IDinsight
DATA. DECISIONS. DEVELOPMENT.

STiR education
Empowering Teachers to be Changemakers

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Impact of STIR's programming on teacher motivation and student learning

Endline report

July 2018[1]

[1] This report has been prepared by the IDinsight team. Please direct all correspondence regarding this report to heather.lanthorn@idinsight.org.

List of tables and figures

Abbreviations

APS	Affordable Private Schools
ASER	Annual Status of Education Report, produced in India by the ASER Centre
BEO	Block Education Officer
CARP	Commitment to Analysis and Reporting Plan
DIET	District Institute for Education and Training
EL	Education Leader
HT	Head Teacher (school principal)
ISIT	In-school Innovation Team
IV	Instrumental Variable
LATE	Local Average Treatment Effect
MDE	Minimum Detectable Effect
OLS	Ordinary Least Square
PM	Program Manager
RE	Randomized Evaluation (sometimes called randomized control trials)
SIEF	Strategic (formerly Spanish) Impact Evaluation Fund
STIR	Schools and Teachers Innovating for Results
ToC	Theory (or model) of change
TOT	Treatment-on-the-treated estimate (teacher-level estimate of STIR's program)
U.P.	Uttar Pradesh, a state in northern India

Acknowledgements

This report has been a collaborative effort. The report has been driven by the core team of the evaluation, namely Pratima Singh, Qayam Jetha, Heather Lanthorn, and Doug Johnson. We would not have the data presented in this report without our intrepid and dedicated Senior Field Managers, Rajkumar Sharma and Pramod Kumar; they have ensured high-quality data collection, time and time again. We would like to thank Andrew Fraker, Jeff McManus, Akshat Goel, Dan Stein, Varun Chakravarthy, and Torben Fischer for their technical advice and input through the course of the analyses and writing the report. We have benefitted from the perspective lent by Ronald Abraham, Neil Buddy Shah, and many others around the organization. We deeply value everyone's contribution and are thankful for their advice.

We would also like to thank the ASER Center, who have provided feedback and have lent their expertise throughout the evaluation. The World Bank's Strategic Impact Evaluation Fund (SIEF) provided funding for the evaluations and has offered key technical support — special thanks to Sangeeta Goyal and Alaka Holla. We also want to thank two anonymous reviewers of this report, who's inputs strengthened it considerably. STIR have been excellent learning partners, and have been patient, understanding, helpful, and supportive through the project's duration. We thank them for all the logistical help during data collection and ensuring smooth surveying over the course of the evaluations. It has been a pleasure interacting with their teams in London, Delhi, and Uttar Pradesh.

Finally, we thank the schools for welcoming us and teachers and students for allowing us to interact with them.

We would like to sincerely acknowledge everyone's contribution and express our deepest gratitude.

1 Executive summary

STIR Program and Evaluation Design

We report results from randomized evaluations of STIR's programming in Delhi and Uttar Pradesh (U.P.). STIR seeks to improve teacher motivation and classroom practice by organizing teachers into local networks. These networks hold monthly, guided meetings where teachers discuss principles of classroom practice and share ideas for how to improve their teaching. In Delhi, STIR worked with teachers at private schools with monthly fees less than $17 (sometimes referred to as "affordable private schools") and STIR staff directly organized and guided the monthly network meetings. In U.P., STIR worked with government schools and trained and coached volunteer government school teachers to organize and guide the meetings. In both Delhi and U.P., schools included grades from 1st to 8th standard and roughly 20% of teachers participated in STIR meetings.[2]

We randomized the offer of STIR programming in two stages. First, schools were randomly assigned to either treatment or control. We then grouped nearby treatment schools into clusters and randomly assigned each cluster of schools to receive either the STIR "standard" model or the STIR "exploratory" model. In addition to the network meetings, teachers in the exploratory model also received non-financial incentives such as recognition from local officials. We collected data on classroom practices, teacher motivation, and student learning outcomes at baseline, one year later, and two years later.

All findings reported below are school-level results. That is, we compare *all* teachers and students in treatment schools, regardless of whether the teacher participated in STIR meetings, to *all* teachers and students in control schools. In the body of the report, we also present estimates of the effect of STIR on teachers who participate in STIR meetings.

Findings from Delhi Private Schools

In Delhi, we find that the offer of STIR programming led to improved math learning outcomes. Students in STIR schools (standard + exploratory combined) increased math learning levels by .1 standard deviations (p-value: 0.02) and students in the standard treatment arm increased math levels by .15 standard deviations (p-value: 0.04) compared to students in control schools. These effects appear to be driven mainly by poor performers achieving a basic math learning level. **We find no effect on Hindi learning outcomes in Delhi.**

In Delhi, we also find suggestive evidence that STIR led to increased teacher motivation. STIR led to a 0.13 standard deviation increase in an overall index measuring teacher motivation among teachers in the standard treatment arm (p-value: <0.01). In addition, we find effects on a sub-index which sought to measure "growth mindset," one of three analyzed sub-indices. STIR led to a 0.15 standard deviation increase on the growth mindset sub-index among teachers in STIR schools and a 0.18 standard deviation increase on this sub-index among teachers in the standard treatment arm. We do not find significant effects for the overall index for STIR schools or for the two other sub-indices (teacher efficacy and positive professional outlook).

[2] According to STIR, participation was limited to roughly 30% of teachers at each treatment school. Our understanding is that, in practice, this cap was rarely binding.

Findings from U.P. Government Schools

In U.P. government schools, we find weak evidence of gains in the amount of time teachers spend teaching. STIR led to a 4 percentage-point increase (p-value: 0.08) in observed teaching time among teachers in STIR schools. In the standard treatment arm, STIR led to an 8-percentage-point increase (p-value: 0.09) in observed teaching time. We characterize this evidence as weak given the large number of outcomes we test for and the relatively large p-values of the results.

In U.P, we do not find statistically significant effects on teacher motivation, student learning outcomes, or other classroom practices.

Conclusions

Our results show that STIR's approach can work but that its effectiveness depends on context, where context may include geography; education systems, financing, and staffing; and program components and approaches to delivery. In Delhi, STIR caused a 0.1 standard increase in math learning outcomes. This result is similar in size to effect sizes from other teacher training and incentive interventions in low- and middle-income countries (McEwan 2015; Snilstveit et al. 2015). In U.P., we find weak evidence that STIR may have increased teaching time and no effect on learning outcomes and several other measures. Unfortunately, we are unable to pin down the source of this difference. There are several large differences in both the context and implementation model between the Delhi and U.P. versions of the program. Our evaluation is unable to disentangle the importance of these differences.

Limitations

This study has three key technical limitations. First, we experienced a high level of teacher and student attrition. We do not detect differential attrition on observables between treatment and control but cannot rule out differential attrition on unobservable characteristics. Second, we analyze many hypotheses which raises the risk of false positive findings. We correct for multiple hypothesis testing within outcomes families with more than four outcomes but do not correct across outcome families. Third, our classroom observations may be subject to observer effects, as some of the child-friendly measures were explicitly highlighted as part of discussions in the community of practices.

2 Introduction

This report is organized as follows. In Section 2, we briefly introduce the STIR Teacher Changemaker Journey, evaluation context, and evaluation design. In Section 3, we turn to a more detailed overview of the programmatic components and stakeholders. In Section 4, we present details of the evaluations including objectives, questions, design, methods and analytical approach. Section 5 lays out the main results from the evaluations and in Section 6, we conclude with a brief discussion of the limitations.

2.1 Background

Nearly 139 million children are enrolled in primary school in India, and there are 4.4 million primary school teachers across government and private facilities (World Bank 2018a). While there has been encouraging progress in getting children into schools — 97% of primary-school-age children are enrolled in school — the performance of Indian children in school is poor relative to most low- or middle-income countries (International Association for the Evaluation of Educational Achievement). Recent estimates suggest that by grade 9, students in urban India are, on average, nearly 4.5 grades behind in math and 2.5 grades behind in Hindi (Muralidharan, Singh, and Ganimian 2017).

This focuses our attention on what is happening in schools. In India (as in many low- and middle-income countries), teacher performance is often poor. Random audits of public school teachers found that teachers were absent 24% of the time and even when present do not spend all their time actively teaching (Muralidharan et al. 2016). Teacher attendance and activity is similarly low in the 'affordable private schools' (APS) proliferating throughout India (Goyal and Pandey 2009).

Teacher motivation — one component of professional mindsets and behaviors — is increasingly viewed as an important route to improving teacher effectiveness in the classroom, for which "no amount of training or inputs can substitute" (World Bank 2018b). The most recent *World Development Report* highlights that "effective teaching depends on teachers' skills and motivation" (World Bank 2018b). Muralidharan, focusing specifically on India, stresses the importance of motivation and professional incentives in encouraging teachers to revise their teaching toward more effective classroom practices (Muralidharan 2012). It is already established that financial incentives do not always sustainably improve teachers' effort or student learning (de Ree et al. 2017). Therefore, finding non-financial routes to increase teacher motivation may be crucial in improving student learning outcomes. This is the problem STIR communities of practice intend to solve.

2.2 Overview of STIR's program

STIR seeks to improve teachers' 'professional mindsets and behaviors' and, in turn, the quantity and quality of their classroom culture and practices, with the aim of improving student learning outcomes.[3] This is illustrated in a basic diagram of the theory of change, shown in Figure 1. A more detailed version is elaborated in Appendix A1. Professional mindsets and behaviors comprise STIR's target suite of nine attitudes and skills: intrinsic motivation (Ryan and Deci 2000), growth mindset (Dweck 2010), efficacy (Bandura 1969), resilience, commitment, collaboration, micro-innovation, reflective practice, and influence.

[3] Over the two years of this study, STIR's and our understanding of the theory of change has evolved. It is now thought upon as a virtuous cycle where the direction of the change is both ways; *i.e.*, there may also be feedback effects from classroom practice and student outcomes on professional mindsets and behaviors. Even the effort to make changes in classroom practice and culture could have a positive or negative impact on a teachers' motivation.

Figure 1: Simple theory of change of STIR communities of practice, with key impact outcomes

To achieve this, STIR organizes primary schools into local networks, made up of teachers in neighboring schools and sometimes in a single STIR school. These networks meet monthly in guided, collaborative sessions, providing the opportunity for teachers to build intrinsic motivation, to shape a growth mindset and professional attitudes, to learn, and to collaboratively overcome day-to-day challenges related to teaching, classroom management, and classroom culture (Dweck 2010). Teachers take these ideas back to their classrooms to try to plan and implement changes intended to be consistent with improving the quality and quantity of classroom practice in ways that promote student learning. Teachers are further encouraged to reflect upon their teaching practices using a portfolio and to influence others around them.

In contrast with most other teacher professional development programs STIR does not focus on specific pedagogic techniques, subject knowledge, or technology. Rather, STIR seeks to use communities of practice to improve teachers' professional behavior and sense of agency to make positive changes in their classrooms and schools. In addition, while most teacher-incentive programs focus on financial or in-kind incentives — often linked to student performance — STIR focuses on non-financial motivators linked to teacher effort.[4]

STIR estimates their per-child annual costs[5] are US$ 1.10 in U.P. government schools and US$ 10.45 in Delhi private schools. The estimated costs per participating teacher per academic year are US$ 33.00 in U.P. government schools and US$ 366.00 in Delhi private schools.

[4] There are no direct financial or career-progression implications for a teacher who participates with STIR, though the literature suggests that adding such implications may lead to higher student test scores following teacher training (Evans, Popova, and Arancibia 2016).

[5] STIR reports that their costs figures are "fully inclusive of all of our costs including program, M&E, operations/logistics, and salaries." IDinsight has not independently verified these figures.

2.3 Overview of the two evaluations: design and context

To assess the impact of STIR communities of practice, IDinsight[6] has undertaken a pair of randomized evaluations (REs) in private schools[7] in East Delhi and government schools in the districts of Rae Bareli and Varanasi in Uttar Pradesh (U.P.).[8] The 'treatment' — in this case, the offer to schools to have their teachers join STIR — was randomized at the school level.[9] Teachers in each treatment school could opt to apply and, if selected, join the program. Across schools, teachers may have experienced differing degrees of support or pressure related to joining.

Delhi and Uttar Pradesh are distinct geographic, institutional, and implementation contexts for STIR's program. In Delhi, included schools are urban and private, which tend to be staffed by younger and less formally qualified teachers. There is no overarching authority or linking system; each private school is a business unto itself. Thus, school interest and engagement in the program and evaluation had to be negotiated and renegotiated with each atomized Head Teacher (principal) and/or school owner. In Delhi, STIR communities of practice and other program activities were carried out directly by STIR staff acting as Education Leaders; for this reason, STIR conceived of Delhi as a 'lab,' with higher control over implementation quality.

In our Uttar Pradesh sample, schools are largely rural and are all government schools, which requires a higher level of teacher qualification than the affordable private schools in which STIR works. The U.P. model evaluated here is embedded in the government structure, with network meetings led by volunteer (or volunteered) government-school teachers acting as Education Leaders in addition to teaching.[10] They received training and support from government officials who have, in turn, been trained by STIR staff and officials (that is, a cascade delivery model). In this systems model, a letter from the Block Education Officer was largely sufficient for individual schools to engage with the programming. In U.P., STIR staff also works with Block Education Officers to help them appreciate the importance of professional mindsets and behaviors for student learning.

---

[6] About IDinsight (evaluator): IDinsight seeks to partner with clients committed to using and generating rigorous evidence to improve their social impact. IDinsight believes that decision-oriented and rigorous approaches to evaluation, monitoring, and measurement are essential to help managers maximize their impact through informing their decisions and actions (Shah et al. 2015). STIR has engaged IDinsight as an evaluation partner since 2013, when IDinsight assisted with process evaluations and support for the development of a theory of change.

[7] STIR only works with schools that have monthly fees of US$ 17 or less; for this reason, STIR and others in India refer to these schools as 'affordable private schools.'

[8] Both REs are supported financially and technically by the Strategic Impact Evaluation Fund (SIEF) under the grant "Impact of Non-Financial Teacher Incentives, India." While STIR's programming now spans three years (in U.P.), this set of evaluations covers two academic years. When this evaluation was first conceptualized, STIR was planning for a two-year Teacher Changemaker Journey For background on randomized evaluations, refer to Appendix A19.

[9] The alternative, of randomizing the offer to join within schools (or putting out the offer to join and then randomizing among interested teachers within schools), presents major implementation problems (teacher jealousy) as well as severe evaluation concerns (potential for contamination between treated and control teachers within the same school).

[10] Unlike in Delhi, STIR did not have direct control over who functioned as an Education Leader in U.P. During the Year 1, ELs in U.P. were often chosen with heavy input from BEOs and other government officials. Groups of ELs in U.P. are managed by Program Mangers (PMs), who are STIR employees. In Year 2, to replace any drop-out ELs, three teachers recommended by the BEO would attend a two-hour session by STIR. Then each nominee was interviewed by two BEOs and one STIR representative, resulting in one teacher selected.

2.4 Programming variations: standard and exploratory models

Each evaluation experimentally investigates two iterations of STIR's approach that represented possible ways forward for programming: a standard and an exploratory model. After we randomly assigned schools to treatment and control, treatment schools were grouped into communities of practice, which were then randomized to receive either the standard or the exploratory model. The contents of these models have changed over the course of each of the two-year evaluations, but schools have remained in the standard or in the exploratory treatment arm throughout.

In Year 1, both models focus on intrinsic motivation[11] among teachers through network meetings, reflective portfolios, and the classroom changes that follow. In Year 1, the standard model[12] used a package of techniques to increase teachers' intrinsic motivation, including: making them feel part of a large, important movement; participating in activities to increase self-actualization as a professional; and a shift in mindset to believe they are responsible for and capable of improving student learning outcomes. In Year 1, the exploratory model[13] built upon the standard model of generating intrinsic motivation by also adding non-financial motivators to boost teacher extrinsic motivation (such as recognition posters) and encouraging teacher participation with STIR (details in Appendix A2).[14]

In Year 2, three key changes were made to the program and evaluation design. First, for programming, STIR took extrinsic motivators considered successful in the exploratory model of Year 1 and incorporated them into all Year 2 programing.[15] Second, as planned, the programmatic focus of Year 2 was on six principles of good classroom practice and culture, each discussed in a network meeting and reinforced through portfolio exercises. In each meeting, Education Leaders introduced a new principle and teachers discussed how to operationalize and implement it in their classrooms through micro-innovations.

Third, in Year 2, STIR had a different evaluative question, centered on how teachers can best operationalize each principle of classroom practice and culture. STIR sought to test whether encouraging teachers to design the innovation would result in increased ownership of the process, as revealed through classroom practices and student learning. To explore this question, in Year 2, teachers in standard model schools chose between a small set of provided, evidence-informed micro-innovations to try in the classroom.[16,17] In the

[11] Intrinsic motivation refers to doing something because it is inherently interesting or enjoyable. Extrinsic motivation refers to doing something because it leads to a separable, desirable outcome (Ryan and Deci 2000).

[12] Called 'core' in Year 1 of the evaluation.

[13] Called 'core-plus' in Year 1 of the evaluation.

[14] All exploratory activities happen outside of the network meeting and reflective portfolio process that is the center of STIR's learning and engagement platforms. Participation in these activities is explicitly kept separate from active participation in STIR's central activities (such as attendance at network meetings or portfolio completion).

[15] For example, the 'local recognition' set of activities were rolled out to all treatment schools across both the arms.

[16] This model is also referred to as the selection model by STIR, as teachers selected from a set of innovations.

[17] In practice, it is not clear how distinct these two treatment arms remained. According to STIR staff in both Delhi and U.P., sometimes the standard networks wanted to develop a different idea and, more commonly, the exploratory networks wanted to choose from some example ideas provided (which were the menu of ideas presented to the standard networks). Despite efforts to reinforce the messages of these two groups distinctively (through groupings

exploratory model, in contrast, teachers were tasked with collaboratively developing their own innovation to operationalize the principle as a specific classroom change.[18] Because STIR has traditionally provided a menu of micro-innovations to teachers, the exploratory model is considered more experimental.[19]

for WhatsApp© groups, through AwaazDe© voice messages, and through reports given to Head Teachers), it is not clear that the standard and exploratory models in Year 2 can be considered distinctly different.

[18] This model is referred to as the co-creation model by STIR, as the teachers collaboratively designed an innovation.

[19] Over the course of the two-year evaluation, teachers in schools randomized to the standard treatment arm received the core model in Year 1 and the selection model in Year 2. Teachers in schools randomized to the exploratory treatment arm received the core-plus model in Year 1 and the co-creation model in Year 2. In Year 2, STIR introduced elements of Year 1's core-plus model into their regular programming and were thus received by both standard and exploratory schools.

3 Details of key programmatic components, activities, and actors

The two-year Teacher Changemaker Journey has many moving parts. We detail our understanding of key elements, which are useful for understanding STIR's program as well the evaluation. We also summarize program components in Table 1.

3.1 Education Leaders

STIR communities of practice are largely implemented by Education Leaders (ELs). ELs play a role in coordinating and facilitating key programmatic activities, which requires them to not only interact with the teachers in their 'networks' but also with key education gatekeepers and stakeholders such as Head Teachers (HTs, particularly in Delhi) and Block Education Officers (BEOs, in U.P.). They coordinate and facilitate the monthly network meetings to bring together teachers within and across schools to share, learn, collaborate, and support one another. In the private school model in Delhi, ELs were STIR staff. In the government school model in U.P., ELs are volunteer government school teachers, who are trained by government officials who are in turn trained by STIR staff. In Year 2, ELs in U.P. received a coaching call from a Program Manager (STIR staff) after each network meeting. In Delhi, ELs visited each school at least once per two network meetings to have coaching check-ins with individual teachers (though this was sometimes resisted by Head Teachers). These visits did not happen in U.P., as Education Leaders are teachers and do not have time to visit other schools and observe teachers in action and support their efforts to change classroom practice and culture.

3.2 Head Teachers

Head Teachers are principals of individual schools. They may or may not engage in teaching themselves. Apart from responsibilities as a teacher (where relevant), they are also responsible for all management and administrative responsibilities in that school. This includes issues related to the infrastructure of the school; to hiring[20], maintaining, monitoring, and managing teachers; and to keeping abreast of new government schemes. Working with Head Teachers is particularly important in Delhi private schools, where these individuals serve a key gatekeeping role for the school. To increase Head Teacher buy-in, STIR worked to make sure that they were active partners in the program, creating an enabling environment that grants teachers "permission to innovate" and reduces system pressures on practice (STIR Education 2017). In U.P. government schools, because STIR has high-level permission from the district and block levels to operate in schools, STIR has not needed to actively engage with Head Teachers to enter schools.

3.3 Block Education Officers

Blocks are administrative units in India, smaller than districts. A Block Education Officer (BEO) is responsible for education in one block. Additional responsibilities include handling administrative and management issues related to the schools in their block and ensuring education outcomes in India. By working with BEOs, STIR works to create an enabling environment that grants teachers "permission to innovate" and reduces system pressures on practice (STIR Education 2017). Working with BEOs and district officials is particularly important for working with government schools, as they serve as key gatekeepers into the school system; a letter from the BEO facilitates STIR's entry into schools in that block.

[20] Note the responsibilities of a Head Teacher may vary from school to school and also across geographies. For instance, in Delhi roles may change according to school structures. In U.P., hiring is done centrally via government channels.

3.4 Years 1 & 2 programming sequence

Year 1 activities in STIR's communities of practice were organized into three phases of roughly equal duration (STIR Education n.d.). All activities during this part of the Teacher Changemaker Journey were intended to "ignite" teachers' passion for teaching and to inspire a sense of agency and self-efficacy toward changing teaching practice (STIR Education n.d.).

The first phase focused on *innovation*. During this phase, ELs focused, in particular, on encouraging teachers to develop a positive, proactive mindset about their ability to improve their teaching, and to begin to explore micro-innovations (defined below). The second phase — *implementation* — required teachers to select one or more micro-innovations to work on putting into practice, and then reflect on the results. The third phase called for teachers to exercise *influence*. This included the formation of In-School Innovation Teams (ISITs, defined below) as well as outreach to the families of five under-performing students (defined below).

In practice, Year 1 of the program was not so linear. For example, motivation and mindsets were stressed throughout the first year rather than only in the first phase. Similarly, while influence was the focus of the third phase, active STIR teachers may have been influencing teachers, parents, and other stakeholders throughout. Nevertheless, the phases are useful in thinking about the focus of key activities (network meetings, portfolios, and classroom/school activities) over the course of Year 1.

Year 2 of STIR communities of practice focused on a set of six principles of good Classroom Culture and Practice (CCP), with each introduced in one of the monthly meetings, during which teachers select or create a corresponding micro-innovation to implement in their classrooms. The six CCP principles are:

1. The example a teacher sets as a learner herself/himself creates the tone for learning in the classroom.
2. An engaging physical classroom environment supports learning.
3. Effective classroom routines facilitate good teaching and learning in the classroom.
4. Teachers need to know their students and students need to feel valued in order to learn.
5. A good learning environment encourages respectful dialogue between teacher and students as well as between students.
6. Punishment discourages learners; applying positive behavior motivates learners.

3.5 Networks and network meetings

A central activity for teachers participating in STIR communities of practice is to attend network meetings. Networks offer an inter-school platform (or community of practice) for teachers to learn from ELs and each other and to collaborate with and support their peers. Networks met monthly (excluding months when school was not in session) during the year of programming evaluated here, for between 45 minutes and two hours per meeting, with meeting time divided between instruction and discussion. These meetings are organized and facilitated by ELs, allowing teachers to learn new concepts, develop the (growth) mindset of a problem-solver, discuss and collaborate over classroom challenges, and receive support and ideas from other teachers.

In general, ELs schedule monthly meetings to accommodate as many teachers' availability as possible and select rotating meeting locations to minimize travel time for teachers. In Delhi, these meetings take place either after school hours or on the weekends. On the other hand, in U.P., the meetings take place during the school day — as required by the government — often requiring participating teachers to miss a full day

of school on a meeting day. In Year 2 in Delhi, there were cases where teachers said they could or would not travel to inter-school network meetings, and so about 20% of meetings were held with all interested teachers in a single school at a time.

3.6 Reflective portfolios

Teachers receive reflective portfolios in their network meetings. These are workbooks for teachers to complete in the month between meetings, helping them to think about their current teaching practice and to prepare for future changes in practice. These workbooks serve four key functions. First, they provide resources and ideas to teachers. Second, they provide teachers a diary in which to plan and track micro-innovations and implementation progress. Third, it encourages reflection by posing questions to teachers related to the Learning Improvement Cycle, such as considering implementation challenges faced, the effects on students, and to think critically about further ways to improve their teaching, classrooms, schools, and school systems. Fourth, they provide an accountability mechanism for STIR and allow ELs to assess how well teachers are internalizing the STIR model. Adequate portfolio completion also plays an important role in determining individual teacher's eligibility for certification (described below) and forms part of the basis for participating in STIR.

3.7 Micro-innovations

Micro-innovations are small changes teachers can make in their classroom practice and environment. These changes may relate to both the quantity and quality of teaching, classroom culture and environment, and classroom managements strategies. Some ideas also extend beyond the classroom, including school-wide initiatives or reaching out to students' families or education stakeholders.

In Year 1, in both the standard and exploratory models, an initial booklet of micro-innovation ideas is presented to teachers who join STIR; the ideas in the booklet are collected and collated by STIR, drawing from local teachers' practices identified through a search process. Implementing micro-innovations provides teachers their (potentially) first experiences leading change in their classrooms and an opportunity to experience both struggles and successes; in theory, this ultimately builds teacher confidence about being agents of change through developing or adapting new ideas and seeing results from their implementation. Some insight on the array of micro-innovations and their aims from our early-2016 process evaluation (during Year 1 of the randomized evaluation) can be found in Appendix A3.

In Year 2, micro-innovations are specifically focused on the principles of Classroom Practice and Culture.

3.8 Local Recognition package

When the local recognition package is in place (in some treatment schools in Year 1 and in all treatment schools in Year 2), it includes the following activities: recognition posters hung in schools (once in Year 2 in Delhi; roughly once every two months in U.P. in Year 2), a 'family day' event to showcase teachers (in Delhi only), and an encouraging AwaazDe© call.[21]

[21] A voice SMS service: https://awaaz.de/

3.9 Influencing other teachers: In-school Innovation Teams

As part of an In-School Innovation Team (ISIT), STIR teachers in Year 1 were supposed to lead two other (not actively participating) teachers in the same school who are interested in implementing micro-innovations and learning about STIR network activities. Teachers who do this successfully are eligible for an ISIT certificate.

Since only some teachers in each 'treated' school become direct participants in STIR communities of practice, STIR established ISITs as a mechanism for participating teachers to influence other teachers in their schools, driving toward a tipping point of school-wide change in professional mindsets and behaviors as well as practice. Participating STIR teachers demonstrate innovative practices to their ISIT and inspire them to try these in their classrooms/schools. These teachers must organize the ISIT meetings; share ideas and solutions in meetings; encourage ISIT teachers to micro-innovate; observe ISIT teachers' practice and invite other ISIT teachers to observe their practice. Through this, participating STIR teachers create an environment of collaborative learning between teachers in their schools. By bringing more teachers into STIR's activities and encouraging them to adopt similar approaches, STIR aims to facilitate the creation of a collective movement of teachers to improve the education system.

3.10 Influencing families: five under-performing students

In Year 1, an additional influence activity that actively participating teachers are meant to undertake is to identify five under-performing students and to try to engage with their families (guardians or caregivers). This recognizes the central role that families play in student motivation, attendance, and effort on homework. Previous research indicates that updates to parents often prove particularly effective in motivating students when they focus not just on learning levels, but also on attendance and performance on individual assignments (World Bank 2018b). When parents are responsive to teachers' efforts, it also provides teachers an opportunity to set realistic goals with the potential for quick successes.

3.11 70-Day Challenge

The 70-Day Challenge reflected a 70-day effort to keep teachers engaged in the program; this came out of a recognition that coaching calls and other elements of programming were being missed. STIR staff wanted to challenge themselves to be operationally efficient as the final term captured by the evaluation ended. In Delhi, this happened at the end of the Year 2 academic term, just before our data collection. In U.P., this happened during the winter and summer breaks; the summer break also immediately preceded our endline data collection. Operationally, this entailed increasing interactions with teachers in Delhi in the last 70 days of the program and ensuring teachers felt that their efforts in the classrooms were being recognized by STIR. In U.P., the STIR staff increased their meetings with the ELs to ensure continuity in their training during the summer and winter breaks and catch up on any pending activities that might have been missed due to paucity of time during the term time.

3.12 Teacher Changemaker Certification (Roehampton Certificate)

In partnership with the University of Roehampton, STIR awards some participating teachers with a certification as a Teacher Changemaker at the end of Year 1 and again at the end of Year 2. This is contingent on the criteria below.

- attending 75% of more of network meetings;

- showing evidence of planning and implementing micro-innovations;
- showing evidence of planning and executing influencing activities to other teachers and students' families; and
- showing evidence of reflecting on these activities through the completion of their portfolio.

As per STIR's data, in Delhi, on average across networks, 25% of participating teachers received the certificate in Year 1 and 28% did in Year 2. In U.P., certification rates differed across the two included districts. In Rae Bareli, 48% of participating teachers received a certification at the end of Year 1 as well as at the end of Year 2. In Varanasi, 30 % of teachers received the certificate at the end of Year 1 and 36% received the certificate and in Year 2.

We illustrate our more detailed understanding of STIR's theory of change and operations in Appendix A1 in the *Report Appendix*. It is based on extensive engagement (including a review of materials, discussions, and workshops including the materials in Appendix A4) with STIR and attempts to reflect both Years 1 and 2 of programming. It also includes engagement with the literature on behavior change, education, adult education, and the production function of student learning outcomes.

Table 1: Description of planned program components in Teacher Changemaker Journey

Standard: Delhi private	Exploratory: Delhi private	Standard: U.P. government	Exploratory: U.P. government
Year 1			
Monthly meetings in collaborative communities of practice, with training in professional mindsets and behaviors. Led by Education Leaders (STIR staff).	Monthly meetings in collaborative communities of practice, with training in professional mindsets and behaviors. Led by Education Leaders (STIR staff).	Monthly meetings in collaborative communities of practice, with training in professional mindsets and behaviors. Led by Education Leaders (STIR-volunteer teachers).	Monthly meetings in collaborative communities of practice, with training in professional mindsets and behaviors. Led by Education Leaders (STIR-volunteer teachers).
Teachers complete reflective portfolios (workbooks) to plan and review classroom changes (micro-innovations)	Teachers complete reflective portfolios (workbooks) to plan and review classroom changes (micro-innovations)	Teachers complete reflective portfolios (workbooks) to plan and review classroom changes (micro-innovations)	Teachers complete reflective portfolios (workbooks) to plan and review classroom changes (micro-innovations)
AwaazDe© call to teachers to share information on upcoming activities	AwaazDe© call to teachers to inform teachers about upcoming activities and recognizing them for their efforts	AwaazDe© call to teachers to share information on upcoming activities (less frequent than in Delhi private schools)	AwaazDe© call to teachers to share information on upcoming activities (less frequent than in Delhi private schools)
WhatsApp© group formed to share upcoming activities, follow-up from meetings, discussion of classroom practice	WhatsApp© group formed to share information related to specific extrinsic, non-financial motivators (elaborated below)	WhatsApp© group formed to share upcoming activities, follow-up from meetings, discussion of classroom practice	WhatsApp© group formed to share information related to specific extrinsic, non-financial motivators (elaborated below)
Teachers work to form In-School Innovation Teams to influence practice of other teachers	Teachers work to form In-School Innovation Teams to influence practice of other teachers	Teachers work to form In-School Innovation Teams to influence practice of other teachers	Teachers work to form In-School Innovation Teams to influence practice of other teachers
Teachers visit parents of five under-performing students to influence families to ensure that student attendance, home work, and checking student progress	Teachers visit parents of five under-performing students to influence families to ensure that student attendance, home work, and checking student progress	Teachers visit parents of five under-performing students to influence families to ensure that student attendance, home work, and checking student progress	Teachers visit parents of five under-performing students to influence families to ensure that student attendance, home work, and checking student progress
EL reports progress to Head Teacher	EL reports on program to Head Teacher		
Education Leader visits teacher in classroom to observe, offer tips	Education Leader visits teacher in classroom to observe, offer tips		
	Treatment schools and/or teachers receive one of four possible non-financial extrinsic motivators (separate from network meetings): • 'Local recognition' (poster of teacher in school, letter home to family, community day celebration) • 'Head Teacher recognition' (skills for Head Teachers) • 'Teacher exposure' (teachers travel to other schools to see peers) • 'Career and personal development' (teachers receive English training)		Treatment schools and/or teachers receive one of three possible non-financial extrinsic motivators (separate from network meetings): • 'Local recognition' (poster of teacher in school, letter home to family, community day celebration) • 'Government and policy exposure' (teachers meet with Block Education Officer) • 'Teacher exposure' (teachers travel to other schools to see peers)
Teachers eligible to receive Roehampton Certificate	Teachers eligible to receive Roehampton Certificate	Teachers eligible to receive Roehampton Certificate	Teachers eligible to receive Roehampton Certificate

Standard: Delhi private	Exploratory: Delhi private	Standard: U.P. government	Exploratory: U.P. government
Year 2			
Monthly meetings in collaborative communities of practice, with non-pedagogical training on classroom culture and practice; teachers select classroom changes from a menu of evidence-informed options related to six 'principles of good classroom practice and culture. Led by Education Leaders (STIR staff)	Monthly meetings in collaborative communities of practice, with non-pedagogical training on classroom culture and practice; teachers collaboratively innovate to operationalize six 'principles of good classroom practice and culture.' Led by Education Leaders (STIR staff)	Monthly meetings in collaborative communities of practice, with non-pedagogical training on classroom culture and practice; teachers select classroom changes from a menu of evidence-informed options related to six 'principles of good classroom practice and culture. Led by Education Leaders (cascade-trained volunteer teachers)	Monthly meetings in collaborative communities of practice, with non-pedagogical training on classroom culture and practice; teachers collaboratively innovate to operationalize six 'principles of good classroom practice and culture.' Led by Education Leaders (cascade-trained volunteer teachers)
Teachers complete reflective portfolios (workbooks) to plan and review classroom changes (micro-innovations), which emphasize the importance of evidence-informed practices	Teachers complete reflective portfolios (workbooks) to plan and review classroom changes (micro-innovations), which emphasize the importance of collaboratively built practices	Teachers complete reflective portfolios (workbooks) to plan and review classroom changes (micro-innovations), which emphasize importance of using evidence-informed practices	Teachers complete reflective portfolios (workbooks) to plan and review classroom changes (micro-innovations), which emphasize the importance of collaboratively built practices
'Local recognition' (poster of teacher in school, letter home to family), highlighting teacher's use of evidence-informed practices	'Local recognition' (poster of teacher in school, letter home to family), highlighting teacher's use of collaboratively developed practices		
Education Leader reports on program to Head Teacher	Education Leader reports on program to Head Teacher		
Education Leader visits teacher in classroom to observe, offer tips	Education Leader visits teacher in classroom to observe, offer tips		
WhatsApp© group formed among teachers to share meeting dates, to highlight importance of evidence-informed practices, and to share classroom practice ideas	WhatsApp© group formed among teachers to share meeting dates, to highlight importance of collaboratively informed practices, and to share classroom practice ideas	WhatsApp© group formed among teachers to share meeting dates, to highlight importance of evidence-informed practices, and to share classroom practice ideas	WhatsApp© group formed among teachers to share meeting dates, to highlight importance of collaboratively informed practices, and to share classroom practice ideas
AwaazDe© call to teachers inform teachers about upcoming activities and highlight importance of evidence-informed practice.	AwaazDe© call to teachers to inform teachers about upcoming activities and highlight importance of collaboratively built and locally tailored solutions.	AwaazDe© call to teachers to inform teachers about upcoming activities and highlight importance of evidence-informed practice	AwaazDe© call to teachers to inform teachers about upcoming activities and highlight importance of collaboratively built and locally tailored solutions
Teachers eligible to become Changemaker Fellows	Teachers eligible to become Changemaker Fellows	Teachers eligible to receive Roehampton Certificate	Teachers eligible to receive Roehampton Certificate
		Education Leaders receive monthly coaching calls from STIR staff	Education Leaders receive monthly coaching calls from STIR staff

4 Evaluation objectives, questions, approach and methods

4.1 Objectives

4.1.1 Evaluation questions

The overall objective of the evaluations is to help STIR understand the extent to which their programming affects teacher professional mindsets and behaviors, classroom practice, and student learning outcomes. We have three guiding evaluation questions and a set of estimates that we produce for each.

1. What is the causal effect of two years of STIR communities of practice on teacher professional mindsets and behaviors?
2. What is the causal effect of two years of STIR communities of practice on the quantity and quality of teaching practices?
3. What is the causal effect of two years of STIR communities of practice on students' Hindi and math learning levels?

Note that this study is not powered to estimate the causal relationships between these outcome sets. More details on the outcomes and the data collection follow further in Section 4.

4.2 Evaluation design

4.2.1 Randomization of STIR communities of practice among schools

We randomly assigned the offer of STIR communities of practice to selected schools; in selected schools, teachers can then apply to join STIR and, if selected, opt to participate actively in STIR communities of practice. Here we give a brief overview of our randomization strategy in both Delhi and in U.P. to show which schools were selected to be invited to STIR communities of practice. Additional details, including visualization, of the randomization are in Appendix A5.

4.2.1.1 Delhi private schools

At the outset, 180 private schools in Delhi met STIR's qualifying criteria and expressed sufficient interest in STIR communities of practices. These 180 schools formed the evaluation's sample. Randomization then proceeded in two stages. The goal of the random assignment procedure was to create three groups of schools equivalent, on average, on pre-program characteristics while avoiding cross-school spillovers and maintaining the geographic patterns required to help STIR build communities of practice: a control group and two treatment groups. The following steps document the randomization procedure used in Delhi:

- Schools were first grouped into 7 (roughly) equally sized strata based on geography.
- Within each stratum, schools were randomly assigned to receive the intervention or to be part of the control group in a 2-to-1 ratio, so that two-thirds of the schools were offered STIR communities of practice and one-third were not.[22]
- Within each stratum, schools assigned to treatment were manually grouped into four smaller geographic clusters. The geographic clusters became the cross-school networks for STIR communities of practice. These cross-school networks only included treatment schools. Two of

[22] In the report, we will refer to the schools who are not offered the program as the control group.

these clusters of schools in each stratum were randomly assigned to receive the STIR standard program. The remaining two clusters received one of the exploratory programming options being trialed in Delhi. This meant there were seven strata, each of which contained two standard STIR communities of practices and two exploratory STIR communities of practice, as well as control schools. Schools were geographically clustered to avoid forcing teachers to travel unnecessary distance to attend network meetings, which could be an obstacle to attendance

We account for this two-stage randomization in our analyses by a) including strata dummies for all analyses and b) clustering our standard errors at the geographic cluster level (for treatment schools) when analyzing results for the standard or exploratory arms.

In Delhi, unlike in U.P., the control group received a one-time placebo treatment provided by STIR appointed staff, consisting of a newspaper subscription, health check-ups, and yoga classes. These interventions were offered to ensure the cooperation of the control schools during data collection[23].

4.2.2.2 U.P. government

In U.P., we worked in two districts: Rae Bareli and Varanasi. Public schools in U.P. are grouped into administrative units called clusters. We considered all clusters with 15 or more schools as part of our potential sample[24]. From these clusters, we randomly selected 16 clusters across the two districts (9 in Rae Bareli and 7 in Varanasi) for inclusion in the study.

Randomization then occurred in two stages.

- First, within each cluster, schools were assigned either to treatment or control at a ratio of 2-to-1. That is, two-thirds of the schools were offered STIR communities of practice and one-third were not.
- Second, all treated schools in a cluster were assigned to one variation of STIR communities of practice (standard or exploratory).

We account for this two-stage randomization in our analyses by a) including strata dummies at the school cluster level for all analyses and b) clustering our standard errors at the school cluster level (for treatment schools) when analyzing results for the standard or exploratory arms.

4.3 Outcome measures and survey instruments

4.3.1 Overview of outcomes and outcome families

In this section, we review our outcomes of interest and discuss the instruments we used to collect measurements on them. Revisions and updates for the endline, as well as a discussion of how our instruments developed, are in the cited appendices.

[23] These activities took place one a year (with the yoga class taking place only once in two years). We do not expect them to impact any of our key outcomes. Furthermore, we cluster our standard errors at the geographic cluster level (by treatment arm) to account for the within treatment clustering.

[24] Clusters with 15 or more schools were selected keeping in mind logistical ease for program implementation and data collection. Schools were either Primary or Upper Primary schools.

We summarize what we measure and how in Table 2. In the first column, we present the overarching concept. These concepts represent the outcome families we use in our evaluation design. Outcome families comprise several measures of a broader overarching concept that is not easily captured by a single indicator. The purpose of this grouping is twofold. First, if STIR communities of practice influence the broader concept, we would expect to observe this for most of the measures of the outcome family. Second, defining outcome families in this way facilitates the correction of statistical inference for asking many similar questions of the data, as detailed in Section 4.5.3.

In the middle column, we present the specific items that we measured in each outcome family, described in more detail below. Under 'measurement strategy,' in the third column, we clarify how we collected the relevant data. Finally, in the right-most column, we clarify the portion of our sample from which we collected data.

Teacher professional mindsets and behaviors

Our first outcome family is of teacher professional mindsets and behaviors (please refer to Appendix A6 for more details). Although for the baseline and midline teacher surveys we used tools designed by IDinsight to measure teacher motivation, for the endline teacher survey, we used a questionnaire created by a team at New York University (NYU) led by Dr. Ed Seidman. STIR felt that this new teacher report survey well-captured the different elements of professional mindsets and behaviors. The questionnaire ('teacher report survey'), was initially used to test the effect of STIR communities of practice on a broader set of professional mindsets and behaviors in Uganda. It was later adapted to and validated in the Indian context by STIR and NYU in late 2016 with the help of focus group surveys. The self-administered questionnaire consists of 46 items, each scored on a 6-point Likert scale. We present these results as an overall index score of all 46 items weighted equally. We also present results for three sub-indices — the derivation of these indices is discussed in Section 5. The three indices are: growth mindset, positive professional outlook, and efficacy. Data for this measure were collected from all teachers in each school.

Classroom practice: quantity

Our second outcome family is quantity of teaching practice. To capture quantity of practice, we used a modified version of the Stallings snapshot tool (Stallings 1977; World Bank 2015) to quantify observed instructional time spent on one of three mutually exclusive categories of activity: teaching, classroom management, or off-task (please refer to Appendix A7 for more details and to Appendix A8 for the endline classroom observation tool). We randomly selected, on average, three teachers in each school for classroom observations. In these classrooms, enumerators sit or stand in the back and code student and teacher activities at intervals of three minutes each for a total of seven snapshots. For this measure, we randomly sampled a subset of teachers in each school for observation (on average, three per school). It is imperative to note up-front that the Stallings instrument is ideally deployed when enumerators can access the classroom at the time the lesson is supposed to start, regardless of teacher presence; time in the classroom before the teacher entered would be categorized as off-task. However, we were rarely able to access classrooms in this way; we needed to wait for a teacher in order to enter. We present results for time teaching and time off-task, with classroom management as the omitted category.

Classroom practice: quality

To capture quality of practice, we measure seven indicators of child-friendliness — six developed by the ASER Centre and one (calling students by their names) developed by IDinsight based on our understanding of STIR communities of practice (NCERT 2005; S. Bhattacharjea, Wadhwa, and Banerji 2011; Suman Bhattacharjea 2017). The version of the endline observation tool for child-friendliness is provided in Appendix A8. For this set out outcomes, we measured the same randomly selected subset of teachers in each school as for quantity of classroom practice. Enumerators sit or stand in the back of classrooms and code student and teacher activities at intervals of three minutes each for a total of seven snapshots. The seven indicators of child-friendliness we capture are:

- Whether teacher smiled, joked, or laughed
- Whether students asked at least one question of the teacher
- Whether teacher incorporated local information into teaching
- Whether the teacher made use of learning aides
- Whether the teacher had the students working in pairs or small groups
- Whether the teacher praised a student or showed-off students' work
- Whether the teacher called any students by their names

Student learning outcomes

To capture learning levels, we used the ASER[25] learning assessment tool for Hindi (local language) and math (please refer to Appendix A9 for more details). The ASER tool is widely established and popular for capturing student learning levels in India. While the ASER tools are administered in the ASER Centre's national survey work to children aged 5 to 16, we added additional levels (additional stories in Hindi and a fractions section in Math) to prevent ceiling effects (ASER Centre 2018). More information on the tools can be found in Appendix A9 and the tools used during endline can be found in Appendix A10. To capture learning levels, we selected a random subset of students from the (baseline) classrooms of those teachers selected for classroom observation; we selected ten students per classroom.

Table 2: Measures, measurement strategy, and measured sample

Overarching concept	Specific construct	Measurement strategy	Sample
Professional mindsets and behaviors	Self-assessed score	Questionnaire completed by teachers on different facets of professional mindsets and behaviors (Appendix A11)	All teachers
	Self-assessed score from 'Positive professional outlook' sub-index[26]		
	Self-assessed score 'Teacher growth mindset' sub-index		
	Self-assessed score from 'Efficacy' sub-index		
	Time spent teaching		

[25] http://www.asercentre.org/p/141.html

[26] The sub-indices were named by NYU after a subjective assessment of the items. While the items themselves were chosen on the basis of the data, the names attributed to the indices were not driven by the data.

Classroom practice: quantity[27]	Time spent off-task	Observation of teacher practice, using a modified Stallings instrument (Stallings 1977) (Appendix A8)	Sample of teachers within schools (Classroom practice sub-set of all teachers[28])
Classroom practice: quality	Whether teacher smiled, joked, or laughed	Observation of teachers and students using ASER child-friendliness indicators (S. Bhattacharjea, Wadhwa, and Banerji 2011) (Appendix A8)	Sample of teachers within schools (Classroom practice sub-set of teachers)
	Whether students asked at least one question of the teacher		
	Whether teacher incorporated local information into teaching		
	Whether the teacher made use of learning aides		
	Whether the teacher had the students working in pairs or small groups		
	Whether the teacher praised a student or showed-off students' work		
	Whether teacher called any student by their name.		
Student learning	Hindi competency	Assessment of student learning using modified ASER learning assessment tool ("Annual Status of Education - Rural" 2005) (Appendix A9 and Appendix A10)	Sample of students within schools (Student learning sub-set of students)
	Math competency		

[27] For this measure, teacher time could be categorized as teaching, managing the classroom, or time off-task.

[28] Power calculations indicated a target of three teachers per school on average. The sample list was created by sub-setting the teacher motivation sample list (list of all teachers in the school).

4.4 Sampling and data collection

4.4.1 Programmatic and evaluation timeline

The academic year in India begins in April and includes both a long summer and a shorter winter holiday. STIR's programming is designed to align with the academic year, such that the Teacher Changemaker Journey evaluated here started in April 2015. To prepare for this, STIR's taster sessions as well as sampling took place in late 2014, followed by data collection starting in January 2015. Randomization took place in April 2015, just before the new school year.

During data collection, we administer the teacher professional mindsets and behaviors questionnaire (PMB) separately from measuring classroom practice (CP), and student learning (SL) survey and measured each of the three outcomes at different points in time. Note that for all data collection activities, the data collectors were blinded to whether they were visiting intervention or control schools and whether the specific teachers with whom they were speaking were active in STIR. For more information on the timing of program implementation and data collection efforts see Table A8 in Appendix A12. In Figure 2 and, below

Figure 3, show the timelines for our evaluations and school terms in the two geographies, including that the endline in U.P. was delayed due to elections.

Figure 2: Evaluation timelines for Delhi

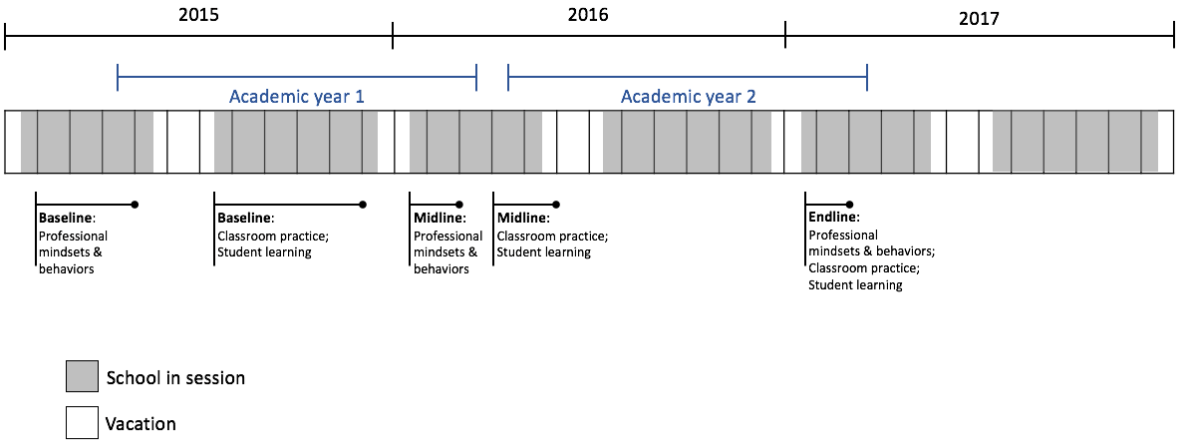
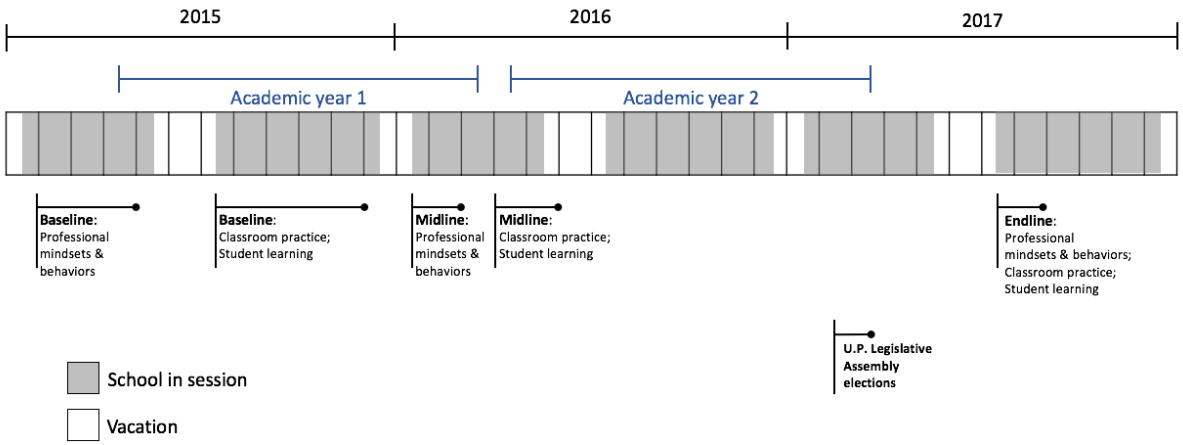


Figure 3: Evaluation timelines for U.P.



4.4.2 Delhi private schools: sampling strategies and baseline and endline samples

In this section, we describe baseline and endline sampling for our different data collection needs. This is summarized in Table 3 and described in more detail following the table. Please find additional sampling details in Appendix A12.

Table 3: Delhi targeted and actual samples

	Teachers for professional mindsets and behaviors questionnaire	Teachers for classroom practice observation	Students for learning outcomes
Target baseline sample	All teachers in STIR and control schools in our sample (no fixed number)	811	8110
Total number of units sampled at baseline	1249	342	3367
Population sample is representative of	All teachers from sample schools	The 811 teachers were all the teachers, in both treatment and control, who expressed interest in STIR by attending an initial taster session. (There were	10 students randomly selected from the main class in which each teacher observation was performed. Students could range from 1st to 8th standard.

		approximately 438 teachers that they did not target. Note that targeting happened prior to randomization.)	
Reason for difference between target and actual sample at baseline	We revisited schools a maximum of five times to ensure all teachers were surveyed. There may be minor differences between total teachers in the schools and teachers we surveyed, but based on our survey tracking we can safely conclude there are no significant differences.	The difference between the target and final number of teachers surveyed was partly due to school level refusals and partly due to teacher attrition (either because the teacher had transferred or refused to participate in the survey).	Due to school and teacher refusals we were unable to sample students from some classes. In addition, some classes had fewer than 10 students in total.
Timeline for baseline data collection	February to April 2015	July to November 2015	July to November 2015
Target endline sample	All teachers in our sample schools	All teachers from the 811 list. If a school has fewer than 2 teachers left from this list, randomly select one or two teachers from among those teachers who were present as on 1st July 2015 but not included in list of 811.	All 3367 students surveyed at baseline
Units attrited since baseline	734	125	1523
Units added since baseline	557	245	158
Total number of units sampled at endline	1072	462	1846
Population sample is representative of	All teachers from our sample schools	All teachers targeted by STIR and still present at the study school. (Plus adding some teachers to the list.)	All students taught by a STIR targeted teacher at baseline still studying in the school at endline. Students could range from 1st to 8th standard.
Reason for difference between target and actual sample at endline	School and teacher refusals, teacher dropouts, and teachers not being available during the data collection window (generally due to a long leave of absence).	School and teacher refusals, teacher dropouts and teachers not being available during the data collection window (generally due to a long leave of absence).	School refusals, students moving to other schools, graduating or dropping out and being absent through the course of the data collection period.
Timeline for endline data collection	January to February 2017	January to February 2017	January to February 2017

Professional mindsets and behaviors (PMB) survey:

We attempted to administer the teacher motivation survey (the precursor to PMB) to all teachers at baseline. A total of 1249 teachers were surveyed after (a maximum of) five visits to the schools. At endline, we again attempted to survey all the teachers in our sample schools by offering the survey to each teacher in the school. The total number of teachers completing the endline teacher PMB questionnaire (for whom baseline data are also available) was 514 (48%). Since all teachers could fill out this questionnaire, even those who joined a school since baseline, we do not have baseline data for all teachers in this sample. All the teachers for whom we have endline data form the sample used for analysis. Please refer to Appendix A12 for details of teacher dropouts.

Classroom practice (CP) survey:

STIR targeted a total of 811 teachers in both treatment and control schools for participation in the program based on interest expressed during a 'taster session' conducted by STIR, which introduced the program to the teachers and took down names of those expressing interest[29]. These 811 initially interested teachers formed the potential sample for classroom practice at baseline. However, in the gap between assessing interest and administering the baseline survey for classroom practice (please refer to Figure 2), we experienced attrition from school refusals and teacher dropouts; the total number of teachers for which classroom practice data were collected was 342[30]. At midline, we again returned to the list of 811 teachers as our target sample. For those schools where the number of teachers available from our 811 list fell below two, new teachers were added based on a random selection from those teachers employed at that school as of 1 July 2015[31]. In total, we ended up observing classrooms of 459 teachers in 143 schools. Among the 459 teachers observed, 311 teachers were from our original list of 811 teachers. The remaining 148 were added on-the-spot. During endline, we used the 459-teacher list from midline as our main sample and added teachers[32] in case of drop outs or in cases where there was not even a single teacher present in the school from our sample. We surveyed 462 teachers in total out of which we have baseline data for 221 (48%) teachers. Please see Appendix A12 for details of dropouts.

Student learning (SL) survey:

To test student learning, 10 students were randomly selected from all the students in the main class that the teacher taught[33]. Students ranged from 1st to 8th standard. Thus, the full target sample for Delhi included 811 teachers (and classrooms to observe) and 8110 students to test for learning levels. Due to the attrition from school refusals and mainly teacher dropouts, the total number of teachers for which classroom practice data were collected was 342 as mentioned above. For these 342 teachers, a total of 3367 students were tested. All students surveyed at baseline formed the potential sample at midline. Among the 3367

[29] A total of 439 teachers from the Delhi TM baseline list were not targeted by STIR due to lack of interest in joining the program, as (not) expressed during the taster sessions. Note, the targeting based on taster sessions happened prior to randomizing schools into treatment and control.

[30] Amongst these 9 teachers were not considered to incomplete surveys and thus had to be dropped. The final number for analysis is 333

[31] Teachers who were in schools as of 1 July 2015 would have been exposed to the program from its start.

[32] New teachers were added based on a random selection from those teachers employed at that school as of 1 July 2016.

[33] A teacher's 'main class' or primary class was defined as the class in which s/he spent the maximum time during the week or were 'class teachers' for. Class teachers have extra administrative responsibilities with respect to their class/grade, such as taking attendance.

students from baseline, 1956 students were tracked and surveyed at midline. Out of the 1956 students surveyed at midline, we were able to successfully survey 1846 during the endline. Please see Appendix A12 for details of dropouts.

4.4.3 U.P. government schools: sampling and baseline and endline

In this section, we describe briefly the baseline and endline sampling for our different data collection needs. This is summarized in Table 4. Please find additional details in Appendix A12.

Table 4: U.P. targeted and actual samples

	Teachers for professional mindsets and behaviors questionnaire	Teachers for classroom practice observation	Students for learning outcomes
Target baseline sample	All teachers in STIR and control schools in our sample (no fixed number)	810	8100
Total number of units sampled at baseline	1145	838	7386
Population sample is representative of	All teachers from our sample schools	On average, 3 teachers were randomly selected from each of the 270 STIR and control schools from the list generated during the teacher motivation survey. There were dropouts and additions to our lists to arrive at 838 teachers	10 students randomly selected from the main class in which each teacher observation was performed.
Reason for difference between target and actual sample at baseline	We revisited schools a maximum of three times to ensure all teachers are surveyed. There may be minor differences between total teachers in the schools and teachers we surveyed; but based on our survey tracking we can safely conclude there are no significant differences	There were dropouts and additions to our lists to arrive at 838 teachers	There were often multiple teachers who taught the same cohort of students; 282 classrooms had fewer than 10 students
Timeline for baseline data collection	February to March 2015	July to September 2015	July to September 2015
Target endline sample	All teachers in the schools in our sample	All 747 teachers surveyed at midline.	All 4560 students surveyed at midline
Units attrited since baseline	563	209	4234

Units added since baseline	551	95	N/A
Total number of units sampled at endline	1133	724[34]	3152
Population sample is representative of	All teachers from our sample schools	All teachers surveyed at baseline and still present at the study school. (Plus adding teachers in cases where all teachers of a school have dropped out.)	All students taught by a STIR targeted teacher at baseline still studying in the school at endline.
Reason for difference between target and actual sample at endline	School and teacher refusals, teacher dropouts and teachers not being available during the data collection window (generally due to long leave of absence).	School and teacher refusals, teacher dropouts and teachers not being available during the data collection window (generally due to long leave of absence).	School level refusals, students moving to other schools or dropping out and being absent through the course of the data collection period.
Timeline for endline data collection	July to August 2017	July to August 2017	July to August 2017

Teacher professional mindsets and behaviors (PMB) survey:

We attempted to administer the teacher motivation survey (the precursor to PMB) to all teachers at baseline. A total of 1145 teachers were surveyed after (a maximum of) three visits to the schools. At endline, we offered the survey to all the teachers in the schools in our sample. The total number of teachers completing the endline teacher motivation questionnaire was 1133 out of which 582 (51%) were present at baseline as well. All the teachers at endline form the sample used for analysis. Please see Appendix A12 for details of dropouts.

Classroom practice (CP) survey:

From each of the 270 schools in our sample, an average of three teachers were randomly selected for observation using the list of 1145 teachers from the teacher motivation baseline. From this list, there were drop-outs and additions. The total number of teachers observed was 838. This was our target for midline classroom practice observations. One teacher was added in schools where all teachers from our 838 list had dropped out. This was done in 13 schools (12 in Rae Bareli and 1 in Varanasi). In total, 747 teachers were surveyed at midline. This formed the sample for endline. Once again, we added teachers in schools where all had dropped out. In total we surveyed 724 teachers at endline, out of which we have baseline data for a total of 629 (86%) teachers. Please see Appendix A12 for details of dropouts.

Student learning (SL) survey:

At baseline, 10 students were randomly selected from all the students in the main class that the teacher was observed at during classroom observation. For these 838 teachers observed during baseline, a total of 7386 students were tested. Of the 7386 students tested at baseline, a total of 4560 students were also tested at midline. These 4560 students formed the sample for endline. We ended up testing a total of 3152 teachers from this sample. Please see Appendix A12 for details of dropouts.

[34] This includes 80 'active' teachers that were added to the sample for the purpose of the observational analysis only. They were excluded for the rest of our analyses as adding them would bias our sample.

4.4.4 Attrition from the sample:

As can be seen from Sections 4.4.2 and 4.4.3, attrition potentially poses a threat to both the evaluations. A quick summary of the attrition numbers – both at the teacher and the student levels are mentioned below in Table 5:

Table 5: Attrition numbers from baseline to endline:

Sample list	Baseline (BL) timeline	Baseline sample	Endline (EL) sample	Endline timeline	Attrited from BL sample	Percentage attrition (BL-EL)
U.P. TM List	Feb-Mar 2015	1145	1133	Jul-Aug 2017	563	49%
U.P. CP List	July-Aug 2015	838	724	Jul-Aug 2017	209	25%
U.P. SL List	July-Aug 2015	7386	3152	Jul-Aug 2017	4234	57%
Delhi TM List	Feb-Apr 2015	1249	1072	Jan-Feb 2017	734	59%
Delhi CP List	July-Nov 2015	342	462	Jan-Feb 2017	125	37%
Delhi SL List	July-Nov 2015	3367	1846	Jan-Feb 2017	1523	45%

At endline, 32 (17%) of Delhi private schools and 19 (7%) of U.P. government schools refused to participate. We also lose teachers from schools that remain in our sample, which can happen for several reasons. In U.P. government schools, about half of our attrited teachers were due to transfers across schools. In Delhi private schools, about half of our attrited teachers had stopped teaching.[35] In addition, teachers also attrit by having a prolonged absence from school, by refusing to participate in the survey and also by retiring or, in rarer cases, dying. By way of comparison, in the Indian state of Andhra Pradesh, Muralidharan and Sundararaman find about 30% of teachers attrit each year (Muralidharan and Sundararaman 2006).

Similarly, students can drop out of our sample for a variety of reasons, ranging from graduating out of primary school, to changing schools, to prolonged illness, to dropping at of school, as well as refusing to participate in our survey. About 20% of our student attrition is due to students graduating.[36] By way of

[35] To provide some sense of why teachers leave, as part of the process evaluation we conducted in Delhi in early 2016, we tried following up telephonically with teachers who had dropped out of private schools between both our two rounds of baseline measurement (from teacher professional mindsets and behaviors to classroom practice). We ended up speaking successfully with 50 teachers. Out of the 50 teachers we spoke to, 38% were no longer working, 22% had moved to teaching in other private schools, 2% were teaching in government schools, 23% had moved onto teaching private tuitions/tutoring, and the remaining had dropped out for other reasons.

[36] We adapted our field protocol to try and maximize the number of teachers and students we captured. A few things we did was to – increase the number of visits per school, work closely with STIR field teams to minimize refusals at the school level, track students and teachers to other schools that were part of our evaluation sample and follow up telephonically with teachers who were absent over a long period of time.

comparison, education studies in India have experienced student attrition ranging from 14% to 25% over the course of two years of study (Banerjee et al. 2007; Muralidharan and Sundararaman 2006; Muralidharan 2012; Linden 2008). Students aging out of primary school is a likely a key driven of difference in the student attrition rates in our study and these other studies.

4.4.4.1 Implications of attrition on the evaluations

Given the high attrition numbers from our samples, we looked closely at the implications that it may have for our evaluations and the results. This section summarizes our findings. For a more detailed understanding and explanation please refer to Appendix A13. We ran the following tests to assess the impact of attrition on our samples:

1. **Tests for differential rates of attrition across the treatment and control schools** by comparing teacher dropout across treatment status (control, standard, exploratory) in our study sites (U.P., Delhi) between baseline and endline survey rounds (Teacher Professional Mindsets and Behaviors, Classroom Practice, Student Learning). If differential attrition was absent, overall trends for attriting teachers and students are expected to be comparable across treatment and control groups. Overall, we do not find evidence of different levels of attrition across treatment and control groups for students or teachers.[37] See Appendix A13 for details.
2. **We test for differential attrition trends by baseline characteristics in a more direct way**. Specifically, we run the following two types of tests:
    I. A test for balance on baseline covariates between treatment and control for those teachers and students who remain in our sample at endline. We do not find evidence of difference in terms of baseline characteristics for teachers or students surveyed at endline at the 5% level of significance. See Appendix A13 for details.
    II. We compare attritors and non-attritors using baseline characteristics to see if they differ systematically between treatment and comparison groups. Based on these comparisons, there is insufficient evidence to suggest that attritors are fundamentally different on baseline characteristics across treatment groups for Delhi and U.P. Again, see Appendix A13 for more details.

In summary, we do not find evidence of differential attrition, which increases our confidence that our results are not confounded by trends in attrition.[38] It may, however, dampen our ability to pick up small but practically meaningful effects. It is further possible that the attrition resulted in imbalance in unobserved teacher characteristics.[39]

[37] We do find one occurrence of differential attrition in one of our samples — dropout rates are significantly different between comparison and standard and exploratory samples for Delhi classroom observation data. However, we do not believe this issue warrants concern given imbalance was found for only one out of several covariates.

[38] Lee Bounds were created during midline analysis to bound the extent of baseline-midline attritors. A decision was made to forgo reporting the bounds as they were too wide to be meaningful. For the same reason, we have not estimated Lee bounds at end line. We also considered re-weighting our sample based on inverse probability weights (IPW), however ultimately decided against it given our limited predictive power to predict dropout. Furthermore, adding IPW weights would likely have limited effect on estimates given the close balance on observable baseline covariates.

[39] To create problems, these unobserved characteristics would either have to be correlated with the observed baseline characteristics that appear balanced or there would need to be a plausible explanation for why observed baseline variables appear balanced while correlated unobserved variables are not.

4.5 General notes on analytic approaches and reporting

In this section, we provide some analytic details that apply across many of the different outcomes we examine.

4.5.1 Measuring the effect of STIR communities of practice: school-wide and teacher-level estimation strategies

To understand the effects of STIR communities of practice on teachers and students, we calculated two broad types of impact estimates: a school-wide estimate (using the *intent-to-treat* estimator) and a teacher-level estimate (using different approaches to approximate the treatment-on-the-treated effect). This study is optimized to produce rigorous school-wide estimates and these are our preferred results. We take the teacher-level effects to be suggestive.

4.5.1.1 School-wide effects

Our measure of STIR's overall causal impact on teacher and student outcomes will be based on the school-wide outcomes, meaning those teachers that completed the professional mindsets and behaviors questionnaire as well as those teachers randomly selected for classroom observation and the students randomly selected in their classrooms.[40] This estimate is traditionally termed *intent-to-treat*; for the remainder of this report, we shall refer to these estimates as school-wide effects. For this estimate, we compare outcomes for teachers in treated[41] schools (schools that received the offer of joining STIR communities of practice) to outcomes for teachers in control schools. This estimation includes teachers in treated schools, both those who were active participants in the program as well as those who weren't. The school-wide estimate, therefore, captures both the direct influence of STIR on active participants as well as the influence on non-active teachers in treated schools. We believe this is the most policy relevant estimate given that STIR communities of practice offers routes to exposure other than being an active participant, such as working with Head Teachers and In-School Innovation Teams.[42] For thinking about what is achievable at scale, we strongly believe that this estimate is the most useful and appropriate estimate of STIR's causal impact on the outcomes of interest.

The key evaluation design feature validating the school-wide estimate is successful random assignment of STIR communities of practice programming to schools (not teachers), with teachers in control schools not receiving any aspect of the Teacher Changemaker Journey.[43] The balance tests on baseline data detailed in Appendix A14 confirm that teachers across all treatment groups in both the geographies are well balanced.

[40] We use the term 'school-wide' to contrast with the effects of STIR on actively participating teachers; however, it may be the case that our selected teachers and students are not fully representative of the teacher or student population in the school, especially after attrition over two years.

[41] In accordance with convention for randomized evaluations, we refer to the offer of STIR's programming as the 'treatment' under investigation.

[42] Note that, operationally, STIR introduced an element of 'rationing' to ensure networks were maintained at a manageable size, such that it is possible that not all teachers who applied and otherwise would have been selected to become active participants were able to do so[42].

[43] Technically speaking, contamination would be present if (some or all teachers in) control schools were exposed to STIR's programming in any way.

4.5.1.2 Teacher-level effects

A second interesting, but more complicated, approach aims to quantify the effect of STIR communities of practice for those teachers that choose to actively participate. This analysis is conventionally termed a *treatment-on-the-treated* (TOT) analysis; we will refer to it as teacher-level analyses for the rest of this report. It is important to emphasize the fact that — by design — active participation in STIR communities of practice is voluntary and subject to a non-random application and selection process. In a model of voluntary participation, we expect teachers who make this decision to be different in important (if unobservable) ways from teachers who ultimately do not participate. Thus, a teacher-level estimate will only help us understand the potential effects of STIR communities of practice among the sub-set of teachers with the characteristics that make them likely to become active participators.

However, given the importance for STIR's internal learning we have looked to estimate the teacher-level effects of STIR communities of practice using two main approaches — an IV/LATE estimate and, at STIR's request, a non-experimental comparison of active and non-active teachers which we refer to as the "observational analysis." More details on these approaches along with the limitations are mentioned in Section 4.6. We urge caution in the interpretation of the magnitude of the IV/LATE estimate, which represents an upper bound on the possible effect size. We urge particular caution in any interpretation of the observational analysis; one should not interpret the results as implying a causal relationship between teacher participation and outcomes.

Note that we conduct teacher-level analysis only for the teacher outcomes, not students. It is difficult to conceptualize an active or exposed student in this program and evaluation. Students move through different grades through the course of the evaluations; they may/not be taught by teachers who are active participants of the STIR program. Further, in schools where students are taught by multiple teachers, it is tough to clearly know the extent to which they were exposed to an actively participating teacher.

Estimating the teacher-effects among participating teachers: IV/LATE

Broadly, any *treatment-on-the-treated* analysis aims to isolate the effects on those who comply with the program. In this case, we aim to isolate the effect of STIR communities of practice for only those teachers who actively participated in STIR. An instrumental variable (IV) offers a strategy to isolate this effect. For this analysis, we define active participation as having attended at least one meeting[44] in both Year 1 and Year 2.

This analysis gives some insight into the causal effect of STIR communities of practice specifically for active teachers.[45] If we assume that for those teachers in STIR schools who did not ever participate in a meeting, there is still some positive benefit (through lessons shared formally and informally among teachers within schools), the instrumental variable estimate provides an upper bound of the treatment-on-the-treated

[44] We did not count the opening 'taster' session as a meeting for this definition.

[45] Note that we do not make use of this analysis for student-level outcomes.

effect.[46,47] Given this, we believe that the true teacher-level effect will lie somewhere between the school-wide estimate and the teacher-level estimates generated using instrumental variables.

Whether we can make a valid claim about STIR's causal impact in this case rests on a key assumption: teachers in schools offered to join STIR communities of practice (treated schools) but who did not individually participate in any STIR meetings were unaffected by the program. Due to the explicit focus on sharing learnings from STIR teachers with non-STIR teachers, this assumption is unlikely to hold in practice. If we assume that all within-school spillover benefits are positive, the estimated treatment-on-the-treated effect represents an upper-bound on the teacher-level effect.

Estimating the effects among participating teachers: Observational analysis

On STIR's request, we also directly compare teachers in the treatment group who participated actively in STIR with teachers in the control group, excluding teachers in the treatment group who didn't participate in STIR communities of practice. For this analysis, we drop all data from teachers in STIR schools who did *not* participate actively in STIR. We then directly compare the participating teachers in STIR schools to all the teachers in control schools. This analysis is an attempt to identify a non-causal relationship between participating in different amounts of STIR communities of practice (*i.e.*, the proportion of total STIR meetings attended) and the outcomes.

For this endline analysis, we use three definitions of 'active participation'. These are: (1) Active defined as teachers attending at least one network meeting in Year 1 and one network meeting in Year 2 (excluding the introductory taster meeting), (2) Active defined as teachers attending at least half the meetings[48] over the two years (excluding the taster meeting), and (3) Active defined as teachers attending at least three-fourths the meetings over the two years (excluding the taster meeting)[49].

To interpret the relationship between teacher attendance and outcomes as causal is not valid. For the causal interpretation to be valid, teachers who participate in STIR must be similar to teachers who did not participate in STIR. In other words, we must assume that teachers who participated in the STIR program (whether they volunteer to join or were selected by school leadership to apply) would have had similar outcomes to the control group teachers if they hadn't participated in the program. However, due to significant personal initiative teachers must demonstrate to join and participate in STIR, **we believe this assumption is unlikely to be true**. As the definition is tightened (*i.e.*, requires that a teacher have attended more meetings), it becomes more difficult to assume that selection bias is not at play (that is, it becomes less likely that the assumption of no meaningful difference between active and inactive teachers is true).

[46] This is for estimates that are positive. For negative estimates, we may consider the estimate a lower bound if we assume the effect on teachers who didn't participate is also negative.

[47] It is critical to note that the result will only apply to teachers of a similar type who might be willing and able to join STIR programming in a new school which is offered; it does not provide a good guide to the expected effect of, say, making STIR mandatory for all teachers in a school.

[48] Given the total of eighteen and sixteen meetings in U.P. and Delhi respectively, teachers who attended more than 9 meetings in U.P. and more than 8 in Delhi would be classified as active

[49] Given the total of eighteen and sixteen meetings in U.P. and Delhi respectively, teachers who attended more than 14 meetings in U.P. and more than 12 in Delhi would be classified as active

4.5.2 Subgroup analyses

For each of the analyses conducted, we consider four main sub-groups (separately for each context) in which we hypothesized we might see heterogeneous treatment effects. Finding 'heterogeneous treatment effects' would mean that STIR communities of practice is differentially effective for different types of teachers. Three of these subgroups — split by teacher sex, teacher years of experience, and teacher baseline motivation levels — were identified before we began analyzing the data.[50] We did not have explicit priors about the direction of influence of these subgroups; for example, we thought there were compelling reasons why male teachers may be able to gain more from STIR communities of practice but equally compelling reasons why this might be the case for female teachers (as detailed in Appendix A1).

In Appendix A15, we provide details on number of teachers per subgroup category.

The final subgroup, dividing the U.P. analysis by administrative blocks, was added after we saw the initial midline results at the request of STIR. STIR thought this would be particularly useful in trying to separate out the influence of program design versus implementation capacity and delivery context, with the hypothesis that administrative units with more supportive BEOs and other local officials would show stronger results. Please see Appendix A15 for details of schools in each block.

Subgroup analysis was only conducted for all treatment schools versus control schools (*i.e.*, not looking separately at standard *v.* exploratory) due to sample size considerations.

When interpreting results from the sub-group analyses we focus less on individual estimates and more on overarching trends; *e.g.*, if STIR communities of practice seems to have a differential impact for all 'low-experience teachers' across all estimates. It would be tough to imagine a situation (and a sound theoretical narrative) where STIR's program would have a differential impact on say female teachers for one of the indicators of the child-friendliness family but not others; or more generally if female teachers are able to influence child-friendly practices within their classrooms but not time-use practices.

[50] We pursued these three sub-groups for the following reasons:

- Teacher baseline motivation —The baseline teacher motivation is used to test whether there is a difference in the impact of treatment across different levels of motivation in teachers. Teachers who are initially more motivated may be more driven to be an active participant in the STIR program. They may also be naturally more eager to adopt what they learn via network meetings in their classrooms. For STIR to achieve their long-term targets it is important that they successfully impress upon 'not so' motivated teachers as well.
- Teacher sex — While in Delhi more than 90% of the sample of teachers is female, in Uttar Pradesh the proportion of male and female teachers in the sample are similar; we only consider this sub-group for U.P. Whether their programing has differential impact for male and female teachers has always interested STIR. Male and female teachers may experience differential effects, given differing incentives and constraints in participating actively in STIR and being able to enact ideas from STIR in the classroom (see Appendix A1).
- Teacher experience — Several researchers have found a transformation from a novice (or rookie) teacher into a teacher with 'more experience' after 3 years of having been a teacher (Araujo et al. 2016; Staiger and Rockoff 2010; Rivkin, Hanushek, and Kain 2005). More experienced teachers may be more set in their ways, and therefore less willing to act on STIR's approach, but they may also be more in need of 're-motivation' and may also be better placed to put STIR's ideas into action.

4.5.3 Multiple hypothesis testing and corrections

In this evaluation, we examine several outcomes (grouped into families) and specifications in accordance with exploring different aspects of the theory of change. Given that we are asking many questions (or testing many hypotheses) we correct for multiple hypothesis testing at the family level in cases where we find statistically significant results. For those interested in refreshing their understanding of statistical inference and the potential for false positives, please see Appendix A16 and for a more detailed discussion on multiple hypothesis correction please refer to Appendix A17.

We use two main approaches to multiple hypothesis testing, depending on the specification we are looking at: the Free Step Down Resampling Method (FSDRM) and the Holm-Bonferroni method. While we consider FSDRM the most powerful we are unable to use this method for comparing standard *v.* control (C) and exploratory *v.* C, given how schools were randomized.[51] Please see Appendix A17 for details. We use a combination of FSDRM and Holm-Bonferroni corrections, as described below.

- **All-STIR (standard and exploratory taken together) *v.* C** — For the analysis taking all treatment schools together (reported as All-STIR) we have used the Free Step Down Resampling Method (FSDRM) when relevant[52]. This is true for the main school-wide estimate as well as the sub-group estimates.
- **Standard *v.* C** — For all analysis where we are comparing purely the schools receiving the standard STIR model to control schools we use the Holm-Bonferroni correction when relevant.
- **Exploratory *v.* C** — For all analysis where we are comparing purely the schools receiving the exploratory STIR model to control schools, we will use the Holm-Bonferroni correction when relevant.

4.5.4 Other notes

All analyses presented below are done using Stata (*STATA* (version 14.0), n.d.). The analysis is conducted separately for Delhi private schools and U.P. government schools, given the contextual, implementation, and programmatic differences between the two settings.

4.6 Analytical models and specifications

4.6.1 School-wide estimates

We use an analysis of covariance (ANCOVA) model to estimate the school-wide effect of the STIR program (McKenzie 2012).[53] The specifications mentioned have been fit separately for teachers from Delhi and Uttar Pradesh. We employ the following specification:

[51] In addition, we do not use the FSDRM to correct multiple inference for the All-STIR IV/LATE results.

[52] Corrections are only relevant in certain cases depending on the number of hypotheses being tested in each family. For instance, we do not correct teacher motivation results as there are few outcomes (4) being tested within each family. Corrections are also only relevant when uncorrected results have one or more significant treatment estimates.

Note that for the observational analysis, we did not conduct multiple hypothesis correction given our skepticism on the validity of these results.

[53] While the randomized design (and primary indicator of treatment *v.* control) allows us to adapt our tools through the course of the evaluation, it does have implications for the analysis. We are limited in our ability to purely compare

$$Y_{1ij} = \sum\alpha_s + \beta_1 * T_j + \beta_2 * Y_{0ij} + \gamma * X_{ij} + \omega_j + \varepsilon_{ij}$$

where,

- Y_{1ij} is an individual teacher's or student's (belonging to school j) outcome at endline
- $\sum\alpha_s$ are strata fixed effects[54]
- Y_{0ij} is an individual teacher's or student's (belonging to school j) outcome at baseline
- T_j is a binary variable for treatment assignment of the school the teacher or student belongs to (which represents pooled treatment *i.e.*, standard and exploratory clubbed together, standard treatment, or exploratory treatment, depending on the regression)
- X_{ij} is a vector of covariates[55][56]. At the teacher-level these include teacher sex, age, qualification, years of experience, baseline teacher motivation, class size, enumerator and network dummies. At the student-level these include student grade, sex, class size, teacher experience, teacher age, teacher sex, teacher qualification, enumerator and network dummies.
- ε_{ij} is an individual level (within schools) error term
- ω_j is a school level error term (except in the case of analyses comparing just the standard or exploratory arms in which case it is a cluster-level error term for treatment schools and a school level error term for control schools)

β_1 will be our estimate of interest (effect size). The standard errors are clustered at the school level.[57] The above specifications would be run for three main treatment (assignment) types. We have broken these into three separate regressions to more closely adhere to the way the original research questions were defined and to ease the interpretations of the results:

1. All-STIR *v.* C: standard (core + selection) and exploratory (core-plus + co-creation) will be clubbed, so that the two-year accumulated effects of STIR communities of practice can be seen.
2. Standard *v.* C
3. Exploratory *v.* C

Having discussed the generic specification, we will now discuss if and how each family of indicators were analyzed using the above.

pre-post values and undertake a 'difference in differences' analysis. Baseline indicators are used as covariates in an ANCOVA model.

[54] In Delhi and U.P., since Education Leaders (EL) are unique at the cluster and strata level respectively, by including cluster (Delhi) and strata (U.P.) level fixed-effects we de-facto control for differences in EL.

[55] For missing values of covariates we used mean imputation (Puma, Michael J., Robert B. Olsen, Stephen H. Bell, and Cristofer Price 2009). As a robustness check, we also fit all our regressions without baseline teacher covariates. Though standard errors are noticeably larger, these results are similar to results presented with baseline teacher controls (which include imputed values for added endline teachers).

[56] For details on the covariates used we request the reader to please refer to Appendix A20.

[57] Standard errors are clustered at the cluster level for Delhi standard and exploratory analyses.

4.6.1.1 Professional Mindsets and Behaviors family

Professional Mindsets and Behaviors index

Our main outcome is the index of Professional Mindsets and Behaviors. We also test for three other indices developed by the NYU team called 'Positive Professional Outlook' index, 'Teacher Growth Mindset' index and 'Efficacy' index. The results are presented in standard deviations.

These indices come directly from the teacher motivation tool mentioned in sections above. The total index was created by IDinsight and the three sub-indices were created by NYU. For the full index of professional mindsets and behaviors, we averaged and standardized all 46 items. To create the three sub-indices, NYU used Exploratory Factor Analysis (EFA) followed by Confirmatory Factor Analysis (CFA). The version of the questionnaire deployed at endline is in Appendix A11. The description of each factor and included items, along with the details of the NYU study are can be found in Appendix A6.

The NYU and STIR team conducted reliability and validity tests for the tool during their study in Uganda as well as through focus groups in Delhi. IDinsight used the tool as well as the three sub-indices as specified by NYU without any changes. Please refer to Appendix A6 for more details.

4.6.1.2 Time use and child friendliness families

The classroom practice observation helped capture outcomes as part of two main families of indicators — time use and the child friendliness.

Time use

For time use, we consider two main outcomes: time teaching and time off-task[58]. For both indicators, we fit regressions using the specification mentioned above. The results are in percentage-point terms. It is important to note that these indicators come from the same question and are hence not independent of each other — teachers are coded as either teaching, engaged in classroom management or off-task. Since this implies a linear relationship between the three activities, we could expect, for instance, an increase in teaching to be accompanied by a reduction in time off-task; albeit the relation may not be one-for-one due to the presence of classroom management (not used as an indicator here).

Child friendliness

In the child friendliness family, we consider seven outcomes (please refer to Table 2 for details). For this family, we do not offer a hierarchy to the indicators. Two indicators of the child friendliness family (namely *refer by name* and *student's work displayed or praised*) were not collected as part of the baseline data collection. Hence their regressions did not include baseline outcomes. The results for the seven child-friendliness outcomes are in percentage-point terms.

[58] Given the way these data were collected (as a snapshot of the classroom) the reader should keep in mind that these are not percentage of times teachers were teaching/ off task but rather percentage of times teachers were coded as teaching/ off-task out of a total of seven observations

4.6.1.3 Student learning family

To see if there is an impact of STIR's program on learning levels we look at the Hindi and math levels obtained from the ASER testing tool. We define a student's learning level as the highest math or Hindi level the student attains. We use a simple OLS regression[59] with the learning level as the outcome variable to gauge the overall effect on learning levels.[60] More details can be found in Appendix A18.[61]

4.6.2 Teacher-level estimates: IV/LATE[62]

The IV/LATE estimation exploits the fact that participating teachers must (by design) be in schools assigned to be offered STIR communities of practice, which (by design) happened in a random fashion. (In technical terms, we use the random assignment of treatment as the instrumental variable for this analysis.) We use the relationship between a school being randomly offered treatment and a teacher in that school taking up treatment (participating in programming) to focus a light on just the outcomes of those teachers who participated in at least one meeting in each of the two years — that is, those teachers that we define as active. We will estimate the following regression specification:

First Stage:

$$\text{IV(Active)}_{ij} = \sum \alpha_s + \beta_1 * T_j + \beta_2 * Y_{0ij} + \gamma * X_{ij} + \omega_j + \varepsilon_{ij}$$

Second Stage:

$$Y_{1ij} = \sum \alpha_s + \beta_1 * \text{IV(Active)}_i + \beta_2 * Y_{0ij} + \gamma * X_{ij} + \omega_j + \varepsilon_{ij}$$

Where,

- Y_{1ij} is an individual teacher's (belonging to school j) outcome at endline
- $\sum \alpha_s$ are strata fixed effects
- Y_{0ij} is an individual teacher's (belonging to school j) outcome at baseline
- T_j is a binary variable for treatment assignment of the school the teacher belongs to (which represents pooled treatment *i.e.*, standard and exploratory clubbed together, standard treatment, or exploratory treatment, depending on the regression)
- X_{ij} is a vector of covariates. For teachers, these include teacher sex, age, qualification, years of experience, baseline teacher motivation, class size, enumerator and network dummies.
- ε_{ij} is an individual level (within schools) error term
- ω_j is a school level error term (except in the case of analyses comparing just the standard or exploratory arms in which case it is a cluster-level error term for treatment schools and a school level error term for control schools)

β_1 in the second stage equation will be our estimate of interest (effect size).

[59] The OLS model is mentioned at the very beginning of Section 4.6.1.

[60] As per our analysis plan, we initially planned to use the ordered logit model to see the marginal effect of treatment on the probability of a child being at a certain learning level (in both Math and Hindi), especially as the gains required to move from one learning level to the next are not the same for all levels. However, before analysis began, we decided to use a simple OLS model based on our understanding of STIR's ToC and desire to make a cleaner learning statement.

[61] Appendix A25 contains a series of visualizations documenting the association between baseline teacher characteristics (motivation & teaching quality & teaching quantity) and student test scores.

[62] Appendix A16 provides details on the observational analysis.

5 Results: school-wide estimates

We present our school-wide results by outcome families of indicators[63]. We also summarize the main results in Table 6; we also report on subgroup results, below. Note that, in Table 6, the 'All STIR' results represent the weighted average of the standard and exploratory results. The regression results we present in this section and the *Results Appendix* include controls for relevant covariates, as per our analysis plan.

Given the sheer number of analysis and specifications we do not comment on each result individually. Full results, including approximations of teacher-level estimates and the full set of subgroup estimates, are presented in full in the *Results Appendix*.[64,65] We provide a balance table of covariates at baseline in Appendix A14. For main and sub-group analyses, we encourage the reader to refer to Tables A25 and A26 in *Report Appendix* A15 indicating the number of teachers while interpreting results.

[63] Note we do not provide a detailed interpretation on the block-level analyses here. We do not find any overarching evidence of differential impact in any one particular block. And with lack of information around specifics of each block, we are unable to provide an interpretation.

[64] In case you do not have the *Results Appendix*, please visit our website or contact Heather Lanthorn (heather.lanthorn@idinsight.org).

[65] Due to the number of regressions our tables provide coefficients for only the estimate of interest (effect size) and do not provide coefficients (and p-values) for covariates. Note also that all the results presented are from specifications with covariates.

Table 6: Summary of all school-wide results

	Delhi private schools			U.P. government schools		
	All-STIR	**Standard**	**Exploratory**	**All-STIR**	**Standard**	**Exploratory**
Teacher professional mindsets and behaviors						
Observations	*1072*	*758*	*664*	*1133*	*750*	*749*
Overall index (*sd*)	0.086 *[-0.04:0.21]*	0.129* *[-0.01:0.27]*	0.010 *[-0.15:0.17]*	-0.023 *[-0.15:0.10]*	-0.054 *[-0.25:0.14]*	0.019 *[-0.16:0.20]*
Growth mindset sub-index (*sd*)	0.149** *[0.01:0.29]*	0.181** *[0.03:0.33]*	0.125 *[-0.53:0.30]*	-0.006 *[-0.14:0.13]*	-0.053 *[-0.39:0.28]*	0.060 *[-0.14:0.26]*
Positive professional outlook sub-index (*sd*)	0.089 *[-0.05:0.22]*	0.098 *[-0.05:0.25]*	0.045 *[-0.11:0.20]*	0.000 *[-0.12:0.12]*	-0.066 *[-0.27:0.14]*	0.070 *[-0.13:0.27]*
Teacher efficacy sub-index (*sd*)	-0.002 *[-0.11:0.11]*	0.041 *[-0.09:0.17]*	-0.076 *[-0.21:0.05]*	0.000 *[-0.13:0.13]*	-0.027 *[-0.15:0.09]*	0.031 *[-0.11:0.17]*
Classroom practice: quantity						
Observations	*462*	*321*	*285*	*644*	*432*	*425*
Observed time spent teaching (*pp*)	0.039 *[-0.05:0.13]*	0.012 *[-0.08:0.11]*	0.057 *[-0.07:0.18]*	0.039* *[-0.01:0.08]*	0.084* *[-0.01:0.18]*	-0.017 *[-0.09:0.06]*
Observed time spent off-task (*pp*)	-0.008 *[-0.02-0.01]*	-0.008 *[-0.03:0.01]*	-0.004 *[-0.02:0.14]*	-0.010 *[-0.03:0.01]*	-0.026 *[-0.06:0.01]*	0.018 *[-0.02:0.06]*
Classroom practice: quality ‡						
Observations	*462*	*321*	*285*	*644*	*432*	*425*
Teacher smiled, joked, or laughed (*pp*)	0.054 *[0.01:0.10]*	0.009 *[-0.03:0.05]*	0.136*** *[0.07:0.20]*	0.020 *[-0.02:0.06]*	0.023 *[-0.04:0.09]*	0.048 *[0.00:0.09]*
Students asked at least one question (*pp*)	0.042 *[-0.02:0.10]*	0.036 *[-0.03:0.09]*	0.046 *[-0.03:0.12]*	0.038 *[-0.02:0.09]*	0.041 *[-0.02:0.10]*	0.059 *[-0.00:0.12]*
Teacher used local materials (*pp*)	0.004 *[-0.05:0.06]*	-0.002 *[-0.07:0.07]*	0.008 *[-0.05:0.07]*	-0.023 *[-0.08:0.03]*	0.006 *[-0.06:0.08]*	-0.024 *[-0.15:0.10]*
Teacher used learning aide (*pp*)	0.043 *[-0.04:0.12]*	0.050 *[-0.04:0.14]*	0.037 *[-0.05:0.13]*	0.037 *[-0.03:0.10]*	0.056 *[-0.01:0.13]*	0.028 *[-0.04:0.10]*
Teacher grouped students (*pp*)	-0.003 *[-0.02:0.01]*	-0.004 *[-0.02:0.01]*	-0.001 *[-0.02:0.02]*	-0.005 *[-0.02:0.01]*	-0.010 *[-0.23:0.00]*	-0.007 *[-0.03:0.02]*
Teacher praised or showed off student work (*pp*)	0.029 *[-0.02:0.08]*	-0.010 *[-0.06:0.04]*	0.063 *[-0.01:0.14]*	0.019 *[-0.01:0.05]*	0.010 *[-0.06:0.08]*	0.048 *[-0.01:0.11]*
Teacher used student names (*pp*)	-0.006 *[-0.07:0.59]*	-0.021 *[-0.12:0.07]*	0.017 *[-0.06:0.10]*	0.070 *[0.012:0.13]*	0.035 *[-0.10:0.17]*	0.101 *[-0.01:0.21]*
Student learning levels						
Observations	*1844*	*1262*	*1165*	*3152*	*2047*	*2146*
Math learning levels (*sd*)	0.102** *[0.01:0.19]*	0.147*** *[0.07:0.23]*	0.053 *[-0.06:0.17]*	-0.011 *[-0.10:0.08]*	0.004 *[-0.08:0.09]*	-0.029 *[-0.09:0.03]*
Hindi learning levels (*sd*)	0.011 *[-0.07:0.09]*	-0.019 *[-0.09:0.05]*	0.048 *[-0.08:0.17]*	-0.046 *[-0.12:0.03]*	-0.060 *[-0.17:0.05]*	-0.015 *[-0.10:0.07]*

Results are presented in either standard deviations (*sd*) or percentage-points (*pp*).

95% CIs are reported under each estimate. Unlike some of the p-values, the CIs have not been corrected.

We denote estimates that achieve conventional levels of statistical significance:
*** denotes estimates significant at the 1% level
** denotes estimates significant at the 5% level
* denotes estimates significant at the 10% level

‡ denotes an estimate that is corrected for within-family testing of multiple hypotheses.

Given the large number of outcomes analyzed in this study, in Table 7 we show the total number of hypothesis tests we performed and the number of statistically significant results we find by estimate type.

Table 7: Summary of number of estimates and significant results for each analytic approach

Specification	School-wide		Teacher-level: IV/LATE		Teacher-level: Observational analysis	
	Total number of impact effects estimated	Subset of impact estimates that are significant	Total number of impact effects estimated	Subset of impact estimates that are significant	Total number of impact effects estimated	Subset of impact estimates that are significant
Main	90	8	78	8	234	38
Subgroup	209	4	209	8	414	40

Notes: The observational analysis has not been corrected for multiple hypotheses; the ITT and IV/LATE results are corrected for multiple hypotheses within family when the family contains > 4 hypotheses. Results have not been corrected across outcome families. For more information, please refer to the appendix. Estimates reported as significant include those with p-values ≤ 0.1. Significant effects mentioned here include both positive (in the expected direction) and negative (in the direction opposite to expected) estimates.

5.1 Teacher participation

To understand what the school-wide results reflect, it is helpful to understand the percent of teachers participating in STIR communities of practice in treatment schools. Recall from Section 4 that we examine two different samples of teachers for different measurement activities. For professional mindsets and behaviors (PMB), we survey all available and willing teachers; this sample approximates the total teachers in treatment schools. For classroom observations (CO), we focus on a sub-set of teachers — on average, three per school. Recall also that for our approximation of IV/LATE, we rely on a definition of a teacher being 'active' participants if they have attended one or more STIR meetings in both Year 1 and Year 2, not including the introductory taster session. In both Delhi and U.P., about 40% of teachers in our classroom observation sample meet the definition, as shown in Table 8. In both Delhi and U.P., this number is closer to 20% in our sample for professional mindsets and behaviors, which we take to roughly reflect the participation rates in the school.[66]

Table 8: Participation of teachers in STIR communities of practice in treatment schools

State	Total meetings held (Year 1 & 2)	Sample	Sample size (All STIR treatment)	Attended at least one training in Year 1 & 2 (%)	Attended half or more trainings in Year 1 & 2 (%)	Attended 75% or more trainings in Year 1 & 2 (%)
U.P.	18	CO	431	41.3	38.5	22.3
		PMB	767	23.1	21.6	13.0
Delhi	16	CO	318	39.3	31.1	18.9
		PMB	722	20.0	16.3	9.7

[66] In U.P., amongst active teachers (attended at least one STIR meeting in both Year 1 and Year 2), the average attendance rate is 13.5 meetings across both years. In Delhi, the same figure is 10.7 meetings.

5.2 Teacher professional mindsets and behaviors

5.2.1 PMB evidence and expectations

To our knowledge, no evidence exists that helps set expectations of whether and by how much teacher professional mindsets and behaviors (including motivation and teacher self-efficacy) may improve over the course of one academic year. The few studies that examine satisfaction or self-efficacy to teach use these as inputs toward student-learning outcomes, rather than as an outcome.

5.2.2 PMB results

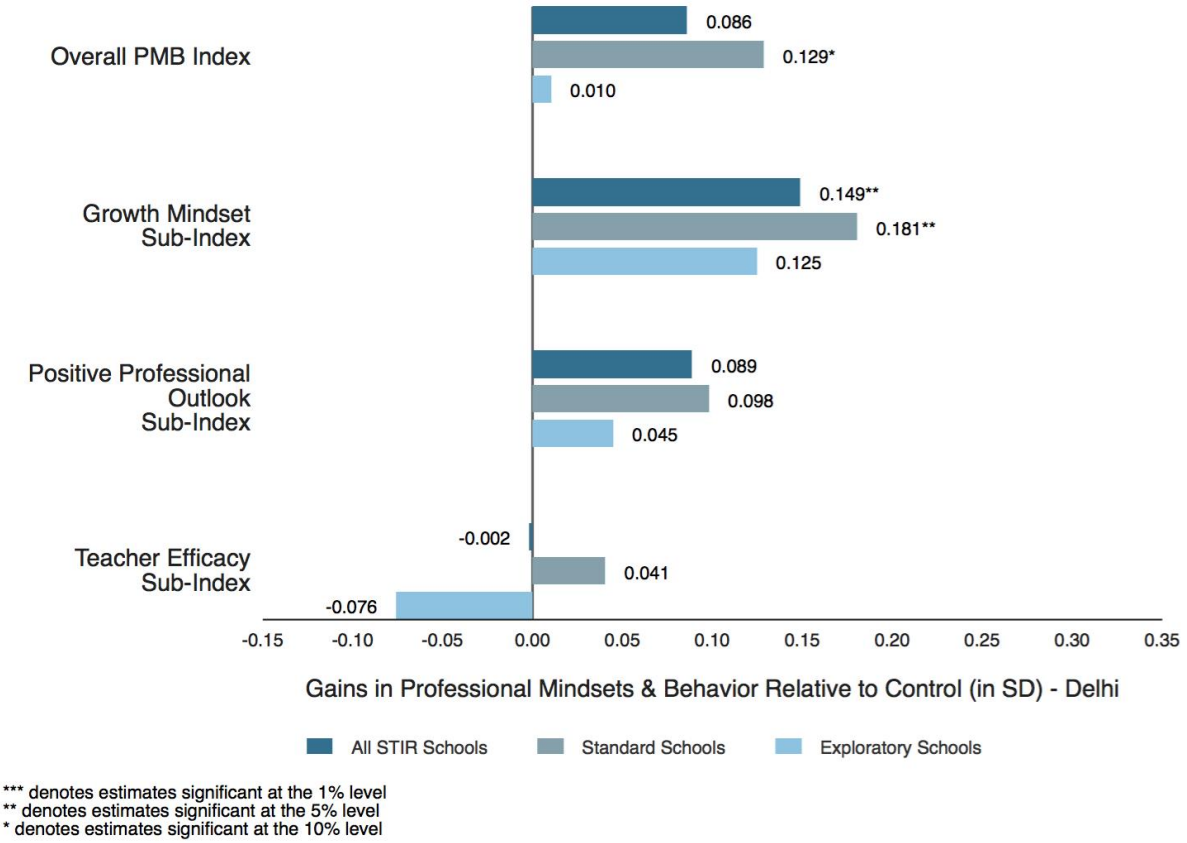
Main results

We present our motivation results by study site and with the treatment first grouped by All-STIR (meaning standard and exploratory combined), then standard and exploratory separately. We present all estimates in standard deviation (and not index value) terms.

In our school-wide estimate for Delhi private schools, we find weak evidence that the offer of STIR to schools increased teacher professional mindsets and behaviors, with particular gains in growth mindset.[67] On average, the total PMB index score for teachers in the standard STIR schools is 0.13 *sd* (p-value: 0.07) higher than teachers in comparison schools, as shown in Figure 4. On the growth mindset sub-index, teachers' scores in all STIR schools is on 0.15 *sd* (p-value: 0.0388) higher than teachers in comparison schools. Teachers in standard STIR schools had growth mindset scores 0.18 *sd* (p-value: 0.0209) higher than control teachers. Overall, results (including non-significant results) for our four measures show gains in PMB. *The IV/LATE findings are significant for the same estimates as the school-wide findings.*

[67] Note that during our midline evaluation, we found a 0.25 *sd* increase in our midline motivation index for the exploratory school teachers that received the local recognition flavor, as compared to the control school teachers in Delhi private schools.

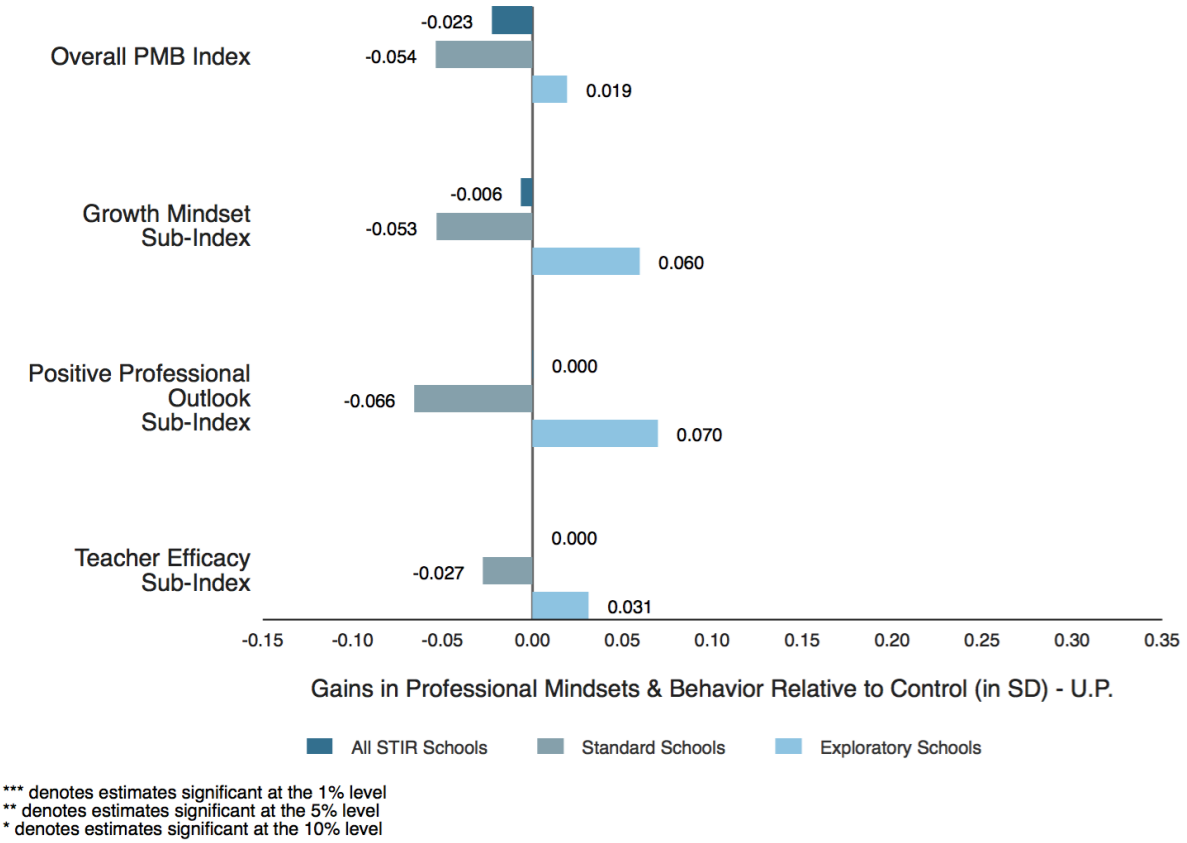
Figure 4: School-wide results for Professional Mindsets and Behaviors (Delhi private schools)



*** denotes estimates significant at the 1% level
** denotes estimates significant at the 5% level
* denotes estimates significant at the 10% level

In our school-wide estimate for government schools in U.P., we find no effect of the offer of STIR to schools on our four measures of professional mindsets and behaviors. These results are illustrated in Figure 5. *This pattern of no effect is repeated in the IV/LATE findings.*

Figure 5: School-wide effects for Professional Mindsets and Behaviors (U.P. government)



*** denotes estimates significant at the 1% level
** denotes estimates significant at the 5% level
* denotes estimates significant at the 10% level

Sub-group results[68]

We find no differential impacts on professional mindsets and behaviors among relevant subgroups in either Delhi or U.P. See Table R4 in the *Results Appendix* for results from the sub-group analysis.

5.3 Classroom practice

5.3.1 CP evidence and expectations

Time use

Instructional time is a key input into learning outcomes; however, scant literature looks at classroom time-use as an intermediate outcome toward changes in student learning (Glewwe and Kremer 2006; McEwan 2015). To our knowledge, only one randomized evaluation (in northern Brazil) looks at the use of classroom time as an outcome (from a scorecard and coaching intervention for teachers); the researchers find that a program focused on changing instructional practices leads to a 6%-point gain in teaching time over one

[68] We would be able to make a clear learning statement for STIR if all indicators (within a family) show a clear trend (both in terms of direction and significance) for a particular category of a subgroup. If for *e.g.*,: Block 'a' displays a positive (or in fact negative) significant result across all three indicators in U.P., STIR could use the evidence to think through potential reasons for the heterogeneity a bit more. Due to the lack of any particular trend in results (across any category), we are unable to offer any conclusive statements on differential impact.

academic year, accompanied by reductions in both classroom management and off-task time (Bruns, Costa, and Cunha 2017). Recall from Stallings that we have a rough benchmark that a 'good' teacher devotes about 85% of classroom time to instruction and 15% to management (World Bank 2015)[69].

Child friendliness

We face a similarly thin evidence base using child-friendly practices as an intermediate outcome (or, indeed, examining child-friendly classrooms and teacher soft skills as an outcome in a quasi/experimental set-up at all) (Suman Bhattacharjea 2017). Recall also that we don't have a clear sense of the desired level for each of the child-friendliness indicators (for example, we know that using pair- and group-work is good — but is likely not desirable to use grouping 100% of the time).

5.3.2 CP results

Time use

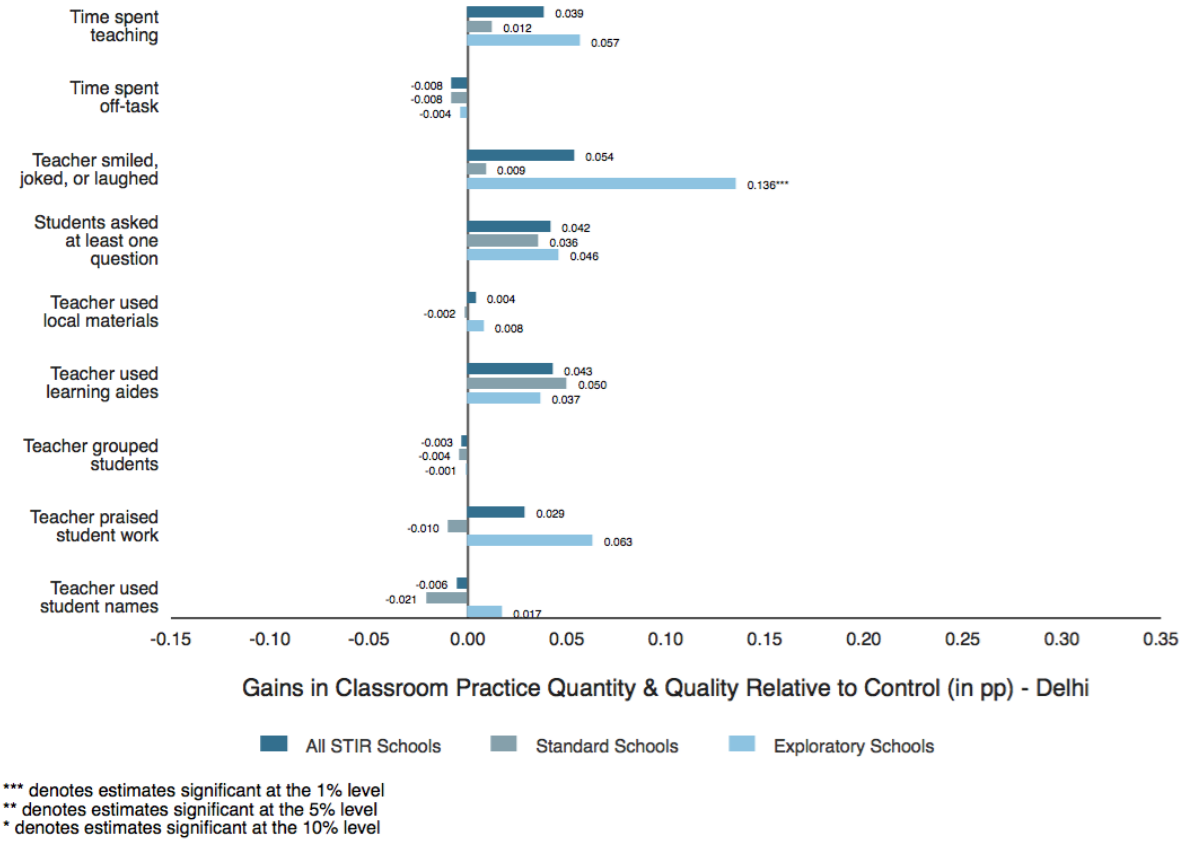
Main results

There are three, mutually exclusive ways in which teachers are recorded as spending their time: teaching, off-task and management. We report changes in teaching and off task time allotments across our three comparisons: STIR *v.* Control (C), standard *v.* C, and exploratory *v.* C.

In Delhi private schools, we find no evidence that the offer of joining STIR communities of practice led to a change in teacher time use. These results are visualized in Figure 6**.** *Similar to the school-wide results, the IV/LATE results are insignificant for teacher time use.*

[69] Note these benchmarks were based on evidence from classrooms in the United States. We do not feel these would map perfectly for the Indian classroom context.

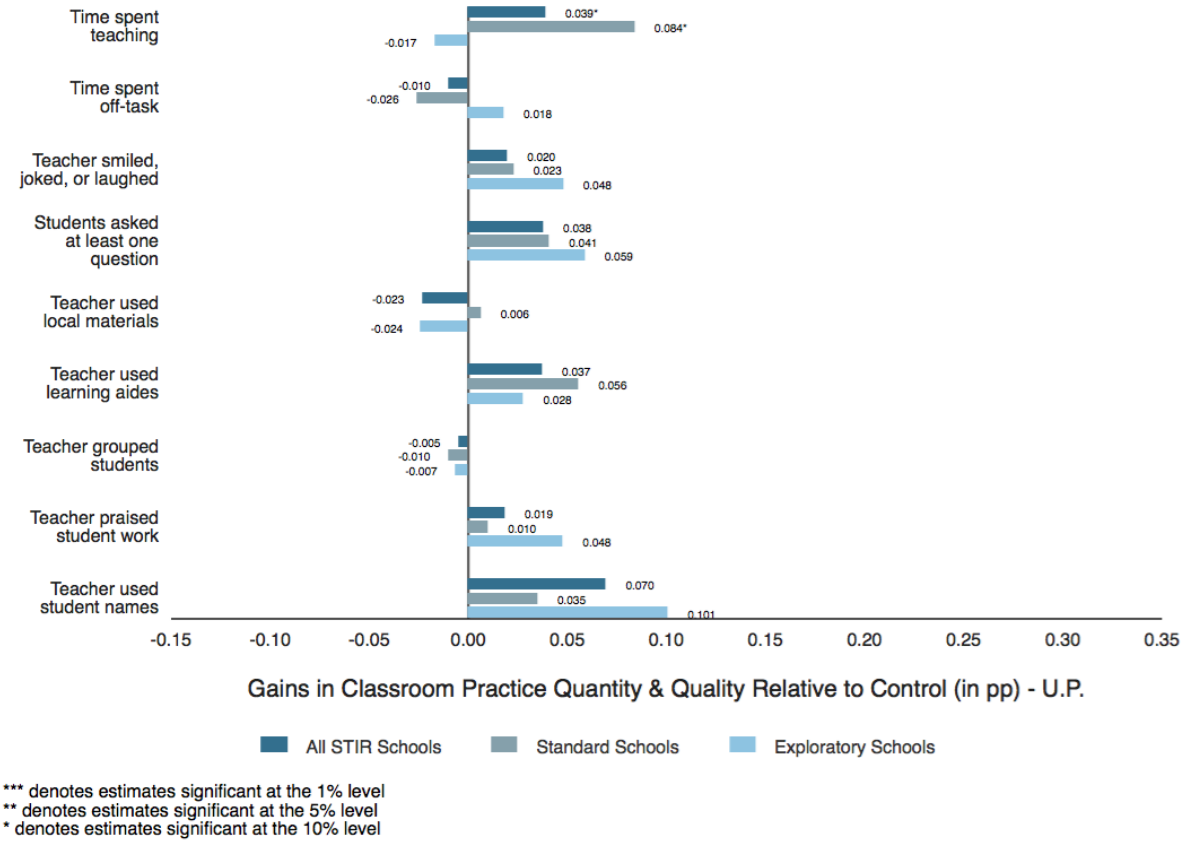
Figure 6: School-wide effects for classroom practice (Delhi private schools)



In U.P. government schools, we find weak evidence that the offer to join STIR communities of practice leads to increases in the amount of time teachers spend teaching[70]. We find a 4 percentage-point increase (p-value: 0.08) in time observed teaching (on average) for teachers in all STIR schools (standard and exploratory taken together) compared to control schools, as shown in Figure 7. In addition, teachers in standard STIR schools are observed teaching 8 percentage-points more than their control counterparts (p-value: 0.09). *The IV/LATE results are of similar magnitude and significance.*

[70] Note that during our midline evaluation, we found a 5 percentage points increase in teaching time and a corresponding 4 percentage point decrease in time off take for teachers in all STIR schools in U.P. public schools as compared to the control school teachers.

Figure 7: School-wide effects for classroom practice (U.P. government schools)



*** denotes estimates significant at the 1% level
** denotes estimates significant at the 5% level
* denotes estimates significant at the 10% level

Sub-group results

In Delhi private schools, (Table R7 in *Results Appendix*) we find no significant differences across sub-groups.

In U.P., (Table R8 in the *Results Appendix*) we find two notable sub-group effects:

- For more experienced teachers, the impact of the program on the time a teacher in U.P. spends off-task is higher by 0.3 percentage points (p-value: 0.05) as compared to less experienced teachers.
- The impact of the program on time spent off-task for female teachers in STIR schools in U.P. are 4.4 percentage-points higher (p-value: 0.06) than for male teachers.

Child friendliness

Main results

In Delhi private schools, we find limited evidence of changes in child-friendliness (Table R9). As shown in Figure 6, teachers in the exploratory STIR schools in Delhi, were observed on average smiling or laughing 13.6 percentage points (p-value: 0.000; corrected for multiple inference, as per Section 4.5.3). *The IV/LATE results are also only significant for smiling and laughing; this estimate is of similar magnitude.*

We find no gains in child friendliness attributable to the offer of joining STIR communities of practice in U.P. (Figure 7; Table R10)**.**

Sub-group results

We find no significant results differential effects on child-friendliness estimates in Delhi private or U.P. public schools by sub-group.

5.4 Student learning

5.4.1 SL evidence and expectations

The literature on improving student learning outcomes can help us set expectations for the effect sizes STIR might produce after one year of programming, calibrating for the length of these interventions and their explicit focus (in some cases) on improving certain types of learning.[71] We focus on four types of findings below: (1) those focused on teacher training and coaching, (2) those focused on teacher incentives, (3) those focused on instructional time and teacher time use, and (4) those focused on teacher behavior and classroom practice. The overall picture from these disparate studies is that these types of changes can lead to 0.1 *sd* gains in student learning.

On teacher incentives, the global impact evidence base from low- and middle-income countries (18 studies assessed through meta-analysis) shows that from teacher incentive programs (of all types), the average effect size is 0.08 standard deviations for gains in math (from 11 studies, significant at $\alpha = 0.10$) and 0.10 *sd*s for gains in language arts (from 7 studies, with an insignificant result at $\alpha=0.05$) (Snilstveit et al. 2015). In another meta-analysis, McEwan estimates that teacher incentives (from 8 studies) lead to 0.09 standard deviation gains in learning (significant at $\alpha = 0.05$) (McEwan 2015).[72]

On teacher training, McEwan estimates that teacher training (from 17 studies) leads to 0.12 standard deviation gains in learning outcomes (significant at $\alpha = 0.001$) (McEwan 2015). 'Teacher training' covers a wide range of intervention types. Popova *et al.* find suggestive but inconclusive evidence that training programs not focused on a specific subject are associated with lower student learning outcomes than trainings focused on specific subjects (Evans, Popova, and Arancibia 2016).[73] The growing literature on teaching coaching in low- and middle-income countries suggests a potential large gains with more (though often cost-effective) investment, with promising recent results in Brazil, Kenya, and South Africa, with gains in student learning ranging from 0.05 in Brazil (though up to 0.23 *sd* with strong implementation) to

[71] After one year, a remedial education program in Bombay targeting under-performing students led to 0.15 (language) and 0.16 (math) *sd* gains in learning outcomes (Banerjee et al. 2007). Another remedial education program, in Andhra Pradesh, led to a 0.74 *sd* increase in (composite language and math) learning levels after two years (Lakshminarayana et al. 2012). A performance-pay program for teachers, also in Andhra Pradesh, led to 0.35 (language) and 0.54 (math) *sd* gains in learning levels for students who experienced all five years of primary school under incentivized teachers (Muralidharan 2012). Despite the large gains in Muralidharan's study, overall, remedial education and structured pedagogy studies in low- and middle-income countries have achieved greater gains in learning outcomes as compared to teacher incentive programs (Snilstveit et al. 2015).

[72] Turning to the literature from high-income countries, a meta-analysis including seven studies of general teacher professional development suggests 0.019 *sd* gains to learning outcomes; more managed, prescriptive professional development leads to gains of 0.052 *sd* in student learning outcomes (Fryer 2016).

[73] They point to one exceptional study, in which an in-service training program focused solely on classroom management resulted in 0.47 standard deviations gains in learning outcomes (Nitsaisook and Anderson 1989).

0.25 *sd* in South Africa (up to 0.75 *sd* in urban areas) (Bruns, Costa, and Cunha 2017; Cilliers and Taylor 2017; Piper and Zuilkowski 2015).[74,75]

On quantity of classroom practice (specifically, instructional time), a few studies examine the causal relationship between time teaching and student learning outcomes; these mostly find positive or null effects and — importantly — suggest a strong role for student ability and school characteristics in moderating the relationship between additional instruction time and test score gains (Cattaneo, Oggenfuss, and Wolter 2016). Using cross-country PISA data, Lavy estimates that instructional time has a positive and significant effect on test scores — but that effect is much lower in low- and middle-income countries; in LMICs, Lavy estimates that an additional hour of teaching time per week raises test score by 0.025 *sd* for 15-year old students (compared to 0.15 *sd* in high-income countries) (Lavy 2015).[76,77] Researchers studying this question in Switzerland draw on student-reported and administrative data on time per subject and find that an additional hour per week of instructional time increases the PISA score by about 0.05 *sd* for 9th graders ($p = 0.005$). The effect is higher for more advanced (higher track) students, suggesting differences in who can make effective use of additional instructional time (Cattaneo, Oggenfuss, and Wolter 2016).

Quality of classroom practice can be measured different ways. In Ecuador, Caridad Araujo *et al.*, find that more 'responsive teaching' (as captured by the Classroom Assessment Scoring System (CLASS)) leads to improved student learning, with a 1 *sd* increase in teaching quality (responsive teaching) linked to 0.13 *sd* and 0.11 *sd* gains in language and math scores (with much larger effects after adjusting for measurement error) (Araujo et al. 2016). On specific child-friendly practices of the type we measure, the ASER Centre finds a "clear correlation between 'child friendly' classrooms and students' learning (S. Bhattacharjea, Wadhwa, and Banerji 2011). However, one (to our knowledge, the only) attempt to make a causal link finds no significant effect of adopting these practices on student learning in rural India (Das 2014).

5.4.2 SL results

We present here the OLS estimates for math and Hindi levels in Delhi and U.P.

[74] Recent work by Cilliers and Taylor allow for direct comparison between one-off teacher training and coaching (Cilliers and Taylor 2017). They find, with an intervention focused on early-grade reading, teacher training alone had insignificant (0.11 *sd*) effects on student learning proficiency, in line with the results of the meta-analyses on teacher training. Adding in monthly coaching visits, in which coaches monitor teachers and provide feedback, leads to 0.25 *sd* gains in reading scores, with even higher gains found in urban areas.

[75] Turning to high-income country evidence, a recent meta-analysis reveals 0.15 *sd* gains in student learning from teacher coaching (Kraft, Blazar, and Hogan 2017).

[76] Further, the effect is higher in schools with accountability mechanisms and the autonomy to hire and fire teachers (Lavy 2015).

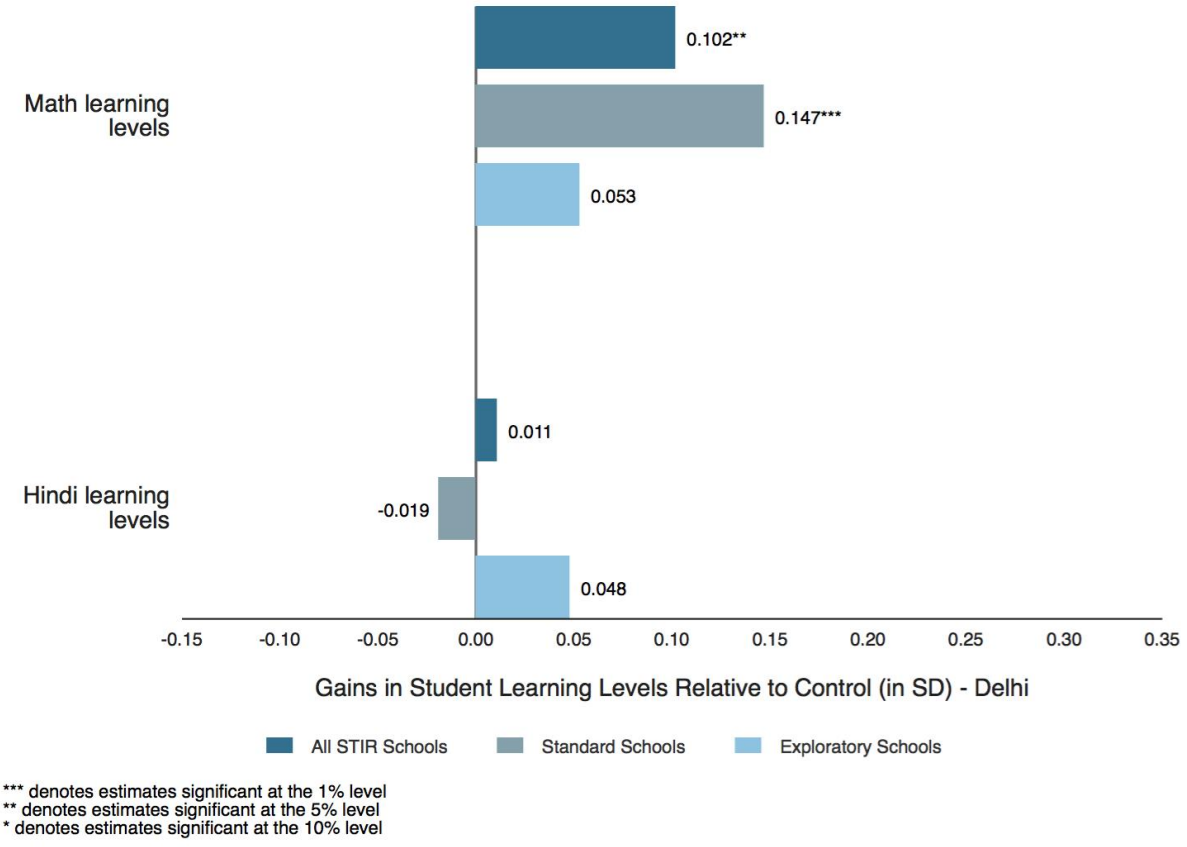
[77] In a non-causal relationship from northern Brazil, Bruns *et al.* estimate the impact of a program that delivers information to teachers about their performance and follows this up with over-Skype© coaching. The researchers estimate a 6%-point gain in teaching time as a result of the intervention. They also estimate the effects on students test scores, with gains ranging from 0.04 to 0.00 *sd*. In the sub-group of schools with the highest implementation fidelity, test score gains range from 0.13 to 0.23 *sd*.

Math levels

Main results

In Delhi, we find significant evidence of student gains in math learning levels, which appear to be driven by gains in the lowest learning levels[78]. *(See Table R15.)* On average, math levels for students in STIR schools were 0.10 *sd* higher (p-value: 0.02) than average levels of students in control schools. There is also an effect in standard model schools, where math levels for students on average is 0.15 *sd* higher (p-value: 0.00) than their control school counterparts, as shown in Figure 8. To better understand if there are certain levels at which these gains are taking place, we also run linear probability models where the outcome variables are binary indicators for whether a student achieved each math learning level (Table R15(a) in the *Results Appendix*). Results from these analyses show that the impact on math learning levels is driven by gains in more basic math concepts, namely single digits and double digits. *We do not calculate IV/LATE estimates for student learning.*

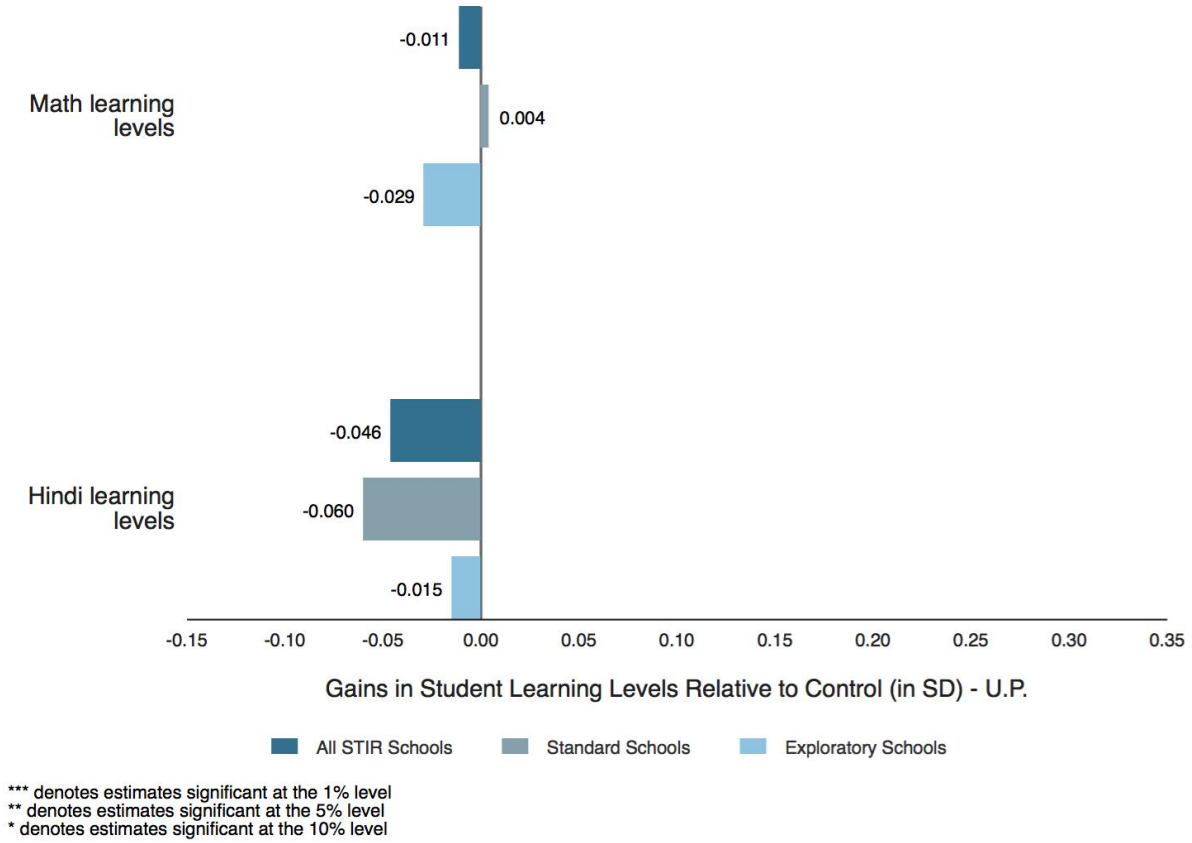
Figure 8: School-wide effects for student learning (Delhi private schools)



In U.P., we find no significant gains in student math learning levels (Figure 9; Table R16)**.**

[78] Note that during our midline evaluation, we found a 0.11 *sd* effect on the average math level of students in standard STIR schools as compared to students in control schools in Delhi.

Figure 9: School-wide effects on student learning (U.P. government schools)



Hindi levels

Main results

In both Delhi and U.P., we fail to find notable impact in standardized aggregate Hindi learning levels and did not see an effect of the STIR program on any of the seven reading proficiency levels.[79] (See Tables R16 and R17 in *Results Appendix*.)

Sub-group results

We did not conduct any sub-group analysis for student learning outcomes, for two reasons. First, we did not collect many covariates at the student level. Second, there is no theoretical reason to expect differential impact for students of different gender or age.

[79] Letter recognition, word recognition, paragraph, story 1, story 2, story 3 and story 4.

6 Discussion and conclusions

6.1 Summary

We report on two-year, endline results from two randomized evaluations of STIR's two-year Teacher Changemaker Journey in Delhi private schools (with monthly fees of US$ 17 or less) and Uttar Pradesh (U.P.) government schools. We focus on the school-wide results. Given the school-level randomization, this study was designed to test school-wide estimates of the effect of offering to a school that teachers can join STIR communities of practice; IV/LATE estimates represent an upper-bound on the likely effect of the program on actively participating teachers.

STIR seeks to improve teachers' professional mindsets and behaviors and, in turn, their classroom culture and practices, with the aim of improving student learning. In the evaluation in Delhi private schools, we see some evidence of gains in school-wide professional mindsets and behaviors, both in an overall index (0.13 *sd*, in the standard model) and in a sub-index on growth mindset (0.15 *sd*, for all STIR schools). While we see very limited effects on school-wide teacher classroom practice in Delhi, we see a significant effect on student math learning levels in STIR schools. A gain of 0.1 *sd* is in line with impacts in the literature on teacher training programs and incentives in low- and middle-income countries (McEwan 2015; Snilstveit et al. 2015). We do not find effects on Hindi learning levels.

In U.P. government schools, we see small, insignificant school-wide changes in teacher professional mindsets in behaviors and in student learning. We find weak evidence of a 10 percentage-point gain in the time a teacher spends teaching relative to control schools — which STIR equates with a gained 40-minute lesson per school day — but consider this result fragile given the overall number of hypotheses tested in U.P. and the yield of only one significant result, with a relatively large p-value. This raises concerns about finding false positives, giving us less confidence in this result than our results for Delhi private schools.

Overall, we are encouraged by the 0.1 *sd* school-level gains in math learning levels in Delhi private schools. We also find weak evidence that the standard model is more effective than the exploratory model in Delhi private schools, suggesting the extra effort and costs associated with the exploratory model may not pay off. Process work at the end of Year 1 and reports from STIR field teams suggest that elements of the exploratory model in both Year 1 and Year 2 are generally hard to implement in both contexts. This further builds the case for using the standard model going forward.

STIR plans to continue only with their government partnerships and a cascade delivery model, seeing this as their path to sustainability and scale. As we find weak evidence of impacts in only one of the outcomes of interest in U.P., STIR should investigate how to strengthen and adapt the standard model in the public education system. To do this, we suggest that STIR continue to carefully examine, collect information on, and test their theory of change in smaller pieces, partnering with external parties to do this. This work can draw on a combination of tools, including monitoring, process evaluation, and smaller-scale experimentation (or 'structured experiential learning'). Using these approaches, STIR can also clarify the indicators that are of most interest, relevance, and that can inform clear action, along with definitions of success for each. Only with these in place and a more stabilized program design does it make sense for STIR to pursue an additional large-scale impact evaluation of their programming on student learning.

6.2 Limitations and reflections for future research

There are three main technical limitations to our evaluations, which should make the reader conservative in interpreting our findings. In addition, there are a set of conceptual and construct limitations that curtail the claims we can make from our randomized evaluations, suggesting considerations for future research.

6.2.1 Technical challenges

First, we experience high levels of teacher and student attrition over the course of the two-year study. While we find no evidence of differential attrition across the treatment and control groups on our baseline covariates, reassuring us that our causal estimates are robust, it is possible that attrition is imbalanced on some unobserved characteristics of teachers or schools. It is further possible that attrition dampens our power to pick up small but practically significant effects, particularly in professional mindsets and behaviors and in classroom practices, where the literature offers limited guidance on what effect sizes we might reasonably expect.

Second, we examine many hypotheses, including across multiple outcomes, treatment comparisons (all-STIR *v.* control, standard *v.* control, and exploratory *v.* control), and sub-groups. While we correct for this multiple hypothesis testing within outcome families with more than four outcomes, we do not correct across families. Thus, we may still be prone to false positives.

Third, our observations of classroom practice may be subject to differential observer effects. Observer effects occur when study participants change their behavior in response to being observed; this is concerning for causal inference when we expect participants in the treatment and control groups to respond differentially. We expect this to be the case for child-friendliness but not for time use. Overall, the presence of our enumerators may induce teachers to, for example, spend more time teaching when we are observing their classrooms; therefore, the absolute levels of teaching time observed in both treatment and control schools are likely to be an above-average case. We do not, however, expect this to be different between treatment and control groups, as STIR does not explicitly tell teachers about how to spend their classroom time. In contrast, many of the practices that form our child-friendliness indicators are explicitly encouraged by STIR during Year 2 programming (focused on classroom culture and practice), such as having students work in groups, calling students by their names, and smiling when addressing the students. Thus, teachers in STIR schools might 'teach to the test' with respect to child-friendly practices in the presence of our enumerators since they are aware of what STIR considers good classroom practices.

6.2.2 Conceptual challenges

We face challenges related to the claims we can make from these randomized evaluations (REs), given their design and power as well as the constructs measured. While most of these challenges were known up-front and deemed acceptable, at the end of the study, we face frustration that we cannot make stronger claims that might inform action.

Study design

We face three main challenges related to the design of this set of evaluations and the claims we can make. Future studies may want to invest in setting up studies that can explicitly explore these dimensions.

1. Causal links between our outcomes: in this study we focus on three main sets of outcomes that have an implied narrative and chorological connection in the theory of change: improved

professional mindsets and behaviors will lead to improved classroom practices in both quantity and quality and that these changes will lead to gains in Hindi and math learning. Our evaluation is designed to test the impact of the (offer of the) participating in STIR communities of practice on each of these outcome sets separately; it is not designed to investigate the causal connections between these outcomes.[80] This is a clear area of interest for future investigation, to understand how changes in professional mindsets and behaviors cause teacher behavior change in classrooms and, in turn, how the types of changes teachers make in their classrooms *cause* changes in student learning levels.

2. Comparisons across contexts: we designed this study to be a set of parallel, independent randomized evaluations examining STIR communities of practice in two different contexts. However, now that we have found results in one setting but not the other, it is naturally tempting to want to draw lessons across these contexts. These contexts differ in multiple ways, including urbanness, baseline learning levels, school type, the delivery model, and engagement with the school's Head Teacher; this makes drawing any conclusions fraught. Future research might more explicitly take test aspects of delivery and context, such that the researchers can make causal claims about their influence.
3. Claims about individual teachers: we designed this study to optimally capture the effect of a school being offered STIR communities of practice, which we take to be the most important question for scale and policy-making. That said, it may be useful to STIR to continue to explore effects among their actively participating teachers, while being careful about how findings are extrapolated. This measurement effort will be complicated by the voluntary nature of participation in STIR and that STIR-active teachers are encouraged to share what they learn from STIR with peers in their school.

Measurement

We face three measurement challenges that may be relevant for future work on STIR and beyond. First, the program itself has undergone small tweaks and one substantive mid-program update over the course of this evaluation. From the outset, STIR planned for programming in Year 1 and Year 2 to be different, as teachers gained the agency, confidence, and skill to progress through the Teacher Changemaker journey (as described in Section 3). We designed this evaluation to estimate the cumulative two-year effects of this journey as a package.

Another change in the conceptualization of the program — and a more general challenge to measurement — related to the idea of teacher motivation. Teacher motivation is an important concept in education research but multi-faceted and difficult to measure. Psychometric measures can potentially be gamed and present comprehension challenges for teachers; proxies such as teacher attendance and presence in the classroom do not capture the whole motivation story. We hope researchers continue to explore how to measure this construct in ways that are compelling to the wider education community and that can inform clear action as outcomes in an impact evaluation. In addition to these general challenges, as STIR progressed through the Year 1 of the evaluation, they homed in on a broader set of interests to target: professional mindsets and behaviors, of which motivation was one of nine components. This is what we measure at

[80] To test whether PMB is a key driver of student learning gains, IDinsight considered conducting a mediation analysis. The intent would be to test whether the treatment improves learning outcomes even after PMB is included as a "mediating" control variable. If the coefficient on treatment remains statistically significant even after inclusion of PMB, then this would point to indicative evidence that PMB is not a mediator between the treatment and student outcomes. Ultimately, the decision was made to not conduct this analysis given possible omitted variable bias — it is highly likely that there are confounding variables correlated with teachers' mindsets and student learning.

endline. More work may be warranted in assessing whether we captured the most meaningful aspects, which are expected to link with changed classroom practice and/or with student learning gains.

Third, there are challenges in measuring the quantity and quality of classroom practice in the classrooms of low- and middle-income countries in ways that are meaningful as impact evaluation outcomes and can inform decision-making. The classroom observation tools we used are premised on a school system with clear times for different lessons and in which we could enter a classroom before a teacher. In the sites we visited in India, these premises did not always hold, presenting some challenges in generating measurements that are comparable with other studies making use of the Stallings snapshot tool. Because of these challenges, our measurement of classroom time-use likely underestimates time off-task. In addition to this, there is room to further refine how to measure quality of classroom practice as an actionable impact measures, capturing aspects most relevant to student learning gains.

6.3 Areas for future action and research

We have focused this set of parallel evaluations on the impact of STIR communities of practice three sets of outcomes: teacher professional mindsets and behaviors, teacher classroom practice, and student learning. To think about how STIR should move forward, it is helpful to zoom out from these outcomes to take in the whole theory of change around and between them (as elaborated in the *Report Appendix*), as the path toward improved student learning is longer than implied by our focus on three sets of outcomes. By laying these out in detail, we do not imply that STIR has not considered these but, rather, aim to help other readers grasp the many moving parts in the change process and how they might be systematically investigated.

Our results (but not necessarily the programming costs or logistical realities) might suggest that STIR continue their standard model of programing in Delhi private schools as a continued learning lab, to understand how to strengthen and realize their desired impacts. However, STIR is invested in their U.P. approach, embedding into government school systems and using cascade delivery models to achieve scale. To strengthen this model, STIR should engage in continued, detailed investigation of the theory of change (elaborated in the *Report Appendix*) through careful, objective measurement. Further, STIR should clarify what gains in student learning are sufficient to justify their approach of focusing on professional mindsets and behaviors as well as classroom practice and culture. Is a 0.1 *sd* gain in math on target? Would a similar gain in Hindi be equally satisfying? Clarity around their effectiveness and cost-effectiveness goals — and working backward from them toward a strengthen program at each step — will help chart the path forward.

References

"Annual Status of Education - Rural." 2005. Annual Status of Education Reports. Delhi: Pratham Resource Centre. http://img.asercentre.org/docs/Publications/ASER%20Reports/ASER_2005/aserfullreport2005.pdf.

Araujo, M. Caridad, Pedro Carneiro, Yyannu Cruz-Aguayo, and Norbert Schady. 2016. "Teacher Quality and Learning Outcomes in Kindergarten." Working Paper IDB-WP-665. IDB Working Paper. Inter-American Development Bank.

ASER Centre. 2018. "Sampling." March 16, 2018. http://www.asercentre.org/Overview/Basic/Pack/History/etc/p/56.html.

Bandura, Albert. 1969. *Principles of Behavior Modification*. Holt, Rinehart and Winston.

Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Linden Leigh. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics* 122 (3): 1235–64.

Bhattacharjea, S., W. Wadhwa, and Rukmini Banerji. 2011. "Inside Primary Schools: A Study of Teaching and Learning in Rural India." Delhi: ASER Centre. http://img.asercentre.org/docs/Publications/Inside_Primary_School/Report/tl_study_print_ready_version_oct_7_201 1.pdf.

Bhattacharjea, Suman. 2017. "Greetings; Question on Child-Friendliness Indicators," January 11, 2017.

Bruns, Barbara, Leandro Costa, and Nina Cunha. 2017. "Through the Looking Glass: Can Classroom Observation and Coaching Improve Teacher Performance in Brazil?" Working Paper. World Bank, Center for Global Development, Stanford University.

Cattaneo, Maria A., Chantal Oggenfuss, and Stefan C. Wolter. 2016. "The More, the Better? The Impact of Instructional Time on Student Performance." Working Paper IZA DP No. 9797. Discussion Paper Series. Bonn: IZA. http://ftp.iza.org/dp9797.pdf.

Cilliers, Jacobus, and Stephen Taylor. 2017. "Monitoring Teachers and Changing Teaching Practice: Evidence from a Field Experiment." Working Paper.

Das, Sushmita Nalini. 2014. "Do 'Child Friendly' Practices Affect Learning? Evidence from Rural India." Working paper. London: Instutute of Education, University of London.

Dweck, Carol S. 2010. "Even Geniuses Work Hard." *Educational Leadership* 68 (1): 16–20.

Evans, David, Anna Popova, and Violeta Arancibia. 2016. "Inside In-Service Training." In *RISE Annual Conference 2016*. Oxford. http://www.riseprogramme.org/sites/www.riseprogramme.org/files/Evans%20Inside%20In-Service%20Teacher%20Training%20-%20CLEAN%20-%20v2016-06-22.pdf.

Fryer, Roland G. 2016. "The Production of Human Capital in Developed Countries: Evidence from 196 Randomized Field Experiments." Working Paper 22130. NBER Working Paper. Cambridge: National Bureau of Economic Research.

Glewwe, Paul, and Michael Kremer. 2006. "Schools, Teachers and Education Outcomes in Developing Countries." In *Handbook of the Economics of Education*, edited by E. Hanushek and F. Welch. Vol. 2. Amsterdam: Elsevier.

Goyal, Sangeeta, and Priyanka Pandey. 2009. "How Do Government and Private Schools Differ? Findings from Two Large Indian States." Delhi: World Bank. http://2010.economicsofeducation.com/user/pdfsesiones/042.pdf.

Kraft, Matthew A., David Blazar, and Dylan Hogan. 2017. "The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence." Working paper. Providence: Brown University. https://scholar.harvard.edu/files/mkraft/files/kraft_blazar_hogan_2016_teacher_coaching_meta-analysis_wp_w_appendix.pdf.

Lakshminarayana, Rashmi, Alex Eble, Preetha Bhakta, Chris Frost, Peter Boone, Diana Elbourne, and Vera Mann. 2012. "Support to Rural India's Public Education System: The STRIPES Cluster-Randomized Trial of Supplementary Teaching, Learning Material, and Additional Material Support in Primary Schools." *PLoS One* 8 (7). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3712986/.

Lavy, Victor. 2015. "Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries." *The Economic Journal* 125 (588). https://onlinelibrary.wiley.com/doi/abs/10.1111/ecoj.12233.

Linden, Leigh. 2008. "Complement or Substitute? The Effect of Technology on Student Achievement in India." Working paper. http://www.leighlinden.com/Gyan_Shala_CAL_2008-06-03.pdf.

McEwan, Patrick J. 2015. "Improving Learning in Primary Schools of Developing Countries: A Meta-Analysis of Randomized Experiments." *Review of Educational Research* 85 (3): 353–94.

McKenzie, David. 2012. "Beyond Baseline and Follow-up: The Case for More T in Experiments." *Journal of Development Economics* 99 (2): 210–21. https://doi.org/10.1016/j.jdeveco.2012.01.002.

Muralidharan, Karthik. 2012. "Long-Term Effects of Teacher Performance Pay: Experimental Evidence from India." Working Paper. San Diego: UC San Diego. http://econweb.ucsd.edu/~kamurali/papers/Working%20Papers/Long%20Term%20Effects%20of%20Teacher%20Performance%20Pay.pdf.

Muralidharan, Karthik, Jishnu Das, Alaka Holla, and Aakash Mophal. 2016. "The Fiscal Cost of Weak Governance: Evidence from Teacher Absence in India." Working Paper. http://econweb.ucsd.edu/~kamurali/papers/Working%20Papers/Fiscal%20Cost%20of%20Weak%20Governance%20(Current%20WP).pdf.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J. Ganimian. 2017. "Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India." Working Paper. http://econweb.ucsd.edu/~kamurali/papers/Working%20Papers/Disrupting%20Education%20(Current%20WP).pdf.

Muralidharan, Karthik, and Venkatesh Sundararaman. 2006. "Teacher Incentives in Developing Countries: Experimental Evidence from India." Job market paper.

NCERT. 2005. "National Curriculum Framework." Delhi: National Council of Educational Research and Training.

Nitsaisook, M., and L.W. Anderson. 1989. "An Experimental Investigation of the Effectiveness of Inservice Teacher Education in Thailand." *Teaching & Teacher Education* 5 (4): 287–302.

Piper, Benjamin, and Stephanie Simmons Zuilkowski. 2015. "Teacher Coaching in Kenya: Examining Instructional Support in Public and Nonformal Schools." *Teaching & Teacher Education* 4 (April): 173–83.

Pritchett, Lant, Salimah Samji, and Jeffrey Hammer. 2012. "It's All About MeE." http://www.wider.unu.edu/publications/working-papers/2012/en_GB/wp2012-104/_files/88827043788095820/default/wp2012-104.pdf.

Ree, Joppe de, Karthik Muralidharan, Menno Pradhan, and Halsey Rogers. 2017. "Double for Nothing? Experimental Evidence on an Unconditional Teacher Salary Increase in Indonesia." *Quarterly Journal of Economics*, November. https://doi.org/10.1093/qje/qjx040.

Rivkin, S., E. Hanushek, and J. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica* 73 (2): 417–58.

Ryan, Richard M., and Edward L. Deci. 2000. "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions." *Contemporary Educational Psychology* 25 (1): 54–67. https://doi.org/10.1006/ceps.1999.1020.

Shah, Neil Buddy, Andrew Fraker, Paul Wang, and Daniel Gatsfriend. 2015. "Evaluations with Impact: Decision-Focused Impact Evaluation as a Practical Policymaking Tool." Working Paper 25. 3ie Working Paper. Delhi: International Initiative for Impact Evaluation (3ie).

Snilstveit, Birte, Jennifer Stevenson, Daniel Phillips, Martina Vojtkova, Emma Gallagher, Tanya Schmidt, Hannah Jobse, Maisie Geelen, Maria Grazia Pastorello, and John Eyers. 2015. "Improving Learning Outcomes and Access to Education in Low- and Middle-Income Countries: A Systematic Review." 3ie Systematic Review 24. London: International Initiative for Impact Evaluation (3ie). http://3ieimpact.org/en/evidence/systematic-reviews/details/259/.

Staiger, D., and J. Rockoff. 2010. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24 (3): 97–118.

Stallings, Jane. 1977. "Learning to Look: A Handbook on Classroom Observation and Teaching Models."

STATA (version 14.0). n.d. College Station, Texas: STATA Corporation.

STIR Education. n.d. "'Our Approach.'" Accessed May 31, 2016. http://stireducation.org/#our-approach.

World Bank. 2015. "Conducting Classroom Observations: Analyzing Classrooms Dynamics and Instructional Time." Washington, D.C.: World Bank.
———. 2018a. "Microdata." 2018. https://data.worldbank.org.
———. 2018b. "World Development Report 2018: Learning to Realize Education's Promise." Text/HTML. World Development Report. Washington, D.C.: World Bank. doi:10.1596/978-1-4648-1096-1. http://www.worldbank.org/en/publication/wdr2018.