.61
Age (in years)	13.16	13.15	13.15	0.88
Scheduled caste, scheduled tribe, or other backward caste (=1)	0.60	0.59	0.64	0.06*
No parent can read and write (=1)	0.14	0.16	0.15	0.64
Participated in Utthan (=1)	0.88	0.89	0.88	0.40
Takes private tuition (=1)	0.73	0.71	0.73	0.71
Baseline test scores				
English	0.01	-0.05	0.05	0.09*
Math	-0.01	-0.02	0.03	0.62
Odia	-0.01	-0.02	0.03	0.52
Baseline competency grade				
English	2.65	2.53	2.58	0.04**
Math	3.45	3.37	3.45	0.49
Odia	3.93	3.84	3.93	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)	52.45	52.44	52.44	1.00
Experience in current position (years)	5.53	5.83	4.70	0.43
Headmaster thinks/considers				
Teacher should follow curriculum (=1)	0.74	0.76	0.80	0.56
Ensuring adherence to curriculum as important job component (=1)	0.95	0.98	0.96	0.45
They have influence on determining delivering curriculum lessons (=1)	0.84	0.79	0.77	0.41
Head of school is the headmaster or headmaster-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: This table shows the reported characteristics of the respondents from the baseline survey. Panel A shows the balance of student-level variables. Panel B reports the balance of headmaster- and school-level variables. We include strata fixed effects, and standard errors are clustered at the school level. ***, **, and * indicate significance at the 1, 5, and 10 percent levels.

Table B2: Balance Table (Teacher)

Panel A: Teachers Surveyed During Baseline	Control (1)	Standard Utkarsh (2)	Flexible Utkarsh (3)	p-value from test of equality (1)=(2)=(3) (4)
Female (=1)	0.48	0.47	0.49	0.96
Age of the teacher (in years)	41.39	41.71	40.87	0.79
Has a teaching certificate (=1)	0.35	0.35	0.35	0.95
Teaching experience (years)	15.77	16.64	15.16	0.48
Teaches an Utkarsh subject (=1)	1.00	1.00	1.00	0.42
Days absent from work	0.33	0.38	0.33	0.73
Works in another school(=1)	0.02	0.02	0.01	0.73
Time spent in preparing for lesson (hours/week)	12.13	12.72	12.78	0.68
Time spent grading (hours/week)	8.46	8.03	8.40	0.64
According to the teacher, percent of student who				
Will pass board exam in first try	63.90	61.86	65.19	0.50
Can write a simple English sentence	55.72	54.79	57.58	0.48
Can do a three digits sum	71.08	73.79	73.02	0.27
Select teaching materials, methods, strategies (=1)	0.95	0.96	0.93	0.38
Allowed to modify course timetable (=1)	0.81	0.80	0.77	0.47
Others do not select evaluation activities (=1)	0.67	0.66	0.65	0.96
Feel pressure to complete curriculum during school year (=1)	0.29	0.28	0.30	0.90
Burnout index	-0.05	-0.15	-0.10	0.64
Autonomy index	0.00	-0.20	-0.16	0.10*
Teacher feels				
Curriculum should be followed even if students have lower learning level (=1)	0.60	0.55	0.62	0.31
That if students' not being ready for board exam would be teacher's own fault (=1)	0.48	0.52	0.47	0.70
Valued and appreciated (=1)	0.69	0.70	0.76	0.09*
Satisfied with job (=1)	0.85	0.82	0.84	0.81
That their opinion seems to count (=1)	0.89	0.92	0.91	0.59
That they have the materials and equipment to teach effectively (=1)	0.63	0.60	0.71	0.08*
Similarly or more effective compared to colleagues (=1)	0.94	0.95	0.96	0.51
Observations	209	189	207	
Panel B: Teachers Surveyed During Monitoring Visit	(1)	(2)	(3)	(4)
Female (=1)	0.47	0.42	0.47	0.55
Age of the teacher (in years)	41.83	42.25	41.60	0.95
Has a teaching certificate (=1)	0.35	0.35	0.35	0.94
Teaching experience (years)	15.77	16.64	15.16	0.48
Teaches an Utkarsh subject (=1)	1.00	1.00	1.00	0.49
Observations	309	306	308	
Panel C: Teachers Surveyed During Endline	(1)	(2)	(3)	(4)
Female (=1)	0.47	0.43	0.47	0.71
Age of the teacher (in years)	41.90	41.99	41.68	0.99
Has a teaching certificate (=1)	0.36	0.35	0.36	0.93
Teaching experience (years)	15.89	16.58	15.25	0.49
Teaches an Utkarsh subject (=1)	1.00	1.00	1.00	N/A
Observations	291	290	289	

Notes: This table shows the reported baseline characteristics of the analysis sample teachers. Panel A is restricted to teachers who were surveyed during the baseline. Panel B reports the characteristics of teachers who were surveyed during the midline survey. Panel C is restricted to teachers who were surveyed during the endline. If any variable is missing from the baseline, we replace it with data from the midline or endline survey. We include strata fixed effects, and standard errors are clustered at the school level. ***, **, and * indicate significance at the 1, 5, and 10 percent levels.

Table B3: Percent of Student in Each Level

Student Level	*Baseline*							
	English		Math		Odia		Science	
	Standard	Flexible	Standard	Flexible	Standard	Flexible	Standard	Flexible
Inception	30.11	31.86	42.45	43.63	27.62	32.46	50.17	50.78
Class 3	24.39	22.62	36.35	37.69	6.52	7.67	26.84	27.05
Class 5	25.51	28.44	6.90	6.70	18.31	18.51	10.87	10.93
Class 8	10.84	8.40	5.12	3.27	36.42	31.39	3.88	2.47
Absent	9.15	8.68	9.18	8.71	11.13	9.97	8.23	8.77
Total Number of Students	7,287	6,569	7,287	6,569	7,287	6,569	7,287	6,569
Student Level	*Endline*							
	English		Math		Odia		Science	
	Standard	Flexible	Standard	Flexible	Standard	Flexible	Standard	Flexible
Inception	15.61	18.50	23.41	23.31	13.69	17.05	37.63	36.76
Class 3	19.25	20.76	34.42	36.40	5.74	5.82	18.07	17.66
Class 5	25.71	25.74	10.83	10.60	12.10	12.28	14.81	16.47
Class 8	24.67	20.76	16.02	14.01	51.67	48.58	14.55	14.98
Absent	14.76	14.23	15.32	15.69	16.80	16.27	14.94	14.13
Total Number of Students	7,284	6,569	7,284	6,569	7,284	6,569	7,284	6,569

Notes: The table shows the share of students at each subject-specific learning level in the baseline and endline of the test administered by the program. (*not the test that the researchers developed*).

Table B4: Treatment Effects on Students' PISA Test Scores

	Overall	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: This table reports the impact on students' standardized IRT scores for PISA questions in the endline tests. PISA questions were only included in the English and math tests. Column 1: overall score based on English and math PISA questions. Columns 2-3: scores for the respective subjects' PISA questions. All regressions include strata, week, and day-of-week fixed effects; standardized IRT scores from baseline English, math, and Odia tests, a dummy variable for whether the student is female, the age of the student, and indicator variables for the endline interview phase. Heteroskedasticity-robust standard errors, clustered at the school level, in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B5: Multiple Hypothesis Test - Treatment Effects on Test Scores

	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
Standard Utkarsh=Control (naive p-value)	0.000	0.000	0.000	0.000	0.001
Standard Utkarsh=Control (adjusted q-value)	0.000	0.000	0.000	0.000	0.001
Flexible Utkarsh=Control (naive p-value)	0.000	0.000	0.000	0.000	0.000
Flexible Utkarsh=Control (adjusted q–value)	0.000	0.000	0.000	0.000	0.000
Flexible Utkarsh=Standard Utkarsh (naive p-value)	0.805	0.904	0.851	0.885	0.210
Flexible Utkarsh=Standard Utkarsh (adjusted q–value)	0.997	0.997	0.997	0.997	0.615

Notes: This table shows adjusted q-values using the Haushofer and Shapiro (2016) implementation of the Anderson (2008) Family-Wise Error Rate (FWER) adjustment for our primary outcome of interest — test scores. Column 1: overall score based on all subjects. Columns 2-5: scores for the respective subjects. All regressions include strata, week, and day-of-week fixed effects; students' standardized IRT scores from baseline English, math, and Odia tests, a dummy for whether the student is female, age of the student, and indicator variables for the endline interview phase. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table B6: Treatment Effect Heterogeneity by Student Gender

	Overall	English	Math	Odia	Science
	(1)	(2)	(3)	(4)	(5)
Standard Utkarsh	0.090***	0.091***	0.133***	0.064***	0.095**
	(0.019)	(0.024)	(0.028)	(0.022)	(0.039)
Standard Utkarsh x Female	0.033	0.055**	-0.026	0.050*	0.018
	(0.021)	(0.028)	(0.033)	(0.029)	(0.050)
Flexible Utkarsh	0.105***	0.114***	0.145***	0.061***	0.126***
	(0.019)	(0.023)	(0.028)	(0.021)	(0.042)
Flexible Utkarsh x Female	0.011	0.006	-0.043	0.050*	0.033
	(0.022)	(0.031)	(0.035)	(0.026)	(0.052)
Female	-0.017	-0.031	-0.086***	0.071***	-0.089**
	(0.015)	(0.022)	(0.023)	(0.020)	(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+Standard Utkarsh x Female=0 (p-value)	0.00	0.00	0.00	0.00	0.00
Flexible Utkarsh+Flexible Utkarsh x Female=0 (p-value)	0.00	0.00	0.00	0.00	0.00

Notes: This table shows the treatment effect heterogeneity on endline test scores by student gender. Column 1 shows overall standardized IRT scores based on endline test scores from all subjects. Columns 2-5: standardized IRT scores for the respective subjects. All regressions include strata, week, and day-of-week fixed effects, students' standardized IRT scores from baseline English, math, and Odia tests, a dummy for whether the student is female, age of the student, and indicator variables for the endline interview phase. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B7: Treatment Effect Heterogeneity by Student Caste

	Overall	English	Math	Odia	Science
	(1)	(2)	(3)	(4)	(5)
Standard Utkarsh	0.100***	0.105***	0.122***	0.075***	0.131***
	(0.019)	(0.026)	(0.026)	(0.024)	(0.038)
Standard Utkarsh x SC/ST/OBC	0.012	0.024	-0.005	0.025	-0.050
	(0.021)	(0.030)	(0.030)	(0.028)	(0.048)
Flexible Utkarsh	0.100***	0.085***	0.142***	0.076***	0.122***
	(0.019)	(0.025)	(0.028)	(0.023)	(0.042)
Flexible Utkarsh x SC/ST/OBC	0.017	0.056*	-0.031	0.017	0.033
	(0.022)	(0.029)	(0.033)	(0.028)	(0.051)
SC/ST/OBC (=1)	-0.014	-0.035	-0.015	-0.002	0.015
	(0.015)	(0.021)	(0.023)	(0.019)	(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+Standard Utkarsh x SC/ST/OBC=0 (p-value)	0.00	0.00	0.00	0.00	0.04
Flexible Utkarsh+Flexible Utkarsh x SC/ST/OBC=0 (p-value)	0.00	0.00	0.00	0.00	0.00

Notes: This table shows the treatment effect heterogeneity on endline test score by student caste. Column 1 shows overall standardized IRT scores based on endline test scores from all subjects. Columns 2-5: standardized IRT scores for the respective subjects. All regressions include strata, week, and day-of-week fixed effects; students' standardized IRT scores from baseline English, math, and Odia tests, a dummy for students belonging to ST/SC/OBC, a dummy for whether the student is female, age of the student, and indicator variables for the endline interview phase. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B8: Treatment Effect Heterogeneity by Student's First Generation Learner Status

	Overall	English	Math	Odia	Science
	(1)	(2)	(3)	(4)	(5)
Standard Utkarsh	0.109***	0.123***	0.117***	0.085***	0.116***
	(0.015)	(0.020)	(0.020)	(0.018)	(0.033)
Standard Utkarsh x First Generation Learner	-0.013	-0.031	0.019	0.029	-0.088
	(0.037)	(0.047)	(0.048)	(0.042)	(0.075)
Flexible Utkarsh	0.111***	0.119***	0.122***	0.076***	0.163***
	(0.015)	(0.018)	(0.021)	(0.017)	(0.034)
Standard Utkarsh x First Generation Learner	0.001	-0.015	0.011	0.076*	-0.146*
	(0.034)	(0.044)	(0.051)	(0.041)	(0.077)
First Generation Learner (=1)	-0.046*	-0.054	-0.068**	-0.090***	0.104*
	(0.025)	(0.034)	(0.035)	(0.031)	(0.057)
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 + Standard Utkarsh x First Generation Learner=0 (p-value)	0.01	0.05	0.01	0.01	0.69
Flexible Utkarsh+Flexible Utkarsh x First Generation Learner=0 (p-value)	0.00	0.02	0.01	0.00	0.81

Notes: This table shows the treatment effect heterogeneity on endline test scores by student's first generation learning status. Column 1 shows overall standardized IRT scores based on endline test scores from all subjects. Columns 2-5: standardized IRT scores for the respective subjects. All regressions include strata, week, and day-of-week fixed effects; students' standardized IRT scores from baseline English, math, and Odia tests, a dummy variable for students being first-generation learner, a dummy for whether the student is female, age of the student, and indicator variables for the endline interview phase. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B9: Treatment Effect Heterogeneity by baseline Test Scores

	Overall	English	Math	Odia	Science
	(1)	(2)	(3)	(4)	(5)
Standard Utkarsh	0.125***	0.166***	0.111***	0.100***	0.090*
	(0.028)	(0.033)	(0.030)	(0.032)	(0.054)
Standard Utkarsh x BL Test Score Middle Third	-0.021	-0.042	0.055	0.001	-0.063
	(0.028)	(0.038)	(0.036)	(0.036)	(0.064)
Standard Utkarsh x BL Test Score Top Third	-0.037	-0.109***	-0.033	-0.044	0.104
	(0.030)	(0.037)	(0.040)	(0.040)	(0.071)
Flexible Utkarsh	0.129***	0.178***	0.126***	0.122***	0.060
	(0.027)	(0.032)	(0.035)	(0.031)	(0.051)
Flexible Utkarsh x BL Test Score Middle Third	-0.030	-0.063*	-0.016	-0.051	0.127**
	(0.027)	(0.037)	(0.038)	(0.038)	(0.062)
Flexible Utkarsh x BL Test Score Top Third	-0.028	-0.122***	0.016	-0.056	0.116*
	(0.031)	(0.035)	(0.044)	(0.038)	(0.064)
BL Test Score Middle Third	0.082***	0.170***	-0.082***	0.155***	-0.090*
	(0.025)	(0.031)	(0.032)	(0.031)	(0.051)
BL Test Score Top Third	0.087**	0.238***	0.048	0.109***	-0.017
	(0.036)	(0.039)	(0.041)	(0.042)	(0.069)
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+Standard Standard x Middle Third=0 (p-value)	0.00	0.00	0.00	0.00	0.53
Standard Utkarsh+Standard Standard x Top Third=0 (p-value)	0.00	0.01	0.01	0.03	0.00
Flexible Utkarsh+Flexible Standard x Middle Third=0 (p-value)	0.00	0.00	0.00	0.00	0.00
Flexible Utkarsh+Flexible Utkarsh x Top Third=0 (p-value)	0.00	0.01	0.00	0.01	0.00

Notes: This table shows the treatment effect heterogeneity on endline test scores by baseline test scores. Column 1 shows overall standardized IRT scores based on endline test scores from all subjects. Column 1 includes an indicator for the tercile of the overall baseline test score and an interaction of treatment status with that indicator. Since there is no science baseline, the overall baseline test score is calculated based on baseline English, math, and Odia tests. Columns 2-5 show standardized IRT scores for the respective subjects. Columns 2-4 include an indicator for the tercile of the respective subject's baseline test score and an interaction of the treatment status with that indicator. Column 5 includes an indicator for the tercile of the overall baseline test score and an interaction of the treatment status with that indicator. All regressions include strata, week, and day-of-week fixed effects; students' standardized IRT scores from baseline English, math, and Odia tests, a dummy for whether the student is female, age of the student, and indicator variables for the endline interview phase. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B10: Attrition by Treatment Arms

	Student's test scores collected at endline (1)	Student's test scores collected at endline (2)
Standard Utkarsh	0.02**	0.02**
	(0.01)	(0.01)
Flexible Utkarsh	0.01	0.01
	(0.01)	(0.01)
Standard Utkarsh x Overall baseline test score		-0.02*
		(0.01)
Flexible Utkarsh x Overall baseline test score		-0.00
		(0.01)
Overall baseline test score		0.07
		(0.06)
Observations	5,756	5,756
Control Mean	0.94	0.94
Standard Utkarsh=Flexible Utkarsh (p-value)	0.27	0.28
Adjusted R-squared	0.11	0.11

Notes: In all columns, the outcome variable is equal to 1 if all four of a student's test scores are collected during the endline survey. All regressions include strata fixed effects, standardized IRT scores from baseline English, math, and Odia tests, a dummy for whether the student is female, and the age of the student. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B11: 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***	0.106***	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***	0.167***
	(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
Control-group change (baseline to endline)	0.19	0.19	0.21	0.21	0.06	0.06	0.21	0.21	N/A	N/A
Standard Utkarsh = Flexible Utkarsh (p-value)	0.87	0.98	1.00	0.49	0.74	0.90	0.87	0.60	0.09	0.28

Notes: This table shows Lee bounds (Lee, 2009). Columns 1-2: overall score based on all subject. Columns 3-10: scores for the respective subjects. Odd columns show the lower bound, while even columns show the upper bound. All regressions include strata, week, and day-of-week fixed effects; students' standardized IRT scores from baseline English, math, and Odia tests, a dummy for whether the student is female, the age of the student, and indicator variables for the endline interview phase. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table B12: Student's Self-Assessment and Expected Educational Outcomes

	Student Attendance	Self-assessed Rank among Peers (1-10)	Expected Class 10 Board Exam Score (Out of 100)					Hopes to Achieve Bachelor's degree or above
			Overall	English	Math	Odia	Science	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Standard Utkarsh	0.017	-0.067	0.496	0.093	0.314	0.576	0.957*	0.030*
	(0.022)	(0.047)	(0.469)	(0.567)	(0.592)	(0.553)	(0.537)	(0.017)
Flexible Utkarsh	0.009	-0.050	0.701	0.525	0.757	0.710	0.806	0.007
	(0.022)	(0.055)	(0.459)	(0.582)	(0.576)	(0.521)	(0.522)	(0.017)
Observations	5,710	5,397	5,397	5,397	5,397	5,397	5,397	5,397
Control mean	0.66	4.06	66.20	62.67	66.77	70.48	64.86	0.51
Standard Utkarsh=Flexible Utkarsh (p-value)	0.70	0.76	0.69	0.47	0.46	0.82	0.80	0.17

Notes: Column 1 measures whether the student was present at school during the monitoring visit. Column 2 measures the self-assessed rank among peers, where 1 indicates the best rank and 10 indicates the worst student. Outcome variables in Columns 3-7: self-reported expected scores (out of 100) on the board exam at the end of class 10. Column 8: whether the student hopes to achieve a Bachelor's degree or higher. Columns 2-8 are measured during the endline survey and include outcome variables measured at baseline. All regressions include strata, survey week, and survey day-of-week fixed effects; standardized IRT scores from baseline English, math, and Odia tests, a dummy for whether the student is female, and the age of the student. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B13: Components of Teaching Practices Index

	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
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Standard Utkarsh	0.012	0.181**	-0.032	0.075	0.008	0.006	-0.009	0.162**	0.130*	0.146**
	(0.012)	(0.073)	(0.049)	(0.076)	(0.027)	(0.028)	(0.029)	(0.075)	(0.078)	(0.071)
Flexible Utkarsh	0.013	0.186**	-0.043	0.051	0.004	-0.005	-0.037	0.149**	0.159**	0.134*
	(0.013)	(0.078)	(0.045)	(0.077)	(0.028)	(0.026)	(0.031)	(0.075)	(0.076)	(0.073)
Observations	290	290	290	290	290	290	290	290	290	290
Control mean	0.99	0.49	0.89	0.53	0.96	0.97	0.98	0.29	0.46	0.57
Standard Utkarsh=Flexible Utkarsh (p-value)	0.91	0.94	0.81	0.76	0.86	0.61	0.40	0.87	0.70	0.85

Notes: All outcomes are teacher-level indicator variables measured during the midline and observed by enumerators while the teacher was supposed to be teaching in the classroom (one teacher observed in each school). All regressions include the teacher's age in years and age squared, the teacher's years of experience and experience squared, a dummy for whether the teacher is female, a vector of dummy variables for the main subject taught by the teacher at baseline, and indicator variables for the monitoring visit phase. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B14: Implementation of Utkarsh

	Outcomes	Standard Utkarsh	Flexible Utkarsh	N	Standard Utkarsh= Flexible Utkarsh (p-value)
		(1)	(2)	(3)	(4)
(1)	Conducted levelling assessment	0.945***	0.956***	299	0.60
		(0.015)	(0.015)		
(2)	Percent of previous 6 days that teacher taught Utkarsh	0.826***	0.844***	298	0.28
		(0.013)	(0.013)		
(3)	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)		
(4)	Implementing the correct phase	0.940***	0.965***	299	0.17
		(0.016)	(0.013)		
(5)	Students who attended FC met the inclusion criteria	0.833***	0.853***	298	0.57
		(0.028)	(0.028)		
(6)	Students who attended SLP met the inclusion criteria	0.804***	0.796***	233	0.87
		(0.039)	(0.036)		
(7)	Students who attended CC met the inclusion criteria	0.781***	0.772***	288	0.83
		(0.032)	(0.033)		
(8)	Share of all FC lessons that are Utkarsh lesson	0.851***	0.849***	298	0.93
		(0.020)	(0.023)		
(9)	Followed Utkarsh lesson exactly as instructed in the lesson guides	0.980***	0.983***	294	0.81
		(0.010)	(0.010)		
(10)	Students currently using handbooks in class (directly observed)	0.980***	0.913***	244	0.08
		(0.025)	(0.032)		
(11)	Classroom has a word wall (directly observed)	0.093**	0.123***	299	0.55
		(0.043)	(0.042)		

...(continue in next page)...

Table B14: Implementation of Utkarsh

	Outcomes	Standard Utkarsh	Flexible Utkarsh	N	Standard Utkarsh= Flexible Utkarsh (p-value)
		(1)	(2)	(3)	(4)
	...(continue from previous page)...				
(12)	Student desks arranged in small group (directly observed)	0.798***	0.781***	249	0.77
		(0.051)	(0.051)		
(13)	Teaching planned lesson during FC	0.862***	0.900***	84	0.50
		(0.058)	(0.047)		
(14)	Percent of previous 6 days that teacher followed planned Utkarsh during FC	0.714***	0.796***	84	0.00
		(0.042)	(0.029)		

Notes: This table shows school-level implementation fidelity of the Utkarsh program. All outcomes are set to zero for all control group schools and range between 0 and 1 for treatment schools. Each regression includes the following school-level variables: average teacher age and its square, average teacher experience and its square, and the share of female teachers. Rows 1-4: teacher-reported responses measured during the spot visit. Rows 5-7: Teacher-reported measures of whether students attending each phase met the inclusion criteria for that phase. These outcomes do not consider any inclusion errors. Rows 5-6: measured during spot visit. Row 6 is measured only if the spot visit took place during the SLP phase. Row 7: measured during the endline. Row 8: teacher-reported responses measured during the spot visit. Row 9: measured during the endline. Rows 10-12: enumerator-observed outcomes measured during the spot visit. Rows 13-14: teacher-reported responses measured during the spot visit and measured at all schools if the spot visit took place during the FC phase. Therefore, the sample sizes for these outcomes are smaller than for other variables. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B15: Headmaster Attendance

	Headmaster in school during monitoring visit
	(1)
Standard Utkarsh	0.083
	(0.055)
Flexible Utkarsh	0.066
	(0.052)
Observations	298
Control group mean	0.14
Standard Utkarsh=Flexible Utkarsh (p-value)	0.76

Notes: This table reports whether the headmaster was present at school during the monitoring visit. The regression includes strata, week, and day-of-week fixed effects; a dummy for the headmaster being female, age and age squared, years of experience and years of experience squared, the school having multiple sections for class 9, total school enrollment, and indicator variables for the monitoring visit phase. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B16: Subject Teacher Surveyed During Monitoring Visit Survey

We surveyed at least one .... teacher	Every Subject	English	Math	Odia	Science
	(1)	(2)	(3)	(4)	(5)
Flexible Utkarsh	0.089	0.088	0.087	0.083	-0.104
	(0.066)	(0.056)	(0.058)	(0.059)	(0.064)
Observations	200	200	200	200	200
Standard Utkarsh Mean	0.36	0.77	0.78	0.74	0.81

Notes: This table shows whether a teacher of an Utkarsh subject was interviewed at the school during the monitoring visit survey. Column 1 shows whether at least one teacher for each subject was surveyed. Columns 2-5 show whether the subject teacher for the respective subject was surveyed. All regressions include the following school-level variables: average teacher age and its square, average teacher experience and its square, and the share of female teachers. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B17: Teacher Flexibility and Discretion with Additional Covariates

	(1) Filled out flexible Utkarsh teaching plan	(2) Followed flexible Utkarsh teaching plan for this week	(3) Had autonomy in using Utkarsh	(4) Teacher can adjust content if students have difficulty
Standard Utkarsh	0.017	0.014	0.945***	0.009
	(0.018)	(0.016)	(0.013)	(0.030)
Flexible Utkarsh	0.200***	0.139***	0.989***	0.060*
	(0.034)	(0.028)	(0.010)	(0.031)
Female (=1)	-0.010	0.002	-0.008	-0.015
	(0.029)	(0.025)	(0.016)	(0.041)
Age of the teacher (in years)	-0.031**	-0.024*	0.002	0.022
	(0.013)	(0.013)	(0.007)	(0.020)
Age squared	0.000**	0.000*	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Teaching experience (years)	0.013**	0.009**	-0.005	-0.012
	(0.006)	(0.005)	(0.003)	(0.011)
Teaching experience squared	-0.000**	-0.000**	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Subject teaches (=1)				
English	0.011	-0.005	0.031	0.032
	(0.048)	(0.039)	(0.025)	(0.065)
Odia	-0.007	-0.021	-0.001	-0.083
	(0.051)	(0.043)	(0.028)	(0.064)
Math1	-0.013	-0.039	-0.027	0.015
	(0.041)	(0.039)	(0.026)	(0.054)
Math2	0.041	0.019	-0.021	-0.044
	(0.039)	(0.035)	(0.029)	(0.056)
Science 1	0.010	-0.024	0.023	0.072
	(0.029)	(0.031)	(0.025)	(0.054)
Science 2	-0.029	-0.038	0.022	-0.048
	(0.037)	(0.030)	(0.026)	(0.047)
History	-0.066*	-0.077***	0.019	-0.029
	(0.035)	(0.029)	(0.025)	(0.060)
Geography	-0.025	-0.030	-0.015	-0.003
	(0.036)	(0.043)	(0.026)	(0.055)
Hindi	-0.072	-0.104**	0.022	0.289***
	(0.053)	(0.050)	(0.039)	(0.103)
Sanskrit	0.064	0.039	0.025	0.025
	(0.088)	(0.091)	(0.025)	(0.115)
Work Experience	-0.123	-0.163	0.022	-0.758***
	(0.128)	(0.118)	(0.039)	(0.093)
EVS	-0.216**	-0.183**	-0.009	0.181
	(0.093)	(0.079)	(0.038)	(0.116)
Other			0.072	0.093
			(0.055)	(0.166)
Has a teaching certificate (=1)	0.014	-0.001	-0.036*	0.068
	(0.029)	(0.029)	(0.019)	(0.042)
Days absent from work	-0.037***	-0.031**	0.000	0.006
	(0.014)	(0.013)	(0.008)	(0.021)
Works in another school(=1)	-0.045	-0.035	0.050	-0.231*
	(0.091)	(0.049)	(0.047)	(0.131)
Time spent in preparing for lesson (hours/week)	-0.002	-0.002	-0.001	0.000
	(0.002)	(0.001)	(0.001)	(0.002)
Time spent grading (hours/week)	0.001	-0.001	-0.001	-0.001
	(0.003)	(0.002)	(0.001)	(0.002)
Select teaching materials, methods, strategies (=1)	0.069*	0.032	0.039	-0.052
	(0.035)	(0.041)	(0.052)	(0.084)
Allowed to modify course timetable (=1)	-0.038	-0.055	0.015	0.233***
	(0.035)	(0.034)	(0.022)	(0.059)
Others do not select evaluation activities (=1)	-0.074**	-0.049	-0.029*	0.019
	(0.031)	(0.031)	(0.016)	(0.041)
Feel pressure to complete curriculum during school year (=1)	-0.053*	-0.029	0.013	0.024
	(0.028)	(0.027)	(0.019)	(0.041)

...(continue in next page)...

Table B17: Teacher Flexibility and Discretion with Additional Covariates (continued)

	(1) Filled out flexible Utkarsh teaching plan	(2) Followed flexible Utkarsh teaching plan for this week	(3) Had autonomy in using Utkarsh	(4) Teacher can adjust content if students have difficulty
...(continue from previous page)...				
Mean autonomy (standardized)	-0.000	-0.012	0.006	-0.016
	(0.015)	(0.018)	(0.010)	(0.023)
According to the teacher, percent of students who				
Will pass board exam in first try	0.000	0.001	0.000	-0.001
	(0.001)	(0.001)	(0.001)	(0.002)
Can write a simple English sentence	0.001	0.001	-0.000	0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Can do a three-digit sum	0.000	0.000	0.000	0.000
	(0.001)	(0.001)	(0.001)	(0.001)
Will eventually pass the class 10 board exams	-0.001	-0.001	0.000	0.000
	(0.001)	(0.001)	(0.001)	(0.002)
Will eventually pass the class 12 board exams	-0.001	-0.001	0.001	0.000
	(0.001)	(0.001)	(0.001)	(0.002)
Will eventually complete bachelors	0.000	-0.000	-0.001	-0.002
	(0.001)	(0.001)	(0.001)	(0.002)
Teacher feels (=1)				
Curriculum should be followed even if students have lower learning levels	0.049**	0.053*	-0.002	0.083**
	(0.023)	(0.029)	(0.015)	(0.041)
That if students are not ready for board exams, it would be the teacher's own fault	-0.001	-0.025	-0.013	0.006
	(0.025)	(0.023)	(0.019)	(0.040)
Valued and appreciated	-0.008	0.002	0.041*	0.021
	(0.030)	(0.026)	(0.021)	(0.042)
Satisfied with job	-0.019	-0.027	0.045	0.033
	(0.033)	(0.026)	(0.028)	(0.051)
That their opinion seems to count	-0.041	-0.053	0.018	0.155**
	(0.042)	(0.042)	(0.031)	(0.071)
That they have the materials and equipment to teach effectively	0.026	0.046*	-0.011	-0.003
	(0.024)	(0.025)	(0.019)	(0.042)
Similarly or more effective compared to colleagues	0.059	0.073	0.056	-0.067
	(0.052)	(0.071)	(0.047)	(0.091)
Observations	569	569	834	834
Control group mean	0.00	0.00	0.00	0.77
Standard Utkarsh = Flexible Utkarsh (p-value)	0.00	0.00	0.00	0.09
Source	Directly Observed	Directly Observed	Self-reported	Self-reported

Notes: This table reports the adoption of teacher flexibility with additional covariates. Column 1 measures whether the teaching plan is filled out. Column 2 is a self-reported measure of whether teachers followed the teaching plan that they filled out for that week. Columns 3 is a self-reported indicator variable of whether the teacher has autonomy in using Utkarsh lessons. Column 4 is a self-reported measure of whether the teacher can adjust the content of the lesson. Columns 1-2, measured at midline, include indicator variables for the monitoring visit phase. Columns 3-4, measured at endline, include an indicator variable for the early endline visit. All regressions include strata, week, and day-of-week fixed effects. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B18: Teacher Flexibility and Discretion Heterogeneity by BL Student Test Score Standard Deviation

	Filled out flexible Utkarsh teaching plan	Followed flexible Utkarsh teaching plan for this week	Had autonomy in using Utkarsh	Teacher can adjust content if students have difficulty
	(1)	(2)	(3)	(4)
Standard Utkarsh	0.037	0.109	0.936***	-0.191
	(0.111)	(0.098)	(0.068)	(0.162)
Standard Utkarsh*BL Test Score SD	-0.020	-0.103	0.009	0.214
	(0.120)	(0.106)	(0.073)	(0.175)
Flexible Utkarsh	-0.012	0.068	0.924***	-0.332**
	(0.165)	(0.127)	(0.053)	(0.164)
Flexible Utkarsh*BL Test Score SD	0.239	0.092	0.070	0.441**
	(0.178)	(0.134)	(0.058)	(0.177)
BL Test Score SD	-0.025	0.041	-0.045	-0.218*
	(0.076)	(0.062)	(0.032)	(0.125)
Observations	569	569	834	834
Control group mean	0.00	0.00	0.00	0.77
Standard Utkarsh+Standard Utkarsh*BL Test Score SD=0 (p-value) (p-value)	0.42	0.71	0.00	0.49
Flexible Utkarsh+Flexible Utkarsh*BL Test Score SD=0 (p-value) (p-value)	0.00	0.00	0.00	0.00
Source	Directly Observed	Self-reported	Self-reported	Self-reported

Notes: This table shows the heterogeneity of teacher adoption of flexibility with respect to the school-level standard deviation (SD) of baseline student test scores, i.e., the SD of the overall baseline test scores within each school. Column 1 measures whether the teaching plan is filled out. Column 2 is a self-reported measure of whether teachers followed the teaching plan that they filled out for that week. Columns 1-2, measured at midline, include indicator variables for the monitoring visit phase. Column 3 is a self-reported measure of whether the teacher has autonomy in using Utkarsh lessons. Column 4 is a self-reported measure of whether the teacher can adjust the content of the lesson. Columns 3-4, measured at endline, include an indicator variable for the early endline visit. 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 whether the teacher is female, and a vector of dummy variables for the main subject taught by the teacher at baseline. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B19: Additional Teacher Outcomes

	Burnout index	Stress index	Anxiety index	Lesson preparation time (hours/week)	Grading time (hours/week)	Teacher enjoyed Utkarsh
	(1)	(2)	(3)	(4)	(5)	(6)
Standard Utkarsh	0.105	0.059	0.025	-0.229	0.410	0.953***
	(0.089)	(0.082)	(0.090)	(0.625)	(0.558)	(0.014)
Flexible Utkarsh	0.135	0.082	0.049	-0.842	0.259	0.940***
	(0.089)	(0.090)	(0.096)	(0.649)	(0.551)	(0.016)
Observations	834	834	834	834	834	834
Control group mean	0.00	0.00	0.00	13.65	8.94	0.00
Standard Utkarsh = Flexible Utkarsh (p-value)	0.75	0.79	0.79	0.34	0.78	0.54

Notes: This table reports additional teacher outcomes measured at endline. Column 1 presents an inverted covariance matrix-weighted standardized index generated following Anderson (2008) from the following variables: teacher's feeling of mental exhaustion from work; feeling fatigued; feeling of having a positive influence on people; feeling very energetic about the job; and feeling satisfied with the job at school. Columns 2 and 3 are measured on the Depression Anxiety Stress Scale (DASS). To construct the respective indices, responses to the relevant questions of the scale are summed and then standardized. Columns 1-3 range between 0 and 1. Columns 4-5 measure self-reported time spent on preparing for lessons and grading, respectively. Column 6 is a self-reported measure of whether the teacher enjoyed Utkarsh. All regressions include strata, week, and day-of-week fixed effects; teacher's age in years and age squared, years of experience and experience squared, a dummy variable for being female, a vector of dummy variables for the main subject taught at baseline, and indicator variables for the early endline visit. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B20: Teacher Competency

	Self-Assessed Effectiveness			Competency Exam				
	At least as Effective as Other Teachers	More Effective Than Other Teachers	Much More Effective Than Other Teachers	Test Score (Percent)		Competency Exam Administered in School	Subject Teacher Tested in the Competency Exam Was Surveyed at Baseline	
				English	Math		English	Math
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Standard Utkarsh	0.012	-0.005	-0.007	-0.476	1.067	0.009	0.069	-0.008
	(0.012)	(0.041)	(0.021)	(2.399)	(1.591)	(0.019)	(0.072)	(0.056)
Flexible Utkarsh	0.003	0.020	-0.021	-0.943	2.909**	-0.015	0.050	-0.078
	(0.013)	(0.040)	(0.020)	(2.655)	(1.413)	(0.023)	(0.076)	(0.058)
Observations	834	834	834	226	303	300	226	303
Control mean	0.97	0.41	0.08	57.83	89.21	0.80	0.65	0.64
Standard Utkarsh=Flexible Utkarsh (p-value)	0.39	0.52	0.51	0.83	0.24	0.24	0.76	0.24

Notes: This table reports teachers' self-reported and researcher-administrated test- based competency. Columns 1-3 are indicator variables capturing teachers' self-assessed effectiveness. Columns 4-5 measure test scores in respective subjects by respective subject teachers on the teacher competency examination (values range between 0 and 100). Columns 1-5 include week-of-survey and day-of-week fixed effects, a dummy for the teacher being female, the age of the teacher and age squared, the teacher's years of experience and experience squared. Columns 4-5 additionally include a vector of dummy variables for the main subject taught at baseline and indicator variables for early endline visits. Column 6 measures whether the school participated in the teacher competency examination. The regression includes a dummy for the headmaster being female, the headmaster's age and age squared, the headmaster's years of experience and years of experience squared, whether the school has multiple class 9 sections, and total enrollment. Columns 7-8 are indicator variables measuring whether the respective subject teacher was surveyed at baseline. Regressions include week-of-survey and day-of-week fixed effects, a dummy for the teacher being female, the age of the teacher and age squared, and the teacher's years of experience and experience squared. All regressions include strata fixed effects. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B21: Detailed Effects on Board Exam Marks

	Grades			Test Scores			
	A	C or above	D or above	English	Math	Odia	Science
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Standard Utkarsh	-0.011**	-0.069***	-0.046**	-0.123**	-0.122**	-0.123***	-0.136***
	(0.005)	(0.022)	(0.018)	(0.051)	(0.056)	(0.044)	(0.051)
Flexible Utkarsh	-0.002	-0.046**	-0.028	-0.094*	-0.045	-0.079*	-0.087
	(0.006)	(0.022)	(0.018)	(0.053)	(0.053)	(0.046)	(0.054)
Observations	18,551	18,551	18,551	18,551	18,551	18,551	18,551
Control group mean	0.05	0.65	0.85	0.00	0.00	0.00	0.00
Standard Utkarsh=Flexible Utkarsh (p-value)	0.15	0.33	0.35	0.60	0.17	0.36	0.37
Raw control mean				0.53	0.58	0.56	0.55
Raw control SD				0.15	0.16	0.16	0.15

Notes: This table shows the effects on board exam marks in detail using administrative data on board exam results. Columns 1-3 indicate whether the student achieved the respective grade or above. Columns 4-7 show standardized total scores in the respective subjects. All regressions include strata fixed effects; standardized IRT scores from baseline English, math, and Odia tests, a dummy for whether the student is female, and the age of the student. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Table B22: Longer-Term Follow-up Survey Details

	Board exam scores of our study sample collected	Analytical Sample Students Surveyed in the Longer-Term 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	0.92	0.22	0.05
Standard Utkarsh=Flexible Utkarsh (p-value)	0.13	0.41	0.64

Notes: This table contains data from a longer-term follow-up survey conducted in November-December 2021. Column 1 is an indicator variable showing whether the board exam scores were collected for students in our sample. Column 2 is an indicator variable showing whether a student in our sample was surveyed in the longer-term follow-up survey. Column 3 reports whether the students surveyed in the follow-up survey reported that they took the offline board exam. All columns report indicator variables. All regressions include strata fixed effects; student's standardized IRT scores from baseline English, math, and Odia tests; a dummy for whether the student is female; and the age of the student. Heteroskedasticity-robust standard errors, clustered at the school level, are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Appendix C. Additional Implementation Details

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 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.

Appendix D. 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 on 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 (four English questions and six math questions) that may not necessarily map to the official curriculum. Tests were group-based and carefully proctored to ensure that students did not cheat. Similar questions were used across both waves of the survey.

We used item response theory (IRT) to construct students' test scores, pooling all ques-

tions from baseline and follow-up. We first checked that there were no questions with poor discrimination properties. We then implemented 1-parameter IRT using Stata. 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 standardize 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 the 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.[27]

[27]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.