IDinsight
DATA. DECISIONS. DEVELOPMENT.

STiR education
Empowering Teachers to be Changemakers

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Non-Financial Teacher Incentives: Impact of the STIR program after one year on motivation, classroom practice, and student learning

Midline report

August 2017[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
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
U.P. ……..Uttar Pradesh, a state in northern India
MDE ……Minimum Detectable Effect
ITT ……Intent to Treat (school-wide estimate of STIR's program)
ToT ……..Treatment-on-the-treated (teacher-level estimate of STIR's program)
IV ……Instrumental Variable
LATE ……Local Average Treatment Effect
OLS ……Ordinary Least Square

Acknowledgements

This report has been a collaborative effort. The report has been driven by the core team of the evaluation namely Heather Lanthorn, Doug Johnson, Pratima Singh, and Varun Chakravarthy. However, we have benefitted from the input and advice of many of our team members at IDinsight. We would like to thank Jeff McManus, Akshat Goel, Dan Stein, and Qayam Jetha for their technical advice and input through the course of the analyses and writing the report. We would also like to thank Torben Fischer for his invaluable contributions and leadership during his internship with IDinsight. We have benefitted from the perspective lent by Ronald Abraham, Neil Buddy Shah, and many others around the organization. Finally, we are lucky to have our amazing field managers Rajkumar Sharma and Pramod Kumar as part of the team, who have ensured high-quality data collection, time and time again. 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 have offered key technical support. STIR have been excellent learning partners, and have been patient, understanding, helpful, and supportive through this 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

We provide estimates of the causal impacts of one year of STIR's programming on key teacher and student outcomes in two locations in India. This midline report represents the mid-points of two parallel evaluations, after one academic year of programming. The endline report will cover the 2-year impacts.

STIR seeks to improve teachers' motivation, mindset, and classroom practice to improve student learning outcomes. This happens through voluntary, in-service professional development and non-financial incentives for teachers. Broadly, STIR works to inspire teachers to become agents of change in their classroom, schools, and education systems. STIR's basic theory of change (as relevant to Year 1 of this evaluation) holds that as teachers change their motivation to teach well and their mindset on their potential to become a better teacher is possible, they will change their level of effort to update their classroom practice, which may change the quantity and/or quality of their teaching. These changes may change student learning outcomes. Changes in classroom practice and student performance will also influence teachers' motivation.

STIR is both the program designer and implementer. We are conducting two parallel randomized evaluations of STIR's programming: one with Affordable Private Schools (APSs) in East Delhi and another with government schools in the state of Uttar Pradesh (U.P.). (Please refer to the baseline report for additional details.) There are differences in the program and evaluation designs across these settings. In East Delhi, STIR staff directly deliver programming to teachers. IDinsight is examining 180 APS, of which 120 were randomly selected to receive the treatment (invitation for teachers in a school to join STIR). In U.P., STIR uses a cascade or "training of trainers" approach: government staff (government school teachers) are trained by STIR staff to deliver programming to other teachers. In U.P., out of 270 government schools, we randomly selected 180 to receive the treatment. In both cases, the participation of individual teachers in treated schools is intended to be voluntary.

In each setting, we test two different variations of STIR's programming – the "core" and the "core-plus" models. Core programming focuses on enhancing intrinsic motivation and professional mindsets among teachers. The core-plus model adds various non-financial incentives to the core programming, thus aiming to increase both intrinsic and extrinsic motivation. The key research questions for both evaluations are to measure the impact of STIR's programming on the following outcomes:

- teacher motivation (using a motivation questionnaire and teacher attendance);
- quantity of teaching practice (using a modified Stallings classroom snapshot) (Stallings 1977; World Bank 2015);
- quality of teaching practice (using a classroom observation tool of child-friendly behaviors);
- student learning (using a modified version of the ASER tool to assess Hindi and math learning levels) ("Annual Status of Education - Rural" 2005).

In each setting, we estimate both school-wide ("intent-to-treat") and teacher-level ("treatment-on-the-treated") effects. School-wide effects capture the overall effect of STIR programming on aggregated school outcomes, including teachers who actively participate in STIR as well as those who don't. As STIR both encourages participating teachers to influence and inspire other teachers in their schools, as well as works in some cases directly with Head Teachers, the school-wide estimate sheds light on the full potential of STIR's programming. The school-wide results therefore capture the combined impacts of teachers who attend STIR meetings as well as those teachers who do not but may be affected through

multiple other channels of influence. We also explore teacher-level effects — *i.e.*, the impact of STIR on the 40-50% of teachers in treatment schools who actively participated in the program — three ways:

- focusing on small schools (where a higher proportion of teachers are active participants);
- using Instrumental Variable/Local Average Treatment Effect (IV/LATE) estimation;
- conducting an observational study style analysis, comparing only teachers who participated actively in STIR's programming in treatment schools with all teachers in control schools.

Out of 52 primary tests of STIR's effects on key outcomes of interest, we find statistically significant effects, in the expected directions, for 4 of the tests. While we correct for multiple hypothesis testing within outcome families, we do not correct for multiple hypothesis testing across these main tests; these results should be interpreted with care. We describe the four statistically significant results below.

- **Student learning**: In core schools in Delhi, we observe a 0.11 standard deviation (sd) gain in math learning levels (significant at α=0.05).
- **Teacher classroom practice**: In U.P, combining core and core-plus schools, we observe a 5 percentage-point increase (significant at α=0.05) in observed teaching time and a 4-percentage point decrease (significant at α=0.05) in the time teachers spend off-task.
- **Teacher motivation**: In the core-plus "local recognition" package in Delhi, we observe a 0.25 (significant at α=0.05) standard deviation (sd) increase on our motivation index.

We provide a summary of the significant impact estimates below, in Table 1.

Table 1: Summary of number of significant results for each analysis

Specification	ITT		IV/LATE		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	52	4	30	3	90	27
Subgroup	266	39	266	25	798	86

Notes: *The IV/LATE and observational analysis are not corrected for multiple hypotheses; the ITT results are corrected for multiple hypotheses within family and not the aggregate levels; 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.*

The subgroups we test include baseline motivation levels (low, medium, high); teacher experience (three and fewer, more than three years of experience); teacher sex (U.P. only); and blocks (administrative units in U.P., representing different key stakeholders and levels of support for STIR's programming). Overall, we find inconsistent results at the subgroup level and do not see a clear trend of differential impact for any particular subgroup. Results from the small-school analysis, IV/LATE analysis, and the observational style analysis similarly reveal mixed evidence across settings and treatment variations.

We emphasize that these are the first year, interim midline results of an evolving program. STIR anticipates having greater impact on classroom practice and student learning in the second year. A fuller understanding of the effect, interpretations, and mechanisms will be available at endline.

2 Introduction

This report is organized as follows. In Section 2, we introduce the evaluations and the program contexts. In Section 3, we turn to an overview of the programmatic components and underlying logic and assumptions in STIR's theory of change (ToC). This builds a solid foundation for understanding the evaluation questions, methodology, and findings. 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 so far and in Section 6, we conclude with a brief discussion of the limitations.

2.1 Background

128 million children are enrolled in primary school in India, and there are 2.4 million teachers in primary schools (India Ministry of Human Resource Development, 2009). The story of India's educational achievements is one of mixed success. While there has been encouraging progress made in raising school participation, India has 46 per cent of the world's illiterate population and is home to a high proportion of the world's out-of-school children and youth (Kingdon 2007). According to the International Association for the Evaluation of Educational Achievement (IEA), the performance of Indian children is poor relative to most low- or middle-income countries.

Teacher effort and ability can have a large effect on student learning outcomes (Rockoff, 2003) and lifetime outcomes (Chetty et al., 2011), yet in India (as in many low- and middle-income countries), teacher performance is generally poor and accountability to citizens and the state is low. Teachers are frequently absent: random audits of public school teachers in India found that teachers were present only 25% of the time and even when present, teachers spent a majority of the time not actually teaching (Kremer et. al., 2005). Less than half of teachers are both present and engaged in teaching on any given school day (Chaudhury et al. 2006). Teacher attendance and activity is similarly low in the affordable private schools (APS) proliferating throughout India (Goyal and Pandey, 2009).

There is evidence that financial incentives can be an effective accountability mechanism to produce desired behavior among teachers (Duflo et. al., 2012) and can increase teacher effort at school as well as student test scores (Muralidharan and Sundararaman, 2011). However, financial incentives have not sustained increases in teachers' effort; focusing on understanding non-financial incentives and its role in teacher motivation may be crucial. Teacher motivation is increasingly viewed as an important link to teacher effectiveness outcomes. Muralidharan and Sundararaman stress the importance of motivation and professional incentives in teachers' revision of their teaching practices towards more effective practices. Teaching effectiveness plays a major role in moving from improved enrolment and attendance to improved learning outcomes (Pritchett 2013; Burgess 2016).

2.2 STIR's program

STIR's[2] programming fits loosely under two broad categories of interventions commonly discussed with respect to improving learning outcomes: teacher training and teacher incentives. However, most evaluated

[2] About STIR (designer and implementer): STIR Education, a non-governmental organization, focuses on helping teachers become central agents of change to overcome a key crisis of learning in low- and middle-income countries: children are increasingly enrolled in school (UN Secretariat 2015) but not learning (Robinson 2011). STIR is headquartered in London and works currently in multiple states in India as well as in Uganda.

in-service teacher training programming focuses on pedagogical techniques, content/subject knowledge, or the use of technology (Evans, Popova, and Arancibia 2016). STIR, instead, focuses on teachers' motivation and sense of agency to make positive changes in their classrooms and schools, accompanied by building soft and professional skills to make — and learn from — changes in classroom management and practice. In addition, 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.[3]

STIR's unique approach is to recognize that teachers' motivation and the classroom culture they foster are central to their effectiveness. STIR focuses squarely on improving teacher motivation to bring about change and the agency and soft professional skills required to do this well (STIR Education 2016). STIR's basic theory of change (as relevant to Year 1 of programming, evaluated in this report) is that as teachers improve their motivation to teach well and their mindset that becoming a better teacher is possible, they will make the effort to change their classroom practice, which may change the quantity and/or quality of their classroom practice. In turn, these changes will facilitate improved student learning outcomes.

Through the course of the first year, 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 be feedback effects from classroom practice and student outcomes on teacher motivation. Efforts to make changes in the classroom and student outcomes could have a positive or negative impact on a teachers' motivation. We provide more details of STIR's program and their theory of change in Section 3.

2.3 Overview of the two evaluations

To assess the impact of STIR's programming, IDinsight[4] has undertaken a pair of randomized evaluations (REs) (supplemented with a set of process evaluations) in Affordable Private Schools (APS) in East Delhi and government schools in the districts of Rae Bareli and Varanasi in Uttar Pradesh (U.P.). The 'treatment', in this case the offer to schools to have their teachers to join STIR, was randomized at the school level.[5]

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 program spans three years,

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

[4] About IDinsight (learning partner): IDinsight seeks to partner with clients committed to using and generating rigorous evidence to improve their social impact. Depending on client needs, we help to: diagnose social sector challenges; design and test potential solutions; and operationalize those solutions found to be the most impactful. 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 a learning partner since 2013, when IDinsight assisted with some early process evaluations and support for the development of a theory of change. Once the program model had matured, STIR decided it was time to pursue a more rigorous evaluation of impact, leading to the current set of evaluations.

[5] 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 (related to teacher jealousy) as well as severe evaluation concerns (related to the potential for contamination between treated and comparison teachers within the same school).

this set of evaluations is planned for two academic years. This report represents the midpoint of these evaluations. For background on randomized evaluations and the formative nature of the current evaluations please refer to Appendix A1.

STIR's program in Delhi and Uttar Pradesh have distinct implementation and institutional contexts. STIR thinks of their APS model as a 'lab,' with higher control over implementation quality since the program is carried out by STIR staff directly. The Uttar Pradesh model is closer to the 'at-scale' model, since it is deeply embedded within the government structure and is led by government school teachers themselves with training and support from STIR staff and officials. We provide more details of the two approaches in Section 3.3.1.

2.4 Overview of program variants

Both evaluations experimentally investigate two models – a core and a core-plus model. The core model is 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.

The core-plus model builds upon the core model by layering on a package of extrinsic motivation techniques including structured recognition for innovation and value-added to student learning from peers and at larger annual events, certification for high-performers, and sharing of student and teacher performance data with relevant local stakeholders. In Delhi, there are four different 'flavors' of the core-plus model, which introduce different source of extrinsic recognition and motivation: namely a local recognition package, a head teacher recognition package, a teacher exposure package, and a career and personal development package. In Uttar Pradesh, there are three flavors: the local recognition package, the government and policy exposure package, and the teacher exposure package. We revisit these in Section 3.3.2 and more details of the packages can be found in Appendix A2.

3 STIR's programming and variations

In contrast to traditional approaches to system improvement which are based on top down accountability and an exclusive emphasis on individual teacher skills, STIR's approach is designed to harness and strengthen teachers' collective motivation to improve learning (STIR Education 2015).

Teachers part of the STIR program enter a three-year 'Teacher Changemaker Journey.'[6] This journey is designed to improve learning by addressing the urgent need to develop teachers' mindsets and soft skills related to classroom and school culture and practice. As they go through the journey teachers become increasingly adept at working together to tackle increasingly complex challenges and barriers to learning and get steadily better at using evidence to inform their teaching.

Teachers are organized into local networks with teachers of other schools (located nearby). These networks meet monthly in network meetings, providing the opportunity for teachers to collaborate with one another, share ideas on overcoming day to day challenges, and learn from others. Teachers take back

[6] The evaluations currently assess the impact of the first two years of this teacher changemaker journey with STIR.

to their respective classrooms the innovative techniques discussed in the meetings. Teachers are encouraged to reflect upon their teaching practices using a portfolio (workbook) and influence others around them.

Education Leaders (ELs) play a pivotal role through this entire journey. They coordinate and facilitate network meetings and encourage and support teachers to undertake other STIR activities.[7] We provide more details of the various components of and actors in the STIR program below.

3.1 Key programmatic components, activities, and actors

STIR's programming, which is designed by STIR, has many moving parts. We detail our understanding of the flow of the first year of STIR's programming in the following section on "Program description and theory of change." However, before this, it is useful to define a few key programmatic elements.

3.1.1 Education Leaders

STIR's programming is implemented at the ground-level 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) and Block Education Officers (BEOs).

3.1.2 Head Teachers

Head Teachers (HTs) are principals of individual schools. They may or may not be in a teaching capacity 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[8], maintaining, monitoring, and managing teachers; and to keeping abreast of new government schemes.

3.1.3 Block Education Officers

Blocks are administrative units in India, operating between the district (above) and gram panchayat (below) levels. A Block Education Officer (BEO) is responsible for one block. Additional responsibilities include handling administrative and management issues related to the schools in their block and ensuring education outcomes in India. They report to the district and state levels of the Ministry of Education.

3.1.4 Year 1 Programming sequence

The first year of STIR's programming, as implemented for the period relevant to this midline assessment, was organized into three phases of roughly equal duration (STIR Education 2016). 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.

[7] Teacher responses from our process evaluation in early 2016 also suggest that ELs are also looked upon as mentors who inspire teachers to continue when they find it tough or do not feel as though they are seeing results for their effort.

[8] Note the role 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.

The first phase is focused on "innovation." During this phase, ELs focused, in particular, on encouraging teachers to develop a positive, growth mindset about their ability to improve their teaching, and to begin to explore micro-innovations (defined below).

During the second phase — "implementation" — teachers selected one or more micro-innovations to work on putting into practice, and then reflect on the results.

During the third and final phase of the first year, teachers added a focus on "influence." This included the formation of In-School Innovation Teams (ISITs) as well as outreach to the families of five under-performing students (both defined below).

In practice, 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 particular focus of the third phase, active STIR teachers may have been influencing other 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.

3.1.5 Networks and network meetings

A central activity for teachers participating in STIR's programming 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 there were 'breaks' in the school year) during the year of programming evaluated here, for between 45 minutes and two hours per meeting, with meeting time spread across 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. ELs schedule monthly meetings to accommodate as many teachers' availability as possible and select rotating meeting locations to minimize travel time for teachers. According to STIR's data, on average, teachers attended roughly 75% of all meetings in U.P. and roughly 40% of all meetings in Delhi.[9]

3.1.6 Reflective portfolios

Reflective portfolios offer a diary- or workbook-like tool for teachers to think about their current teaching practice and to prepare for future changes in practice. These workbooks serve two key functions. First, they provide teachers a diary in which to track and reflect upon their progress and also pose questions to encourage teachers to think critically about ways to improve their teaching, classrooms, schools, and school systems. Second, they provide an accountability mechanism for STIR and allow ELs to assess how well teachers are internalizing the STIR model, changing their mindsets, and changing their classroom practices. Adequate portfolio completion also plays an important role in determining eligibility for certification (described below) and forms part of the basis for actively participating in STIR.

[9] Note there is variation in attendance rates at the individual teacher level. The averages come from STIR's network attendance data.

3.1.7 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.

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 can be found in Appendix A3.

3.1.8 Influencing other teachers: In-school Innovation Teams

As part of an In-School Innovation Team (ISIT), an actively participating STIR teacher leads two other (not actively participating) teachers in the same school who are interested in implementing micro-innovations and learning about STIR network activities. Since only some teachers in each 'treated' school become active participants in STIR's programming, STIR established ISITs as a mechanism for actively participating teachers to influence other teachers in their schools, driving toward a tipping point of school-wide change in motivation and practice. Actively 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, actively participating STIR teachers create an environment of collaborative learning between teachers in their schools. While this is an explicit activity during Year 1 programming, our understanding is that teachers who are active participants in STIR are expected to keep formally and informally working to influence other teachers in their school in the second year. Due to this influence component, the effect of STIR's program is not limited to only actively participating teachers. From an evaluation standpoint, this makes is difficult to evaluate the effects of the program on solely the actively participating teachers since we don't have a way to measure the intentional spillover[10] from the ISITs.

3.1.9 Influencing families: five under-performing students

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 whether students attend class and put in effort on their homework. It also provides teachers an opportunity to set realistic goals with the potential for quick successes.

[10] We think of spillovers as the impact (either positive or negative) on teachers who are not active participants. Ignoring spillover effects results in a biased impact estimate. If there are spillover effects then the group of beneficiaries is larger than the group of participants.

3.1.10 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. This is contingent on, to the best of our understanding:

- 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.

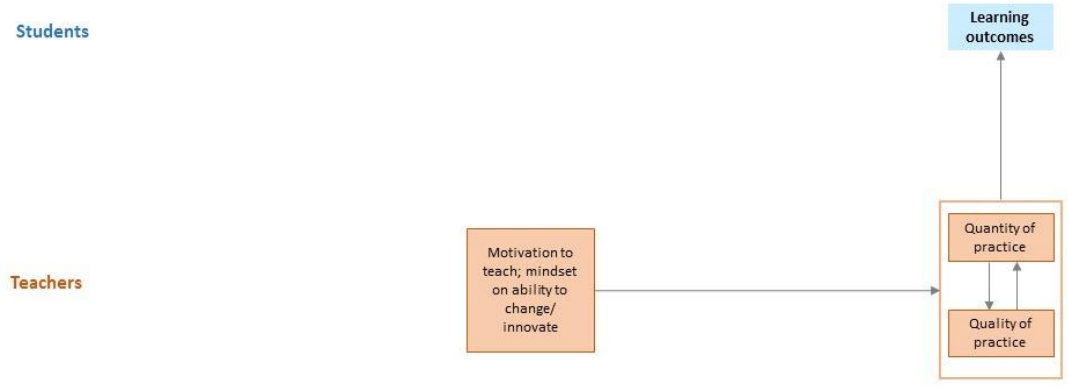
3.2 Program description and theory of change

STIR's programming is — in some ways — light touch, with ground-level implementer effort focused on the monthly activities. However, it has several moving parts that rely on a kaleidoscope of stakeholders and actions to be effective. Its success requires implementation capacity and quality (among ELs and Program Managers (PMs)), teacher within-school influence, and community and stakeholder support (HTs, BEOs, and families of students as well teachers).

3.2.1 Simple theory of change

STIR's basic theory of change (as relevant to Year 1 of programming) is that as teachers improve their motivation to teach well and their mindset that becoming a better teacher is possible, they will make the effort to change their classroom practice, which may change the quantity and/or quality of their classroom practice. In turn, these changes will facilitate improved student learning outcomes. We illustrate this basic program logic in Figure 1 and then continue to add in programmatic, conceptual, and ecosystem detail in the remainder of this section.

Figure 1: Simple theory of change for STIR's Year 1 programming



3.2.2 Detailed theory of change, with implementation processes

Getting familiar with the diagram

We illustrate our more detailed understanding of STIR's theory of change and operations in

Figure 2; this is based on extensive engagement (including a review of materials, discussions, and workshops including the materials in Appendix A4) with STIR. It also includes engagement with the literature on behavior change, education, adult education, and the production function of student learning outcomes. We recognize the diagram presented in

Figure 2 may, at first pass, appear complex. However, we encourage the reader to engage with the diagram alongside the text in this section. Understanding the program is critical to the evaluation and to expectations of what could be achieved in the program's first year. The details may also raise useful questions for future programmatic, monitoring, and evaluation work.

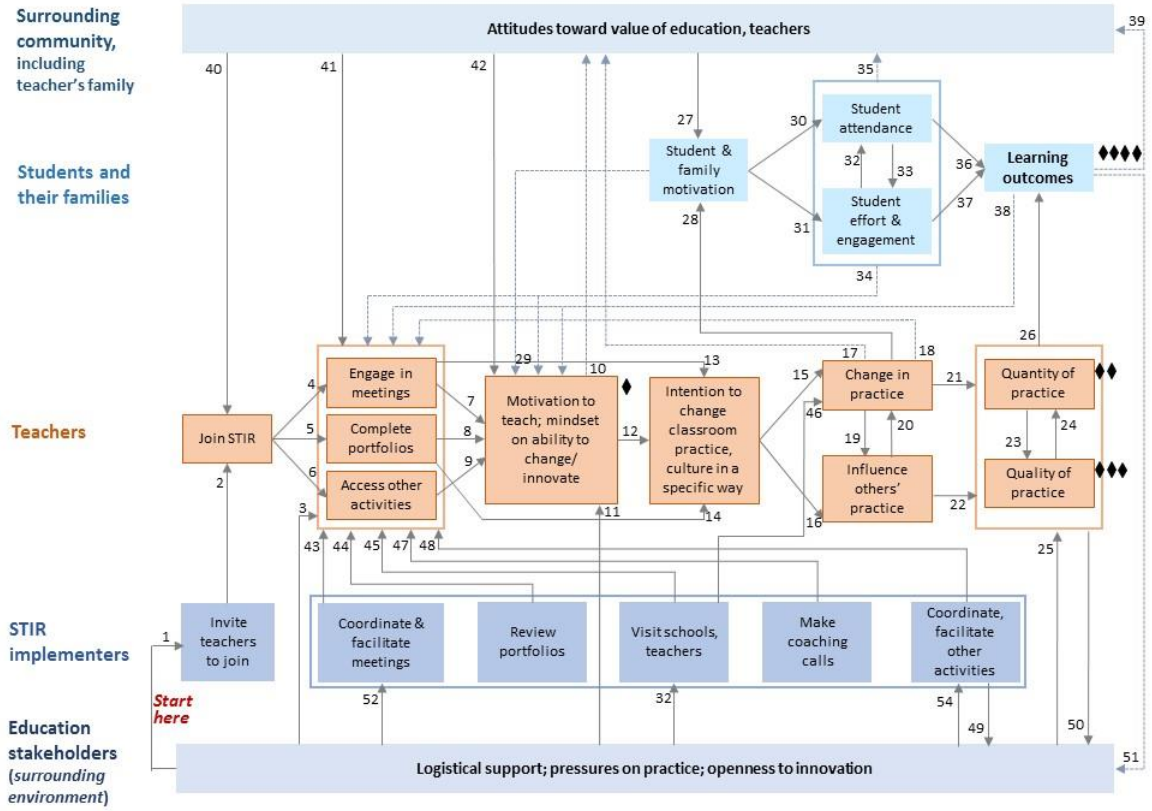
To provide some guidance for reading

Figure 2: running down the left side are a series of key actors in the ecosystem in which STIR operates: the wider community, students and their families, teachers, direct implementers of STIR programing, and education stakeholders such as Head Teachers and government officials. In each of the associated rows are actions and perspectives of these actors relevant for the implementation and success of STIR's programming. For teachers and students, these follow a left-to-right causal sequence; for communities, STIR implementers, and education stakeholders, these are discrete attitudes and actions.

The arrows provide the links between key attitudes and/or actions. We use solid lines for forward progression of the program and dashed lines for feedback loops. We number the arrows in a narratively coherent order to help guide the reader through the diagram, starting at the lower left of the diagram. We also denote, with filled-in diamonds, the points in the theory of change that we measure in the randomized evaluation. For further guidance and detail, in Appendix A5, we write out a more detailed narrative, adding specific description to the numbered arrows shown in

Figure 2.

Figure 2: Detailed theory of change and action for STIR's Year 1 programming



We have numbered the links (arrows) in the theory of change to help guide the reader through the diagram in what we feel is a narratively coherent order, starting in the lower-left corner. Solid arrows indicate forward progression through the program while dashed arrows indicate feedback loops. We also denote, with filled-in diamonds, the points in the theory of change that are the focus of measurement for the randomized evaluations.

Theory-of-change narrative

STIR's programming to enact its theory of change is largely implemented by ELs. They coordinate and facilitate the monthly network meetings to bring together teachers from across schools to share, learn, collaborate, and support one another — promoting a growth mindset and professional attitude but not a specific pedagogical approach (Dweck 2010). Teachers also work to become more thoughtful and intentional practitioners through completing questions and activities in their reflective portfolios; these workbooks are reviewed by ELs.

Teachers then take the enthusiasm, mindset, and ideas gained from network meetings and portfolio completion into their schools. During Year 1 of the program (as evaluated in the present report), much of the focus on classroom practice was on undertaking micro-innovations — small changes in and around the classroom and school. Teachers also work to spread the lessons learned through STIR to non-participating teachers in their school, through the formal channel of ISITs, as well as through existing channels of peer learning and sharing. There are also supplementary programmatic elements included in the core-plus models, described in Section 3.3. All aim at improving (extrinsic) teacher motivation and/or at improving the enabling environment (such as working with HTs) to allow teachers to more easily act on their motivation to make changes in their classrooms, schools, and education systems.

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. By further engagement with key system gatekeepers and stakeholders, STIR works to create an enabling environment that grants teachers "permission to innovate" and reduces system pressures on practice. STIR estimates that their approach costs US$ 70 per participating teacher per academic year and US$ 2.33 per student per academic year.

3.3 Program variations: context and content

3.3.1 Context

We have alluded briefly to the differences in the two different contexts in which we are conducting randomized evaluations (REs): private schools in Delhi and government schools in U.P.

School and system entry

Working with Affordable Private Schools (APS) versus government schools mandates different entry and negotiation points for the STIR program. In the government system in U.P., STIR establishes high-level government buy in (at the block level and above). In APS, however, there is no overarching authority or linking system; each APS is a business unto itself. Thus, interest in STIR had to be assessed on a school-by-school basis through talks with the HT (principal) and/or school owner.

For continuing programmatic activities as well as for the evaluation, this difference in entry points remains important. While in the government school system a supportive letter from the BEO is sufficient for entering schools in that block, entering a private school requires frequent renegotiation with individual HT.

Education Leaders and implementation

ELs are recruited and employed in a different manner in Delhi APS and in U.P. government schools. In Delhi, ELs are STIR employees and are thus recruited directly by STIR. By contrast, in U.P., ELs are government teachers with full-time jobs who volunteer (or were told to) to work with STIR in addition to their teaching and bureaucratic responsibilities. During the year of programming evaluated here, ELs 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 sum, the Delhi APS model relies on direct facilitation of teachers while the U.P. government model relies on a facilitation-of-facilitators model of delivery (Evans, Popova, and Arancibia 2016).

One other distinction between Delhi and U.P. is the timing of network meetings. 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, often requiring participating teachers to miss a full day of school on a meeting day.

3.3.2 Intrinsic and extrinsic motivating content: core and core-plus models

STIR wanted to test both a core and a core-plus model. Both models focus on intrinsic motivation[11] among teachers through network meetings, reflective portfolios, and the classroom changes that follow. The additional activities in the core-plus model focused on different extrinsic, non-financial motivation strategies that would directly influence teachers or would encourage changes in the enabling environment (namely, senior education stakeholders) to allow teachers to more easily act on their motivation to make changes in their classrooms, schools, and education systems (Ryan and Deci 2000). Participation in these activities is explicitly kept separate from active participation in core STIR activities (such as attendance at network meetings or portfolio completion). This stems from a concern that 'carrots and sticks' may be demotivating (Pink 2010). Said another way, the core-plus activities are not designed to increase teacher motivation but instead work instrumentally to increase a teacher's active participation in STIR. This is the role served by the Changemaker Certificate.

All core-plus activities (labelled as "other activities" in

Figure 2) happen outside of the network meeting and reflective portfolio process that make up the core of STIR's learning and engagement platforms. We detail the different flavors of core-plus activities in Appendix A2.

[11] Intrinsic motivation refers to doing something because it is inherently interesting or enjoyable, which can lead to high-quality learning and creativity. Extrinsic motivation refers to doing something because it leads to a separable, desirable outcome. Some types of extrinsic motivation are impoverished forms of motivation while others represent active, agentic states (Ryan and Deci 2000).

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 is having the expected effect at key points along the theory of change, namely, teacher motivation, 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 the first year of STIR's programming on teacher motivation (Delhi and U.P.) and teacher attendance (U.P. only)?
2. What is the causal effect of the first year of STIR's programming on teaching time and child-friendly practices?
3. What is the causal effect of the first year of STIR's programming on students' Hindi and math competency?

More details on the outcomes and the data collection will follow further in Section 4.

4.2 Evaluation design

4.2.1 Randomization of STIR's programming among schools

We randomly assigned the offer of STIR's programming to selected schools, in which teachers can then opt to participate actively in STIR's programming. 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's programming. Additional details, including visualization, of the randomization are in Appendix A6.

4.2.1.1 Delhi APS

Random assignment of treatment status happened in two stages in Delhi among the 180 APSs that STIR identified as meeting the qualifying criteria and that expressed sufficient interest in STIR's programming.

- 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 comparison group in a 2-to-1 ratio, so that two-thirds of the schools were slated to receive STIR's programming and one-third were not.[12]
- After this, within each stratum, schools assigned to treatment were manually grouped into four geographic clusters. Two of these clusters of schools were randomly selected to receive the STIR core program. The remaining two clusters were randomly assigned to receive one of the four core-plus programming options being trialed in Delhi. The reason for geographically clustering schools before assigning the specific treatment was to minimize the travel time required by teachers to attend network meetings.

[12] In the report, we will refer to the schools who are not offered the program as the comparison group. For readers familiar with RCT terminology, this is the same as the 'control' group.

In Delhi, unlike in U.P., the comparison group received a placebo treatment consisting of a newspaper description, health check-ups, and yoga classes. These interventions were offered to ensure the cooperation of the comparison schools during data collection.

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[13]. From among 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 comparison at a ratio of 2-to-1. That is, two-thirds of the schools were slated to receive STIR's programming and one-third were not.
- Second, all treated schools in a cluster were assigned to one variation of STIR's programming (core or core-plus).

4.3 Outcome measures and survey instruments

4.3.1 Overview of outcomes and outcome families

In this section, we review what we measure for the purposes of the evaluations and discuss the instruments we used to collect these measurements. We also discuss development of our measurement strategies and instruments and any revisions or updates to these made for midline data collection in the mentioned appendices for interested readers.

We summarize what we measure and how in Table 2. This table links with the constructs illustrated in the theory of change diagram in Figure 2; specifically, the diamonds next to key boxes in the diagram match with those in left-most column of the table below. In this table, we consider the overarching concept from the theory of change (shown in rectangles in the diagram) and then detail the specific constructs under its umbrella. In most cases, we measure an overarching concept in multiple ways.

In general, 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's programming influences 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.

[13] 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.

Table 2: How outcomes are measured

Theory of change notation (see Figure 2)	Overarching concept	Specific construct	Measurement strategy	Sample
♦	Motivation to teach	Self-assessed motivation index score	Questionnaire completed by teachers on different facets of motivation (Appendix A7)	All teachers
♦	Motivation to teach	Teacher attendance	Count of teachers present at school upon visits	All teachers
♦♦	Classroom practice: quantity	Time spent teaching	Observation of teacher practice, using modified Stallings instrument (Stallings 1977) (Appendix A8)	Sample of teachers within schools (CP sub-set of teacher[14])
♦♦	Classroom practice: quantity	Time spent off-task	Observation of teacher practice, using modified Stallings instrument (Stallings 1977) (Appendix A8)	Sample of teachers within schools (CP sub-set of teacher)
♦♦♦	Classroom practice: quality	Whether teacher smiled, joked, or laughed	Observation of teachers and students using ASER child-friendliness indicators (Bhattacharjea, Wadhwa, and Banerji 2011) (Appendix A8)	Sample of teachers within schools (CP sub-set of teacher)
♦♦♦	Classroom practice: quality	Whether students asked at least one question of the teacher	Observation of teachers and students using ASER child-friendliness indicators (Bhattacharjea, Wadhwa, and Banerji 2011) (Appendix A8)	Sample of teachers within schools (CP sub-set of teacher)
♦♦♦	Classroom practice: quality	Whether teacher incorporated local information into teaching	Observation of teachers and students using ASER child-friendliness indicators (Bhattacharjea, Wadhwa, and Banerji 2011) (Appendix A8)	Sample of teachers within schools (CP sub-set of teacher)
♦♦♦	Classroom practice: quality	Whether the teacher made use	Observation of teachers and students using ASER child-	Sample of teachers within schools (CP

[14] 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).

		of learning aides	friendliness indicators (Bhattacharjea, Wadhwa, and Banerji 2011) (Appendix A8)	sub-set of teacher)
♦♦♦	Classroom practice: quality	Whether the teacher had the students working in pairs or small groups	Observation of teachers and students using ASER child-friendliness indicators (Bhattacharjea, Wadhwa, and Banerji 2011) (Appendix A8)	Sample of teachers within schools (CP sub-set of teacher)
♦♦♦	Classroom practice: quality	Whether the teacher praised a student or showed-off students' work	Observation of teachers and students using ASER child-friendliness indicators (Bhattacharjea, Wadhwa, and Banerji 2011) (Appendix A8)	Sample of teachers within schools (CP sub-set of teacher)
♦♦♦♦	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 (SL sub-set of students)
♦♦♦♦	Student learning	Math 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 (SL sub-set of students)

We briefly mention the survey tools used below:

- **Teacher motivation** (please refer to Appendix A11 for more details): To our knowledge, there is no off-the-shelf, agreed-upon measure of self-assessed teacher. As such, we built a questionnaire to capture elements of job satisfaction — motivators — as a means of getting at motivation. We acknowledge from the outset that this is an imperfect approach, though it appears in line with other efforts (Guajardo 2011). We also acknowledge imperfect overlap between the construct of motivation we capture with these questions and STIR's evolving understanding of motivation and professional mindsets and behaviors.[15] This questionnaire draws on two types of questions (statements and situations) for 10 motivators identified from the literature and piloting, to develop an index score which we call the teacher motivation index. For each question type, we provide a positively framed and a negatively framed iteration (to mitigate

[15] Moreover, preliminary psychometric analyses suggest that we may be capturing more than one underlying construct, as discussed in Appendix A11.

acquiescent response bias), for a total of 40 questions that feed into the index (Groves et al. 2004). Items are scored on a 4-point Likert scale. The index is constructed by weighting the teacher responses to the statement items for a given motivator by the ranking (importance to teaching effort) they provided in the situation items for the same sub-category. Note that, drawing on *ex post* Principal Components Analysis, we are probably capturing multiple underlying constructs related to motivation rather than a single construct. The version of the questionnaire deployed at midline is in Appendix A7.

- **Teacher attendance** (please refer to Appendix A12 for more details): We did not track teacher attendance at baseline and only did so in U.P. at midline, as a way of trying to get a more objective measure of revealed motivation[16]. We use enumerator spot checks to measure attendance in two ways: attendance at school and presence in a classroom. For each visit to each school, enumerators recorded whether a given teacher from our sample list was present in the school and in the class as they entered and exited the school.
- **Classroom practice** (please refer to Appendix A13 for more details): We used a classroom observation tool to capture quantity and quality of teaching practice in a classroom. We used a modified version of the Stallings snapshot (Stallings 1977; World Bank 2015) tool to quantify instructional time spent on teaching, classroom management, or off-task. We added to this indicators of child-friendliness as developed by the ASER center (NCERT 2005; S. Bhattacharjea, Wadhwa, and Banerji 2011; Suman Bhattacharjea 2017). Enumerators 'sit-in' in classrooms and code student and teacher activities at intervals of five minutes each. While we did not make any changes to the time-use portion of the classroom practice tool, we did, however, make some changes to supplementary components of the tool (details in Appendix A13). The version of the observation tool used for midline is provided in Appendix A8. It is imperative to note up front that this tool is ideally deployed when enumerators can access the classroom at the time the lesson is supposed to start, regardless of teacher presence. However, we were very rarely able to access classrooms in this away.
- **Student attendance** (U.P. only): To our knowledge there is limited evidence on student attendance or an established way of capturing student attendance (Glewwe and Muralidharan 2015). Based on our process evaluation results from earlier in 2016, we found that many teachers thought one of the biggest changes coming from implementing STIR-inspired practices was a rise in student attendance. As student attendance is a pre-requisite for student learning, we captured data on attendance during midline data collection in U.P. (but had not collected it during baseline). These data were collected from the Head Teacher in this school based on administrative records. Firstly, we captured reported enrollment in the school[17] and then captured the reported school-level attendance numbers for the day.
- **Student learning outcomes** (please refer to Appendix A9 for more details): To capture learning levels we used the ASER[18] tool for Hindi (local language) and math. The ASER tool is widely

[16] We have previously tried collecting self-reported attendance, but found it tough to collect and unreliable. We also considered collecting administrative attendance data. But this information is considered sensitive and schools are hesitant in sharing these with us or allowing us to look through attendance logs.

[17] While we suspect these numbers are not perfectly accurate (given the benefits that accrue to schools on account of their enrollment numbers), we have no reason to think that reported enrollment will be differentially inaccurate across treatment and comparison schools.

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

established and popular for capturing student learning levels in India. Given that our evaluations targeted students till 8th grade as well as students from urban private schools (in Delhi), we added additional levels (stories in Hindi and a fractions sections in Math) to prevent ceiling effects. More information on the tools can be found in Appendix A9 and the tools used during midline can be found in Appendix A10.

4.4 Sampling and data collection

4.4.1 Programmatic and evaluation timeline

The new 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.

We measure different outcomes at different points in time. We administer the teacher motivation questionnaire (TM) separately from measuring teacher attendance (TA), classroom practice (CP), and student learning (SL) survey. The baseline teacher motivation survey was conducted prior to the classroom observations and both were conducted prior to the beginning of STIR's implementation of its program to minimize the chance of influencing teacher attitudes and perceptions. Note that for all data collection activities, the data collectors were blinded to whether they were visiting intervention or comparison 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 A14.

At the time of writing this report, we are in the second academic year of the teacher changemaker journey for this evaluation, reflecting on the first year. The network meetings for this academic year are scheduled to finish by mid-January 2017. The endline evaluation will take place following the completion of these meetings.

4.4.2 Delhi APS: sampling strategies and baseline and midline samples

In this section, we describe baseline and midline sampling for our different data collection needs. This is summarized in Table 3. Please find additional sampling details in Appendix A14.

Table 3: Delhi targeted and actual samples

	Teachers for motivation questionnaire	Teachers for classroom practice observation	Students for learning outcomes
Target baseline sample	All teachers in STIR and comparison 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	The 811 teachers were all teachers that STIR targeted for participation in their program based on interest in a taster session. (There were	10 students randomly selected from the main class in which each teacher observation was performed.

		approximately 439 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 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.	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 level 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 midline sample	All 1249 teachers surveyed at baseline	All teachers from the 811 list. If a school has fewer than 2 teachers left from this list, randomly select one or two teachers randomly from among those teachers who were present as on 1st July 2015	All 3367 students surveyed at baseline
Total number of units sampled at midline	657	459	1956
Population sample is representative of	All teachers	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 midline.
Reason for difference between target and actual sample at midline	The major reason of attrition was due to school and teacher level refusals, teacher dropouts and teachers not being available during the data collection window (generally due to long leave of absence).	School and teacher level 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 midline data collection	April to May 2016	July to September 2016	July to September 2016

Teacher motivation (TM) survey:

We attempted to administer the teacher motivation survey to all teachers at baseline. A total of 1249 teachers were surveyed after (a maximum of) five visits to the schools. At midline, we attemped to resurvey these 1249 teachers. The total number of teachers completing the midline teacher motivation

questionnaire (for whom baseline data are also available) was 657 (53%). These teachers form the sample used for analysis. Please refer to Appendix A14 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[19]. These 811 interested teachers formed the potential sample for classroom practice at baseline. Due to the attrition from school refusals and teacher dropouts, the total number of teachers for which classroom practice data were collected was 342[20]. 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[21]. 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. Please see Appendix A14 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[22]. Thus, the full 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 students from baseline, 1956 students were tracked and surveyed at midline. Please see Appendix A14 for details of dropouts.

4.4.3 U.P. government schools: sampling and baseline and midline samples

In this section, we describe briefly the baseline and midline sampling for our different data collection needs. This is summarized in Table 4. Please find additional details in Appendix A14.

Table 4: U.P. targeted and actual samples

	Teachers for motivation questionnaire	Teachers for classroom practice observation	Students for learning outcomes

[19] 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 comparison.

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

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

[22] 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.

Target baseline sample	All teachers in STIR and comparison 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	On average, 3 teachers were randomly selected from each of the 270 STIR and comparison 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 midline sample	All 1145 teachers surveyed at baseline	All 838 teachers surveyed at baseline.	All 7386 students surveyed at baseline
Total number of units sampled at midline	755	747	4560
Population sample is representative of	All teachers	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 midline.
Reason for difference between target and actual sample at midline	School and teacher level refusals, teacher dropouts and teachers not being available during the data collection window (generally due to long leave of absence).	School and teacher level 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 midline data	April to May 2016	July to September 2016	July to September 2016

collection			

Teacher Motivation survey:

We attempted to administer the teacher motivation survey to all teachers at baseline. A total of 1145 teachers were surveyed after (a maximum of) three visits to the schools. At midline, we looked to resurvey these 1145 teachers. The total number of teachers completing the midline teacher motivation questionnaire was 755 (66%). These teachers form the sample used for analysis. Please see Appendix A14 for details of dropouts.

Classroom practice 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.[23] 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. Please see Appendix A14 for details of dropouts.

Student learning 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 (62%) students were also tested at midline. Please see Appendix A14 for details of dropouts.

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 midline

Sample list	Baseline timeline	Baseline sample	Midline sample	Midline timeline	Attrition	Percentage attrition
U.P. TM List	Feb-Mar 2015	1145	755	Apr-May 2016	390	34%
U.P. CP List	July-Aug 2015	838	747	Jul- Sept 2016	91	11%
U.P. SL List	July-Aug 2015	7386	4560	Jul- Sept 2016	2826	38%
Delhi TM List	Feb-Apr 2015	1249	657	Apr-May 2016	592	47%
Delhi CP List[24]	July-Nov 2015	333	248	Jul- Sept 2016	85	26%
Delhi SL List	July-Nov 2015	3367	1956	Jul- Sept 2016	1411	42%

[23] Unfortunately, the precise details of how we ended up with 28 more teachers than we were targeting are lost to poor documentation and staff turnover.

[24] Note the numbers mentioned here tracks the 333 teachers from baseline to midline. As mentioned in the previous section, more teachers were added at midline. The final number of teachers sampled at midline is 459.

Teacher dropout from schools happens for a variety of reasons — in U.P. the most common reason was official transfer of teachers to other government schools whereas in Delhi the most common reason was dropout from the APS or teaching altogether.[25] Similarly, students can drop out of school for a variety of reasons, reaching from changing schools to drop out of school for work or family support. At the same time, teachers and/or students might still be in school, but may be absent on the day or refuse to participate in follow-up data collection. In all these situations, midline data are not available and we consider attrited. It is important to note that in our studies the attrition numbers represent attrition from our sample list and maintain this distinction between dropouts from schools and dropouts from the list. The latter apart from including the former also includes absenteeism and refusals[26].

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. Mentioned in this section is the summary of our findings. For a more detailed understanding and explanation please refer to Appendix A15. We ran the following tests to check for trends in our samples:

1. **Tests for differential attrition across the treatment and comparison groups** by comparing teacher dropout across treatment status (control, core, core-plus) in our study sites (U.P., Delhi) for the survey rounds (Teacher Motivation, Classroom Practice, Student Learning). If differential attrition was absent, overall trends for attriting teachers and students are expected to be comparable across treatment and comparison groups. Overall, we do not find evidence of differential attrition across treatment and comparison groups for students or teachers.
2. **We test for differential attrition trends by baseline characteristics in a more direct way**. Specifically, we try to explain teacher and student level dropout as a function of baseline characteristics, allowing for differential trends in (the two types of) STIR and comparison schools. We do not find evidence of difference in terms of baseline characteristics for teachers or students surveyed at midline at the 5 % level of significance.
3. **We also reassessed our power calculations** in order to gauge how the minimum effect size we were powered to pick up was affected by our attrition. Please check Appendix A16 for details.

In summary, we do not find evidence of differential attrition which increases our confidence that our results are not driven by trends in attrition. While the attrition does have an implication on the studies' power; the effect was limited.

[25] 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 APS schools between both our two rounds of baseline measurement (from teacher motivation 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 onto teaching in other APS schools, 2% were teaching in government schools, 23% had moved onto teaching private tuitions/tutoring, and the remaining had dropped out for other reasons.

[26] 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.

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. We focus on the impact estimation. All descriptive statistics are in Appendix A17 (for classroom practices), Appendix A18 (for student learning outcomes), and Appendix A19 (for teacher motivation index).

4.5.1 Measuring the effect of STIR's programming: ITT and TOT estimation strategies

To understand the effects of STIR programming 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 effect using different approaches to estimate the treatment-on-the-treated effect.

4.5.1.1 The intent-to-treat (ITT) estimator

Our measure of STIR's overall, system-wide causal impact on teacher and student outcomes will be based on the school-wide outcomes ("intent-to-treat" (ITT) approach).[27] The ITT approach compares overall outcomes for teachers in "treated" schools (which received the offer of STIR programming) to overall outcomes for teachers in comparison schools. This estimation includes teachers who were active participants in the program and those who weren't. The intention-to-treat results therefore capture the direct impact of STiR as well as the spill over effects on non-active teachers. Operationally, STIR introduced an element of 'rationing' to ensure networks were maintained at a manageable size. In the lack of this rationing element one would expect take up of the program to be higher[28]. The key assumption validating the ITT approach is random assignment of the STIR programming on the school level and excluding "contamination" between treated and comparison schools.[29]

Due to the randomized nature of the study, this estimate yields consistent estimates for all teachers in the study, *i.e.*, not just limiting to those who are active participants. For thinking about what is achievable at scale, we think that this estimate is the most useful and appropriate estimate of STIR's causal impact on the outcomes of interest.

4.5.1.2 The treatment-on-the-treated (TOT) estimator

A second interesting, but more complicated, approach aims to quantify the effect of STIR's programming for those teachers that choose to actively participate. This analysis is conventionally termed a "treatment-on-the-treated" (ToT) analysis. It is important to emphasize the fact that — by design — active participation in STIR's programming is voluntary. 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 ToT estimate will only help us understand the potential effects of STIR's programming among the sub-set of teachers with the characteristics that make them likely to become active participators.

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

[28] For more details of rationing and implications please refer to Appendix A5.

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

Given the importance for STIR's internal learning we have looked to estimate the teacher level effects of STIR's programming by using three main approaches — an IV/LATE estimate; small-schools analysis and an observational analysis. More details on these approaches along with the limitations are mentioned in Section 4.6.

Note that the ToT analysis was conducted only for the teacher level indicators. Students move through different grades through the course of the evaluations and may or may not be taught by teachers who are active participants of the STIR program. Further still 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. Given this it is tough to conceptualize what being an actively participating student means.

Estimating the effects among participating teachers: focusing on small schools

Although technically another sub-group analysis — planned after the CARP — we consider our small schools analysis separately, since it serves a different analytical purpose. We divide the sample into smaller and larger schools, with the cut-offs determined by drop-offs in the proportional participation rate. (We define "small schools" as those with 5 or fewer teachers in Delhi and 3 or fewer teachers in U.P.) Schools with fewer teachers tend to have higher rates of participation in STIR's program as a proportion of total teachers in the school. As such, looking only at smaller schools may allow us to estimate the effect of the STIR program when the participating proportion of teachers in a school is relatively high. Thus, in a way this analysis may shed additional light on the 'treatment-on-the-treated effect,' which is the effect for teachers who directly participate in STIR's programming by attending meetings. It is, roughly, a treatment-when-more-are-treated effect. In Delhi, the average proportion of teachers in STIR schools who are active in STIR (attend at least one network meeting) is 66% (compared to 44% in larger schools) and in U.P., the average proportion is 54% (compared to 40% in larger schools). As school size is a baseline covariate, this is a valid subgroup analysis. As with all subgroup analyses, results are only valid for these smaller schools and may not generalize if there are systematic differences between smaller and larger schools.

Estimating the effects among participating teachers: IV/LATE

Broadly, any "treatment-on-the-treated" analysis aims to isolate the effects of STIR's programming for only those teachers who actively participated in STIR. An 'instrumental variable' offers a strategy to isolate this effect.

This analysis gives some insight into the causal effect of STIR's programming specifically for active teachers as well as for the students of active teachers.[30] 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 true treatment-on-the-treated effect.[31,32] Given this, we believe that the

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

[31] 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.

'true' treatment-on-the-treatment effect will lie somewhere between the ITT estimate and the treatment-on-the-treated estimate 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 STIR programming (treated schools) but who did not individually participate in any STIR meetings received zero benefit from 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 true treatment-on-the-treated effect.

Estimating the effects among participating teachers: Observational analysis

On STIR's request, we will also directly compare teachers in the treatment group who participated actively in STIR with teachers in the comparison group, excluding teachers in the treatment group who didn't participate in STIR. For this analysis, we will 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 comparison schools. This analysis is an attempt to a relationship between participating in some amount of STIR programming and the outcomes for different teachers.

For this analysis, we will use three definitions of 'active participation'. These are 1) Active defined as teachers attending at least one meeting (excluding the taster meeting) 2) Active defined as teachers attending at least half the meetings[33] (excluding the taster meeting) 3) Active defined as teachers attending at least three-fourths the meetings (excluding the taster meeting)[34].

In order for results from this analysis 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 would have had similar outcomes to the comparison group teachers if they hadn't participated in the program. However, due to significant personal initiative teachers must demonstrate to participate in STIR, **we believe this assumption is unlikely to hold**. As the definition becomes stricter (*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 difference between active and inactive teachers is true).

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's programming is differentially effective for different types of teachers. Three of these subgroups – split by teacher sex, teacher years of experience, and teacher baseline

[32] 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.

[33] Given the total of eight meetings, teachers who attended more than 4 meetings would be classified as active

[34] Given the total of eight meetings, teachers who attended more than 6 meetings would be classified as active

motivation levels – were detailed before we began analyzing the data.[35] 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's programming but equally compelling reasons why this might be the case for female teachers (as detailed in Appendix A5).

In Appendix A20, 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 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 A20 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 core *v.* core-plus) due to sample size considerations.

To conclude, it is important to note that in interpreting the sub-group analyses we would be less focused on individual estimates but rather look for an overarching trend; *e.g.*, if STIR's programming seems to have a differential impact for all 'low-experience teachers' across all estimates. This would help us make the clearest learning statement for STIR. 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 say time-use practices.

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. For those interested in refreshing their understanding of statistical inference and the potential for false positives,

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

- Teacher baseline motivation — Using the baseline teacher motivation index, teachers are split into three categories (low, medium, high). 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 A5).
- 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.

please see Appendix A21 and for a more detailed discussion on multiple hypothesis correction please refer to Appendix A22.

In this evaluation, we take two main approaches, depending on the specification we are looking at. While we consider the Free Step Down Resampling Method (FSDRM) as the most powerful, we are unable to use this method for core *v.* control (C) and core-plus *v.* C specifications, given how schools were randomized. Please see Appendix A22 for details. We use a combination of FSDRM and Holm-Bonferroni corrections, as described below and as summarized in Table 6.

- **All-STIR (core and core-plus taken together) *v.* C** — For the analysis taking all treatment schools together (reported as "All-STIR") we have used for the Free Step Down Resampling Method (FSDRM). This is true for the main ITT estimate as well as the sub-group estimates.
- **Core *v.* C** — For all analysis where we are comparing purely the "core" schools to comparison schools we will use the Holm-Bonferroni correction.
- **Core-plus *v.* C** — For all analysis where we are comparing purely the core-plus schools to comparison schools we will use the Holm-Bonferroni correction.

Table 6: Summary of multiple hypotheses correction

Family	All-STIR *v.* C		Core *v.* C		Core-plus *v.* C	
	Main	Subgroups	Main	Subgroups	Main	Subgroups
Teacher motivation	NA	FSDRM	NA	NA	NA	NA
Time use	NA	FSDRM	NA	NA	NA	NA
Child friendliness	FSDRM	FSDRM	Holm-Bonferroni	NA	Holm-Bonferroni	NA
Student learning	NA	NA	NA	NA	NA	NA

Note: For the IV/LATE and observational analysis, we did not conduct multiple hypothesis correction[36]. The readers should keep that in mind as they are interpreting results. **All the other results presented are adjusted for these corrections as needed.**

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 APS 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 ITT estimates

Analysis of Covariance (ANCOVA) has been used to estimate the school level ITT effect of the STIR program (McKenzie 2012).[37] The specifications mentioned have been run separately for teachers from Delhi and Uttar Pradesh. We employ the following specification:

[36] Given that the uncorrected p-values were largely insignificant, we felt the multiple hypotheses corrections were not necessary. What this implies is that we would still expect one out of twenty (at the 5% level of significance) and one out of 10 (at the 10% level) hypotheses to come up significant even though they may be driven by statistical noise.

$$Y_{1ij} = \beta_0 + \beta_1 * T_j + \beta_2 * Y_{0ij} + \gamma * X_{ij} + \omega_j + \varepsilon_{ij}$$

where,

- Y_{1ij} is an individual teacher's or student's (belonging to school j) outcome at midline
- 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.*, core and core-plus clubbed together, core treatment, or core-plus treatment, depending on the regression)
- X_{ij} is a vector of covariates[38][39]. 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

β_1 will be our estimate of interest (effect size). The standard errors are clustered at the school level. The above specifications would be run for three main treatment (assignment) types. We have broken these intro 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: core (1.0) and core-plus (2.0) will be clubbed, so that the effects of STIR programming in general can be seen.
2. Core *v.* C
3. Core-plus *v.* C

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

4.6.1.1 Teacher motivation family

Motivation index

Our main outcome is the teacher motivation index. Before fitting the regression, this index has been standardized (with mean = 0 and standard deviation = 1). The results presented here are thus in standard deviation terms.

Our teacher motivation index comes directly from the teacher motivation tool mentioned in sections above. We performed several checks to test, ex post, the validity of the teacher motivation survey. The results suggest lower reliability on the tool than STIR and IDinsight had previously expected and it

[37] While the randomized design (and primary indicator of treatment vs comparison) 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 pre-post values and undertake a 'difference in differences' analysis. Baseline indicators are used as covariates in an ANCOVA model.

[38] For missing values of covariates we used mean imputation and categorization (Puma, Michael J., Robert B. Olsen, Stephen H. Bell, and Cristofer Price 2009)

[39] For details on the covariates used we request the reader to please refer to Appendix A18.

appears that questions used in the teacher motivation index seem to be measuring underlying concepts. Please refer to Appendix A11 for more details.

Teacher attendance (U.P. only)

For observed teacher attendance we look at two indicators – attendance at school and presence in a classroom. Furthermore, we restrict our analysis to only U.P. since we were unable to collect these data from schools in Delhi (please see Appendix A12 for details). Given that we added these indicators to data collection only at midline, we could not use baseline levels as a covariate.

4.6.1.2 Time use and child friendliness family

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

Time use

For time use, we consider two main outcomes: time teaching and time off-task[40]. For both these indicators, regressions are fit 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 off each other – teachers are coded as either teaching, engaged in classroom management or off-task. Hence, we could expect for instance an increase in teaching to be accompanied by a reduction in 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 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 only collected as part of the midline data collection. Hence their regressions did not include baseline outcomes.

4.6.1.3 Student attendance

The student attendance regression varied from the others. The student attendance data were collected at the school level. These data were collected by enumerators by asking principals/ head teachers about grade wise enrollment of students and total number of student present in the school today. Student attendance was defined as the percentage of total enrolled students (sum of grade wise enrolled) present in the school today. Note we conduct this analysis only for U.P. These data were only collected as part of the midline data collection.

This meant that, compared to the regression specification mentioned above, the student attendance regression would not include baseline values. At the same time, we do not use teacher level covariates[41]. Hence, we use only network and enumerator dummies as control.

[40] 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 four observations

[41] Note given that the level of the regression is the school, this roughly means that we do not include any individual level information (denoted by subscript 'i'). There is only one error term, at the school level.

4.6.1.4 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.

Initially we had planned to use the ordered logit model to see the marginal effect of treatment on the probability of child being at a certain learning level (in both Math and Hindi). However, we eventually decided to use a simple OLS regression with the learning level as the outcome variable to gauge the overall effect on learning levels. This was done based on our understanding of STIR's ToC and to help make a cleaner learning statement. More details can be found in Appendix A23.

The OLS model is mentioned at the very beginning of Section 4.6.1. The learning level variables were standardized before fitting the regressions; and all effect sizes presented are in standard deviations.

4.6.2 IV/LATE estimates

The IV/LATE estimation exploits the fact that participating teachers have to be in schools that were assigned to be offered STIR's programming, 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). Crudely, 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. We will estimate the following regression specification:

First Stage:

$$\text{IV(Active)}_{ij} = \beta_0 + \beta_1 * T_j + \beta_2 * Y_{0ij} + \gamma * X_{ij} + \omega_j + \varepsilon_{ij}$$

Second Stage:

$$Y_{1ij} = \beta_0 + \beta_1 * \text{IV(Active)}_i + \beta_2 * Y_{0ij} + \gamma * X_{ij} + \omega_j + \varepsilon_{ij}$$

Where,

- Y_{1ij} is an individual teacher's (belonging to school j) outcome at midline
- 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.*, core and core-plus clubbed together, core treatment, or core-plus treatment, depending on the regression)
- X_{ij} is a vector of covariates. At the teacher level 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

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

4.6.3 Observational analysis

For the observational analysis, we first restrict the sample by removing teachers in treatment schools who did not participate in STIR and then use the following specification:

$$Y_{1ij} = \beta_0 + \beta_1 * \text{leastActive}_{ij} + \beta_2 * \text{partiallyActive}_{ij} + \beta_3 * \text{fullyActive}_{ij} + \beta_4 * Y_{0ij} + \gamma * X_{ij} + \omega_j + \varepsilon_{ij}$$

Where,

- Y_{1ij} is an individual teacher's (belonging to school j) outcome at midline
- Y_{0ij} is an individual teacher's (belonging to school j) outcome at baseline
- leastActive_{ij} is a dummy variable which is 1 if the teacher (in a STIR school j) falls in the category of least active participation *i.e.*, has attended only one meeting
- $\text{partiallyActive}_{ij}$ is a dummy variable which is 1 if the teacher (in a STIR school j) falls in the middle category of active participation *i.e.*, has attended at least half (but less than three fourths) of the meetings.
- fullyActive_{ij} is a dummy variable which is 1 if the teacher (in a STIR school j) falls in the highest category of active participation *i.e.*, has attended at least three-fourths of the meetings.
- X_{ij} is a vector of covariates. At the teacher level 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

β_1, β_2 and β_3 are the coefficients of interest (effect size) for least, partially and fully active teachers respectively. Note each teacher can be part of only one of the three categories. In the results, we report all three; *i.e.*, for each regression we compare outcomes of teachers in each of the three categories to teachers in comparison schools.

5 Results and interpretations

We divide this section into: (1) the school-wide (intent-to-treat (ITT) results) and (2) our three ways of approximating teacher-level effects for active teachers. Our review of the school-wide results is organized by the *families* (or groups) of indicators. For each family, we provide a summary of similar results from other studies (wherever possible). We then provide an overview of the results (main ITT and subgroup[42]). At this stage, midway in the evaluations, we refrain from discussions around the interpretation and practical significance[43] (Kirk 1996) of the results. We feel it would be best to wait till endline results before offering interpretations and potential implications. Our review of the teacher-level results includes a quick summary of the additional analyses conducted (namely the IV/LATE, small schools and observational analyses). Note, given the sheer number of analysis and specifications we do not comment on each result individually (please see Table 1 for a summary of the number of total and significant estimates).

The results are presented in full in the *Results Appendix*.[44,45] We fit regressions both with and without covariates. In most cases the inclusion of covariates altered our estimates of the impact only slightly while significantly increasing the precision of the estimates. Hence here we present results only for the regressions with controlling for covariates.

For the school-level and sub-group analyses, we encourage the reader to refer to tables A25 and A26 in the Appendix A20 indicating the number of teachers while interpreting results.

[42] **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.

[43] Apart from statistical significance, another important concept is that of practical significance. Statistical significance indicates the difference in the means of the comparison and treatment groups that is not driven by statistical 'noise'. However statistical significance is influenced by sample size – statistically significant differences can be found even with very small differences if the sample size is large enough and vice versa. At endline, we will thus try and extend our interpretation beyond statistical significance. Practical significance asks the larger question of whether the results are large enough to have *real* meaning. Come endline, we will refer to contemporary literature and our confidence intervals (standard errors) to help provide a coherent and comprehensive picture. We will comment on what effect sizes we can confidently rule out at the 95% level of confidence, we will say what effect size (and above) we can confidently say STIR's program is not having.

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

[45] 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.

5.1 ITT Estimates

5.1.1 Teacher motivation

5.1.1.1 Evidence and expectations

Motivation

To our knowledge, no evidence exists that helps set expectations of how much teacher motivation or the related concepts of job satisfaction of 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 important outcome in itself.

Teacher attendance (U.P. only)

In a recent meta-analysis exploring how to improve teacher attendance, Snilstveit *et al.* consider the evidence from low- and middle-income countries. Pooling the results of three teacher incentive studies that measure teacher attendance[46], they find a pooled change of 0.07 standard deviations (not significant at α=0.05) (Snilstveit et al. 2015; Duflo, Hanna, and Ryan 2012; Muralidharan 2012; Glewwe, Ilias, and Kremer 2010). For example, a performance-pay (based on student performance) program in India led to no relative (to a comparison group) gains in teacher attendance (Muralidharan 2012).

5.1.1.2 Results

Motivation index

School-wide results

We present our motivation results by study site and with the treatment first grouped by All-STIR (meaning core and core-plus combined), then core and core-plus separately, and finally disaggregated to flavors within core-plus (Table R2 in the *Results Appendix*); the core-plus flavors are described in more detail in Appendix A2. We present all estimates in standard deviation (and not index value) terms.

Amongst the 13 estimates, we only find a significant impact in 1: the 'local recognition' core-plus flavor in Delhi APS. On average, the teacher motivation index for teachers in the local recognition STIR schools is 0.25 sd higher than the average teacher motivation index for teachers in comparison schools. This difference is significant at the 5% significance level[47]. In interpreting the results, we would again like to acknowledge the limitations of the teacher motivation tool as mentioned in Section 4 and Appendix A11 and encourage the reader to consider implications in interpreting the results.

[46] The three studies that included teacher attendance as an outcome include one of installing cameras in the schools to monitor attendance and two in which teachers were rewarded financially (at an individual or school-wide level) for student performance. These results are available in the *Results Appendix*.

[47] From here on significance at 10% will be represented by α=0.10 (p-values marked with one asterisks (*) in tables); significant at 5% will be represented by α=0.05 (p-values marked with two asterisks (**) in tables) and significance at 1% will be represented by α=0.01 (p-values marked with three asterisks (***) in tables)

Sub-group results[48]

For Delhi APS, we find no significant results among the 5 subgroup estimates. Please see Table R3 in the *Results Appendix*.

In U.P. Govt. schools (Table R4 in the *Results Appendix*) we find significant results in four (including two block estimates) amongst 18 subgroup estimates[49]. Teacher motivation index values for teachers in the 'high' motivation category in STIR schools are on average 0.26 sd (significant at α=0.05) *less* than their comparison school counterparts (with 'high' motivation). Similarly, teacher motivation index values for female teachers in STIR schools are on average 0.19 sd (significant at α=0.10) *less* than their female counterparts in comparison schools.

Teacher attendance (U.P. only)

School-wide results

For both the indicators used as proxies for teacher attendance namely observed present (in school) and observed in class we find no significant results among the 6 estimates (Table R5 in the Results Appendix).

Sub-group results

We find significant estimates for three out of the 36 estimates (for both indicators combined); all three of these are for specific blocks. See the *Results Appendix* (Table R6) for details.

5.1.2 Classroom culture and practice

5.1.2.1 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; the researchers find that a program focused on changing instructional practices leads to a 7%-point gain in teaching time over one academic year, accompanied by reductions in

[48] 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.

[49] The interpretation of estimates and significance will be slightly different from a 'normal' regression. For example, both our treatment indicator (for comparison and pooled-treatment) as well as say teacher sex are dummy variables. Given how these parameters 'interact' in our regression, the final estimates (presented in the tables) are generated by linearly combining the coefficient value of the constant term (which represents treatment ==0 and sex==0) and coefficient values of the others (say treatment ==1 and sex==1). Thus, the interpretation of any sub-group estimate and its statistical significance is relative to the comparison/category combination of the same category. For example, the estimate and significance for say low motivated teachers in Table R3 and Table R4 in the *Results Appendix* represents the estimate and significance of low-motivation category teachers in treatment schools compared to low-motivation teachers in comparison schools. The important distinction to maintain here is that comparison is not with the entire comparison group but a specific sub-section of it.

off-task time (Costa, Cunha, and Bruns 2016). 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)[50].

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).

5.1.2.2 Results

Time use

School-wide results

In this section, we present the regression results on our time-use indicators. As mentioned in the previous section, there are only three ways in which teachers divide their time, namely teaching, off-task and management; we look for changes in teaching and off task time allotments.

We find no significant estimates of the time-use indicators out of a total of 6 estimates in APS Delhi (Please refer to Table R7 in the *Results Appendix*)**.**

In U.P. government schools, we find significant estimates for three out of 6 estimates. As shown in the *Results Appendix*, teachers in core-plus schools on average are observed as teaching 5 percentage points *more* (significant at α=0.05) and correspondingly on average are observed as off-task 4 percentage points *less* (significant at α=0.05) as compared to their comparison school counterparts. There is also evidence of a 2 percentage-point *reduction* (significant at α=0.10) in time observed off-task (on average) for teachers in STIR schools (core and core-plus taken together) compared to comparison schools.

Sub-group results

In Delhi APS, (Table R9 in the *Results Appendix*) we find two significant estimates out of a total of 10 estimates (across both indicators). Teachers in STIR schools with 'medium' baseline motivation are observed on average as teaching 12 percentage points *more* (significant at α=0.05) than comparison school teachers with 'medium' baseline motivation whereas, teachers with 'low' baseline motivation are observed on average to be off-task 3 percentage points *more* (significant at α=0.05) than their comparison school counterparts.

In U.P., (Table R10 in the *Results Appendix*) we find significance in five out of the total of 36 estimates.

- Teachers in STIR schools in U.P. with 3 or more years of teaching experience are, on average, observed teaching 3 percentage points *more* and are on average observed off-task 3 percentage points *less* (both significant at α=0.10) than their comparison school counterparts.

[50] 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.

- Male teachers in STIR schools in U.P. are observed on average as teaching 6 percentage points *more* and are on average observed as off-task 5 percentage points *less* (both significant at α=0.05) than male teachers in comparison schools.

Child friendliness

School-wide results

We find no significant results on any of the child-friendliness indicators, for any of the program variants (STIR, core, and core-plus) in either Delhi or U.P. (Table R11 and Table R12). Note: all estimates presented here are corrected for multiple hypothesis testing (please check Section 4.5.3 for details)[51].

Sub-group results

In Delhi, we find significant estimates in four amongst 35 total estimates (Table R13 in *Results Appendix*). Teachers in STIR schools with 'medium' baseline motivation are observed on average as smiling or laughing (at least once during the observation window) 11 percentage points *more* and are observed on average as praising or displaying students' work (at least once during the observation window) 10 percentage points *more* than teachers in comparison schools with 'medium' baseline motivation (both the results being significant at α=0.05). Teachers with low experience were on average observed as smiling or laughing (at least once during the observation window) 16 percentage points *less* (the result being significant at α=0.05) in STIR schools compared to comparison schools while teachers with high experience on average were observed as using learning aides (teaching learning materials) 10 percentage points *more* (the result being significant at α=0.10) in STIR schools as compared to comparison schools.

In U.P. (Table R14 in the *Results Appendix*), we see significant estimates in nineteen out of a total of 126 estimates.

For the motivation subgroup, results are summarized as below:

- 'Low' motivation teachers are on average observed as smiling or laughing (at least once during the observation window) 7 percentage points *less*; are on average observed as praising or displaying students' work (at least once during the observation window) 5 percentage points *less*; and on average are observed as referring to students by name (always during the observation window) 8 percentage points *less* in STIR schools as compared to comparison schools (all three of the results being significant at α=0.10).
- Teachers in STIR schools with 'medium' baseline motivation are on average observed as smiling or laughing (at least once during the observation window) 7 percentage points *more* (result being significant at α=0.10) than 'medium' motivated teachers in comparison schools. At the same time however, they are observed as praising or displaying students' work (at least once during the observation window) 6 percentage points *less* (result being significant at α=0.05) than 'medium' motivated teachers in comparison schools.

[51] We ran our base regression (STIR v C) for a 'child friendliness index' (combining all the individual indicators) as a robustness check. We do not find any evidence of impact. Hence, we can conclude that there are no patterns across indicators or when they are aggregated. In general, we felt an index would not be the most useful from a learning perspective for STIR. Index values may be driven by a particular-indicator (as opposed to all indicators displaying the same trend). Given that some of the indicators map more directly to STIR's ToC as compared to others, we felt presenting results at the most granular level would be more useful.

- Teachers in STIR schools were observed as smiling or laughing (at least once during the observation window) 5 percentage points *more* (result being significant at α=0.10) than 'high' motivated teachers in comparison schools.

Teachers in STIR schools (with high experience) are on average observed as praising or displaying students' work (at least once during the observation window) 4 percentage points *less* (result being significant at α=0.05) than (high experience) teachers in comparison schools.

Teachers in STIR schools are on average observed as praising or displaying students' work (at least once during the observation window) 5 percentage points *less* (result being significant at α=0.05) than female teachers in comparison schools; and for the 'learning aides' indicator where male teachers in STIR schools on average are observed as using learning aides (teaching learning materials) 10 percentage points *more* (the result being significant at α=0.05) than male teachers in comparison schools.

5.1.3 Student attendance (U.P. only)

5.1.3.1 Evidence and expectations

There seem to be limited studies that look at the impact on student learning, for an intervention similar to STIR. We present here results from varied interventions. Across three studies of teacher incentives that measure student attendance[52], Snilstveit *et al.* find a 0.01 standard deviation increase in student attendance relative to students who receive no intervention (not significant at α=0.05) (Snilstveit et al. 2015). By contrast, to boost student attendance, cash transfers (38 pooled studies) lead to a 0.13 standard deviation increase in attendance (significant at α=0.05) and school feeding programs (6 pooled studies) lead to a 0.09 standard deviation gain in attendance (significant at α=0.05) (Snilstveit et al. 2015; Duflo, Hanna, and Ryan 2012; Muralidharan 2012; Glewwe, Ilias, and Kremer 2010).

5.1.3.2 Results

The student attendance analysis was run at the school level (please refer to Section 4.3.1 for details on how these data were collected). While we do suspect that schools may misreport student attendance, we do not expect differential misreporting between treatment and comparison schools allowing us to make a valid comparison.

While student attendance is not one of our 'primary' indicators, it sheds light on important link between improved classroom practice and student learning. In our process evaluation, many teachers reported increased student attendance as a 'change' they observed in their classrooms as part of their journey with STIR's program.

School-wide results

As can be seen in Table R15 in the *Results Appendix*, **there are no significant estimates among the 3 estimates.**

[52] The three studies that included teacher attendance as an outcome include one of installing cameras in the schools to monitor attendance and two in which teachers were rewarded financially (at an individual or school-wide level) for student performance.

Sub-group results

We cannot conduct the same subgroup analyses used for other outcome families since this analysis focuses only at the school level.

5.1.4 Student learning

5.1.4.1 Evidence and expectations

While there is limited literature exploring the precise effects of teacher motivation or child-friendly practices on student learning outcomes, there is a broader literature on improving student learning outcomes. These 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.[53, 54]

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 standard deviations 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 training (from 17 studies) leads to 0.12 standard deviation gains in learning outcomes (significant at $\alpha = 0.001$) while teacher incentives (from 8 studies) lead to 0.09 standard deviation gains in learning (significant at $\alpha = 0.05$) (McEwan 2015).[55]

5.1.4.2 Results

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

[53] After one year, a remedial education program in Bombay targeting under-performing students led to 0.15 (language) and 0.16 (math) standard deviation gains in learning outcomes (Banerjee et al. 2007). Another remedial education program, in Andhra Pradesh, led to a 0.74 standard deviation 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) standard deviation 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 around the world have achieved greater gains in learning outcomes as compared to teacher incentive programs (Snilstveit et al. 2015).

[54] Note that 'teacher training' and 'in-service professional development' cover a broad range of activities, which are not always precisely described in evaluations. 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). 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). Certain classroom management strategies, especially interventionist behavior management strategies, are associated with gains in student learning outcomes, at least in a small sample of elementary school teachers in the United States (Sowell 2013). 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 standard deviation gain in teaching quality (as measured by CLASS) associated with 0.11 standard deviations in language and math scores (Araujo et al. 2016).

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

Hindi levels

School-wide results

In both Delhi and U.P. (Tables R16 and R17 in *Results Appendix*), we find significance in none of the 6 estimates (in total) for Hindi learning levels.

Sub-group results

We did not conduct any sub-group analysis for student learning outcomes. This was due to two reasons. Firstly, we did not collect many covariates at the student level – we only collected age and sex, which already enter our regression as covariates. Secondly, given STIR's ToC, there is no theoretical reason to expect differential impact for students of different sex.

Math levels

School-wide results

In U.P. (Table R19)**, we find no significant estimates.**

In Delhi (Table R18)**, we find significance in two out of three estimates for math levels.** We find an 'overall' effect as seen from the OLS estimates. On average math levels for students in STIR schools was 0.08 sd *higher* (the result being significant at α=0.10) than average levels of students in comparison schools. There is also an effect in core model schools, where math levels for students on average is .11 sd *higher* (the result being significant at α=0.05) than their comparison school counterparts.

Sub-group results

We did not conduct any sub-group analysis for student learning outcomes. This was due to two reasons. Firstly, we did not collect many covariates at the student level – we only collected age and sex, which already enter our regression as covariates. Secondly, given STIR's ToC, there is no theoretical reason to expect differential impact for students of different sex.

5.2 TOT Estimates

The reader is reminded that for the three teacher level analyses, we restrict the analysis only to teacher level indicators[56].

The small school analysis was run similarly to other sub-group analysis (*i.e.*, with an interaction term). The hypothesis was to see if there was higher impact in schools with fewer teachers (which tend to have higher rates of participation in STIR's program). Please check the *Results Appendix* for details (Section 4)**.** The IV/LATE analysis tries to understand the teacher-level effects of STIR's programming; please refer to the section above for details[57]. The results from this analysis closely mirror the ITT estimates (please

[56] **Note:** In interpreting the results for the teacher level estimates, the reader is cautioned that these results have not been corrected for multiple hypotheses. We would expect 1 out of 20 (at the 5% level of significance) and 1 out of 10 (at the 10% level of significance) p-values would turn up significant driven by statistical noise even though there may not be a 'true' effect.

[57] Note, as part of the IV/LATE analyses, we also included the 'small-school' analysis above as a subgroup.

refer to Section 2 of the *Results Appendix*). The results of the observational style analysis can be found in the *Results Appendix* (Section 3).

6 Limitations

While we have tried to follow the most rigorous approach in trying to assess the impact of STIR's program, there are limitations and challenges to the evaluations:

1. **Capturing teacher motivation:** It is inherently tough to quantitatively capture motivation, which is tough to define/ identify; and (potentially) fluctuates over time. To our knowledge, there is no readily available tool designed to capture motivation. Based on the literature and in conjunction with STIR, we have tried to come up with a tool that adequately captures motivation while overlapping with STIR's understanding of motivation (Details in Section 4.4.2). We acknowledge that there may be imperfect overlap between construct of motivation we develop with the questions and STIR's evolving understanding of motivation and professional mindsets and behaviors. We also acknowledge that the questions used in the teacher motivation index may be measuring numerous underlying concepts. Results from the tool validity tests are as follows (please refer Appendix A11 for details):
    a. The Cronbach's alpha to assess the overall reliability of the questions taken together fall in the "questionable" and "acceptable" range for the baseline and midline tool respectively, but not in the "good" range. Results were similar from the Cronbach's alpha for questions split by category.
    b. Correlation analysis between the positive and negatively framed questions suggested a low degree of reliability.

    There is still scope for improvement and continue to work with STIR to refine our tools.
2. **Observer effects:** Observer effects occur when study participants change their behavior in response to being observed. Observer effects are likely to influence our classroom practice data. We consider our classroom practice data (and teachers' performance) as a reflection of an 'above average' (as opposed to normal) classroom and do not consider it perfectly representative of the 'everyday' scenario. Having said that, however, we do not feel limited from an analysis standpoint, since we do not anticipate differential observer effects between treatment and comparison given that STIR does not directly tell teachers to spend more time teaching or, in Year 1 of the program, directly to engage in the practices we capture as child-friendly.
3. **Teacher-level effects:** Given STIR's theory of change and program design, we consider the school-level estimates as the 'primary' indicators of impact. However, the teacher level effects are also important to STIR from a learning standpoint. While the evaluations, by design, are suited to answer the questions at the school-level, we will continue to work with STIR to come up with creative ways to get at the teacher level estimates while still ensuring technical rigor and theoretical backing of the hypotheses.
4. **Attrition:** As mentioned in Section 4.5.5 and Appendix A15, there is large attrition in both evaluations at the teacher and the student levels. The sensitivity analyses suggest that attrition, despite being high, likely does not effects the results of the evaluation. We will continue to work with the support of STIR field staff to minimize attrition from our samples and continue to assess the effect of attrition on our evaluations as we move towards endline.

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.

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.

Burgess, Simon. 2016. "Human Capital and Education: The State of the Art in the Economics of Education." IZA DP No. 9885. Bonn: IZA.

Costa, Leandro, Nina Cunha, and Barbara Bruns. 2016. "Our RCT in Brazil."

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102 (4): 1241–78.

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, Nauman Ilias, and Michael Kremer. 2010. "Teacher Incentives." *American Economic Journal: Applied Economics* 2 (July): 205–7.

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.

Glewwe, Paul, and Karthik Muralidharan. 2015. "Improving School Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications." Working Paper. RISE Working Paper. RISE-WP-15/001: RISE: Research on Improving Systems of Education.

Groves, Robert M., Floyd J. Fowler, Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2004. *Survey Methodology*. Hoboken: Wiley-Interscience.

Guajardo, Jared. 2011. "Teacher Motivation – Theoretical Framework, Situation Analysis of Save the Children Country Offices, and Recommended Strategies." London: Save the Children. http://www.teachermotivation.org/blog2/resource-teacher-motivation-theoretical-framework/.

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/.

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. doi: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.

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.

Pink, Daniel H. 2010. *Drive: The Surprising Truth About What Motivates Us*. Canongate Books.

Pritchett, Lant. 2013. *The Rebirth of Education: Schooling Ain't Learning*. CGD Books.

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

Robinson, Jenny Perlman. 2011. "A Global Compact on Learning: Taking Action on Education in Developing Countries." Washington, D.C.: Brookings Institute.

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

Sowell, Hope Kathryn. 2013. "Classroom Management Strategies: The Impact on Student Achivement." Dissertation, Lynchburg: Liberty University. http://digitalcommons.liberty.edu/cgi/viewcontent.cgi?article=1824&context=doctoral.

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. 2016. "'Our Approach.'" Accessed May 31. http://stireducation.org/#our-approach.

UN Secretariat. 2015. "Millennium Development Goals Report." New York City: United Nations. http://www.un.org/millenniumgoals/2015_MDG_Report/pdf/MDG%202015%20rev%20%28July%201%29.pdf.

World Bank. 2015. "Conducting Classroom Observations: Analyzing Classrooms Dynamics and Instructional Time." Washington, D.C.: World Bank.