Public Disclosure Authorized

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

Endline Report Appendix

May 2018[1]

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

This document is the main appendix to the endline report for the impact evaluations of the STIR program, conducted by IDinsight. This document has been prepared by IDinsight. We recommend that the reader views this document along with the other two documents prepared: the main report and the results appendix[2].

[2] In case you do not have access to these documents please reach out to Heather at heather.lanthorn@idinsight.org

Abbreviations

APS	Affordable Private Schools
ASER	Annual Status of Education Report, produced in India by the ASER Centre
BEO	Block Education Officer
CARP	Commitment to Analysis and Reporting Plan
DIET	District Institute for Education and Training
EL	Education Leader
HT	Head Teacher (school principal)
ISIT	In-school Innovation Team
IV	Instrumental Variable
LATE	Local Average Treatment Effect
MDE	Minimum Detectable Effect
PM	Program Manager
OLS	Ordinary Least Square
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

Appendix A1: Detailed theory of change and program description

In this section, we narratively walk through STIR's programmatic design and underlying theory of change to familiarize the reader with the contents of the STIR's programming, as evaluated by IDinsight from April 2015 to August 2017. The theory-of-change narrative that follows builds on STIR's documentation as well as IDinsight's understanding of the program, built through discussions and workshops with STIR (STIR Education 2015).

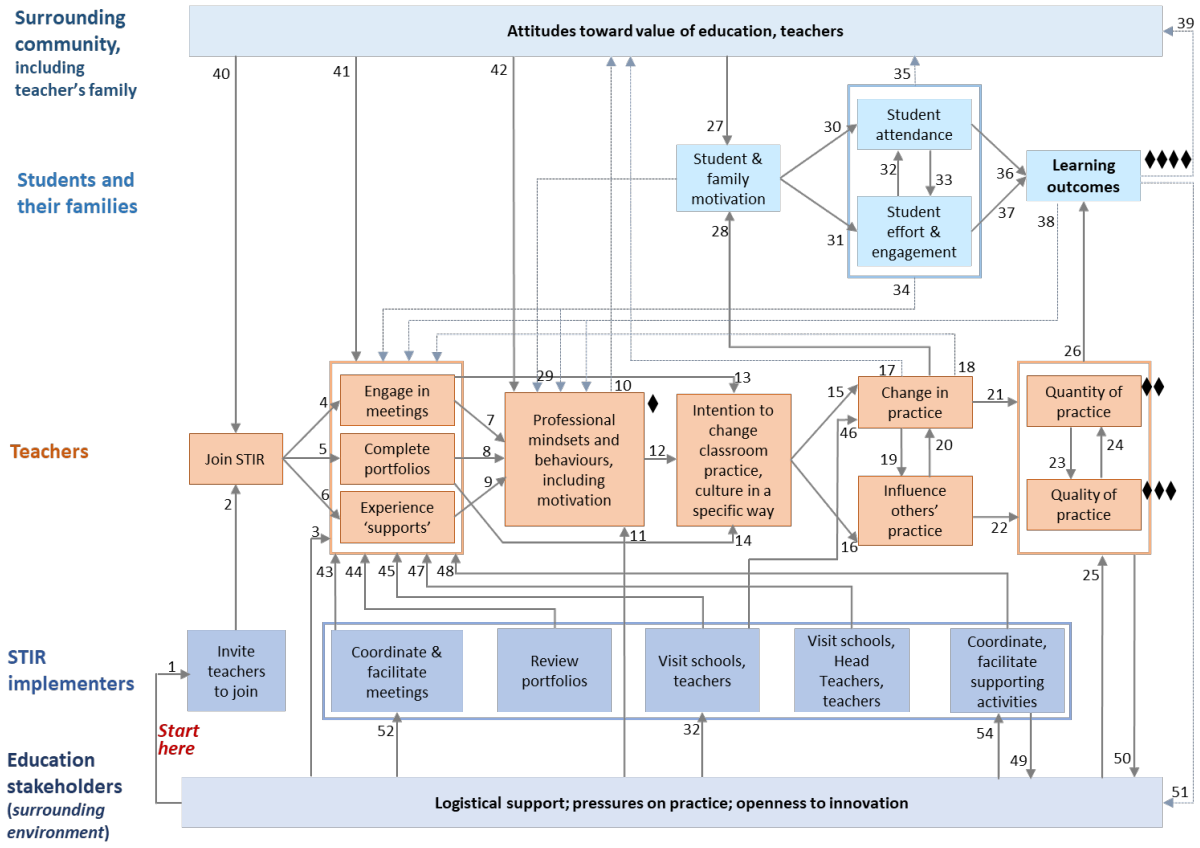
This narrative follows the order of — and is supplemented by — the following detailed diagram; in Table A6, we provide additional detail about the illustrated links and assumptions. Our goal in Table 5 is to allow the reader to focus in on particular links (numbered arrows) of interest, which may pinpoint specific areas for interrogating whether and how the program currently works. The links in the figure provide the connection with Table A6, with one row per link; the diamonds in the figure correspond with key measurement points for the randomized evaluation.

We recognize the diagram presented in **Error! Reference source not found.**may, at first pass, a ppear 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 over two years of programming. The details may also raise useful questions for future programmatic, monitoring, and evaluation work.

To provide some guidance for reading**Error! Reference source not found.**: running down the l eft 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.

Figure 1: 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.

Table A1: Steps, links, and assumptions in the theory of change of STIR's Year 1 programming (corresponds with Figure 1 above)

Link (arrow) number	'From' construct	'To' construct	Linking logic and assumptions
1	Senior education stakeholders provide logistical support, pressures on practice, openness to innovation	STIR invites teachers to apply to join their programming	For STIR to operate in schools and school systems, they require permission and buy-in from key gatekeepers, who must see value in STIR's programming and be able to lend the necessary support.
2	STIR invite teachers to join their programming	Teachers join STIR	To join, teachers need to be aware of STIR and understand how to apply to join; to be interested and able to apply given their understanding of the

			program; and to have their applications selected by STIR.
3	Senior education stakeholders provide logistical support, pressures on practice, openness to innovation	Teachers participation level in STIR (engaging in meetings, completing portfolios, accessing other activities)	Senior education stakeholders implicitly or explicitly display their dis/interest in STIR activities; this can range from actively discouraging teachers on the one end to participating in STIR programming themselves on the other. They can also not/give logistical support, such as providing meeting space.
4	Teacher joins STIR (becomes active)	Teacher engagement level in monthly network meetings	Teachers weigh the personal as well as systemic costs and benefits of travelling to meetings and participating in the discussion and activities.
5	Teacher joins STIR (becomes active)	Teacher completion level reflective portfolios each month	To the extent that teachers value and complete their portfolios — given time, skills, and confidence — they plan for changed practice and reflect on their successes and ways to improve.
6	Teacher joins STIR (becomes active)	Teacher level of accessing other activities offered to them	All other activities take place outside of the network meeting time and outside school time, so teachers weigh the personal as well as systemic costs and benefits of travelling to and engaging in these activities.
7	Active teacher's engagement level in network meetings	Teachers experience changed motivation to teach; mindset on ability to change/innovate	To the extent that meetings are well-facilitated, among an engaged group of peers, and present new information: teachers gain a sense of professional purpose and pride, a mindset that they can improve their skills, and self-efficacy to be an agent of change in their schools and the school system.
8	Active teacher's completion level of reflective portfolios	Teachers experience changed motivation to teach; mindset on ability to change/innovate	Completing their portfolios makes teachers think more deeply about their practices and motivates the teachers to use new innovations than they would have without the portfolios.
9	Active teachers access other activities	Teachers experience changed motivation to teach; mindset on	Additional activities can be motivating to extent that teachers experience and respond to: seeing new

		ability to change/innovate	environments/ practices, being recognized for effort, mastering new ideas, or feeling part of system decision-making.
10	Teacher's motivation to teach; mindset on ability to change/innovate	Surrounding community, including teacher's family attitudes toward value of formal education, quality of teachers	To the extent that teacher's changed motivation is visible to the surrounding community, community perceptions of the quality and value of the school and formal education may change.
11	Senior education stakeholders provide logistical support, pressures on practice, openness to innovation	Motivation to teach; mindset on ability to change/innovate	To the extent that senior education stakeholders provide teachers a sense of agency and creativity in their classrooms (contrasted with hewing to the curriculum and syllabus), and to the extent that teachers derive motivation and satisfaction from this, teacher motivation may be influenced by senior education stakeholders.
12	Motivation to teach; mindset on ability to change/innovate	Intention to change classroom practice, culture in a specific way	Translating general motivation to improve as a teacher into a planned intention to change a specific aspect of classroom practice or environment depends on feeling one has the ideas, confidence, resources, planning skills, and sense of agency to do so.
13	Teacher's engagement in STIR meetings	Intention to change classroom practice, culture in a specific way	To the extent that meetings — through the facilitated lecture and the interaction with other teachers — teachers gain ideas to try, they may intend to try them even without a more general gain in motivation. This could be due to other pressures on classroom practice.
14	Teachers completion level of their reflective portfolios	Intention to change classroom practice, culture in a specific way	To the extent that completing reflective portfolios provide teachers with ideas and plans to make classroom changes, it may generate intention to change, even without a more general gain in motivation.
15	Teacher's intention to change classroom practice, culture in a specific way	Change in teacher's classroom practice (quantity or quality of teaching)	For an intended change to be actualized, teachers need to be physically present in classrooms; to have the required skills, resources, and

			self-efficacy about effecting change; and to have sufficient permission and agency to innovate.
16	Teacher's intention to change classroom practice, culture in a specific way	STIR teachers influencing the practice of other teachers	As teachers plan to make changes in their own classrooms, they may communicate these plans to other teachers in the school, who make consider making similar changes based on their assessment of the idea and their respect for the proposing teacher.
17	Change in teacher's classroom practice (quantity or quality of teaching)	Surrounding community attitudes toward value of formal education, quality of teachers	To the extent that changes in teacher's effort becomes visible to the surrounding community (including students' families and the teacher's own family), the community may update their opinion of schools, education.
18	Change in teacher's classroom practice (quantity or quality of teaching)	Teachers engage in STIR activities (meetings, portfolios, other activities)	As teachers try to make changes in their classroom, and to the extent that they find this enjoyable and that they can make changes, it may encourage increased understanding of and engagement in STIR activities.
19	Change in teacher's classroom practice (quantity or quality of teaching)	STIR teachers influencing the practice of other teachers	To the extent that one teacher's changed practice is visible to or shared with other teachers and is judged to be a worthwhile practice, other teachers may update their own practices. This can be not/facilitated by the extent to which teachers typically share with each other, through informal mechanisms or, if a STIR teacher is willing and able to establish one, through the formal mechanism of ISITs.
20	STIR teachers influence on the practice of other teachers	Change in teacher's classroom practice (quantity or quality of teaching)	As other teachers in a school are encouraged to make changes in their classrooms, and to the extent that these are visible and appealing to individual STIR teachers, they may incorporate these new practices.
21	Change in teacher's	Teacher classroom	Depending on the changes

	classroom practice	practices (quantity and quality of teaching)	teachers make, they may redistribute their time across teaching, classroom management, and off-task activities. They may also make changes that improve the value of time they spend teaching or managing the classroom.
22	STIR teachers influencing the practice of other teachers	Teacher classroom practices (quantity and quality of teaching)	Depending on the changes teachers make, they may redistribute their time across teaching, classroom management, and off-task activities. They may also make changes that improve the value of time they spend teaching or managing the classroom.
23	Quantity of classroom practice (distribution of time on teaching, classroom management, off-task)	Quality of practice (value of time spent relative to outcomes of interest)	If teachers spend more of their time in the classroom productively, they may be able to incorporate more high-quality practices.
24	Quality of practice (value of time spent relative to outcomes of interest)	Quantity of classroom practice (distribution of time on teaching, classroom management, off-task)	As teachers engage in a given quality of teaching and classroom management strategies, they may redistribute the way they spend their classroom time and may be more motivated to do so.
25	Senior education stakeholders provide logistical support, pressures on practice, openness to innovation	Teacher classroom practices (quantity and quality of teaching)	To extent to which teachers redistribute their time and energy to more high-quality practices will be influenced by incentives from education stakeholders to do so, the extent to which senior education stakeholders require strict adherence to the syllabus, as well as openness to innovation among senior stakeholders as well as ideas and modeled behavior from such stakeholders.
26	Teacher's classroom practice (quantity or quality of teaching).	Student learning outcomes	Student learning outcomes may respond to new classroom practices to the extent that the changes are directly relevant to the learning outcomes (*e.g.*, specific skills) and/or create an environment that generally facilitates learning.

27	Surrounding community, including teacher's family attitudes toward value of education, teachers	Student & family motivation to be formally educated	As community values and norms around education change, to the extent that these matter to a given household, it may influence their views and motivation around the value of formal education relative to competing demands.
28	Teacher's classroom practice (quantity or quality of teaching).	Student & family motivation to be formally educated	To the extent that teachers' practices are visible to and matter to a given household, when a teacher updates classroom practice, students' families they may update their valuation of education.
29	Student and family motivation to have children attend and engage at school (relative to competing household demands on time and money)	Teacher's motivation to teach; mindset on ability to change/innovate	To the extent that teachers derive a sense of professional purpose and satisfaction (as well as find their job easier) when students are in attendance and attentive in school and supported at home, students and their families can influence teacher motivation.
30	Student and family motivation to be formally educated (relative to competing demands on household time and money)	Student attendance level	As students and families change their views on the value of education, it may lead to differential effort to get students to attend school.
31	Student and family motivation to be formally educated (relative to competing demands on household time and money)	Student effort & engagement in school	As students and families change their valuation of education, the may change the effort they exert in making sure a student is prepared for school, is able to be attentive during school, and is able to complete school work at home.
32	Student effort & engagement in school	Student attendance	As students change the effort put into their school work, they may feel more/less interested in attending school.
33	Student attendance level	Student effort & engagement in school	As students alter their attendance, it may change their interest, confidence, and ability to engage in the classroom.
34	Student practice (attendance, engagement, effort)	Teacher participation in STIR (engage in meetings, complete reflective portfolios, access other	As students change their practices, teachers may in turn have changed motivation to work to improve as a teacher, especially if they feel they influenced students' practices.

		activities)	This can include changing their participation level with STIR.
35	Student practice (attendance, engagement, effort)	Surrounding community, including teacher's family attitudes toward value of education, teachers	As students change their attendance and engagement at school — to the extent that this is visible to the community — the community may change their perception towards the value of education.
36	Student attendance level	Student learning outcomes	To the extent that a student's presence in school/classroom translates into knowledge and skills relevant for a given outcome test, scores may change.
37	Student effort & engagement	Student learning outcomes	To the extent that students put in changed effort into classes and classwork relevant for a given outcome, scores may change.
38	Student learning outcomes	Teacher participation in STIR (engage in meetings, complete reflective portfolios, access other activities)	As students' learning outcomes change — to the extent that these are visible to a teacher and to the extent that a teacher attributes these changes to her own effort — a teacher may have changed motivation to try to improve as a teacher. This may lead to changed participation with STIR.
39	Student learning outcomes	Surrounding community, including teacher's family attitudes toward value of education, teachers	As students' learning outcomes change — to the extent that these are visible to the community — the community may update its valuation of formal education.
40	Surrounding community, including teacher's family attitudes toward value of education, teachers	Teacher joins STIR	To the extent that a teacher feels her profession and skills are valued by the surrounding community and that this matters to her, she may have differing interest in investing time, energy into improving these skills by joining STIR.
41	Surrounding community, including teacher's family attitudes toward value of education, teachers	Teacher participates in STIR (engage in meetings, complete portfolios, accesses other activities)	To the extent that a teacher feels her profession and skills are valued by the surrounding community and that this matters to her, she may have differing willingness and ability to make monthly decisions to invest time and energy to

			improving these skills.
42	Surrounding community, including teacher's family attitudes toward value of education, teachers	Motivation to teach; mindset on ability to change/innovate	To the extent that a teacher feels her profession and skills are valued by the surrounding community and that this matters to her, this may reinforce a teacher's sense of motivation to become a better teacher.
43	STIR implementers (ELs) coordinate & facilitate meetings	Teachers participate in STIR, including engaging in meetings.	Teachers can participate in STIR meetings to the extent that ELs organize the logistics of the meeting in a manner convenient to teachers and then facilitate the meeting to encourage participation.
44	STIR implementers (ELs) review teachers' reflective portfolios	Teachers participate in STIR, including completing reflective portfolios.	Teachers may change their interest in exerting time, energy, and thought in their reflective portfolios to the extent that ELs clarify the purpose of this activity and provide feedback that teachers deem useful and encouraging.
45	STIR implementers (ELs) visit schools, teachers	Teachers participate in STIR (attending meetings, completing portfolios, accessing other activities).	Teachers may be more excited about, persuaded by, confident in, or simply reminded of STIR activities to the extent ELs visit their schools while they are teaching.
46	STIR implementers (ELs) visit schools, teachers	Change in teacher's classroom practice (quantity or quality of teaching)	The presence of a STIR EL or PM visiting a classroom may directly induce changes in teaching quantity or quality, at least during the visit.
47	STIR implementers (ELs) make coaching calls to teachers.	Teachers participate in STIR (attending meetings, completing portfolios, accessing other activities).	Teachers may be more excited about, persuaded by, confident in, or simply reminded of STIR activities if ELs call them between activities and may benefit from explicit coaching and feel more inclined to participate in STIR, to the extent that ELs are effective during these calls.
48	STIR implementers (ELs) coordinate, facilitate other activities	STIR teachers participate in STIR, including accessing other activities.	Teachers can access STIR activities to the extent that ELs organize the logistics and facilitate the activities to encourage participation, learning.

49	STIR implementers (ELs) coordinate, facilitate other activities	Senior education stakeholders provide logistical support, pressures on practice, openness to innovation	For STIR activities that directly involve senior education stakeholders — such as Head Teachers and Block Education Officers — these experiences can influence their support of STIR and the innovative and professional principles promoted by STIR.
50	Teacher's classroom practice (quantity or quality of teaching)	Senior education stakeholders provide logistical support, pressures on practice, openness to innovation	As teachers change their practice, to the extent that senior education stakeholders are aware of these changes and see them as valuable, they may update their openness to and encouragement of such innovation.
51	Student learning outcomes	Senior education stakeholders provide logistical support, pressures on practice, openness to innovation	To the extent that student learning outcomes are visible and impressive to senior education stakeholders — and to the extent that they attribute changes in these outcomes to STIR-like activities — they may update their views on STIR, the curriculum, and permission to innovate.
52	Senior education stakeholders provide logistical support, pressures on practice, openness to innovation	STIR implementers (ELs) coordinate and facilitate meetings	Senior education stakeholders may need to provide permission and sometimes active support in order for Els to find and reserve a suitable meeting time and place. To extent of support can influence the EL's success in coordinating network meetings.
53	Senior education stakeholders provide logistical support, pressures on practice, openness to innovation	STIR implementers (ELs) visit schools, teachers	Senior education stakeholders often need to provide permission for ELs to visit schools and go into classrooms. The extent of support from these stakeholders can influence whether ELs can do school visits.
54	Senior education stakeholders provide logistical support, pressures on practice, openness to innovation	STIR implementers (ELs) coordinate and facilitate other activities	To the extent that 'other activities' require space and logistical permissions from education stakeholders, stakeholder support could facilitate/block whether these experiences were accessible to STIR teachers.

Selecting schools and teachers for programming and evaluation

After negotiating interest among senior education stakeholders at the block- and/or school-level (in U.P. and Delhi, respectively), STIR offered a 'taster' introductory session to generate interest in STIR and to invite teachers to apply to join, which included expressing interest and providing examples of innovative classroom practice. Inviting teachers to join is the first step shown in the theory-of-change diagram in Figure 2 in the main report ("invite teachers to join," on the lower left side of the diagram, marked with "start here").

In schools selected (randomly assigned) to receive STIR programming, all teachers had the option to voluntarily apply to participate with STIR. From the perspective of the evaluation, treatment assignment (the offer for some teachers in a given school to join STIR) took place at the school-level. From a programmatic perspective, STIR planned to only select some of the interested teachers in each treated school to participate. This selection, based on application materials, submitted to STIR, depended on the teacher's fit with STIR's programming and the quality ideas of innovative practice ideas submitted with the application. In addition, to help keep networks at what was expected to be a manageable size for ELs, STIR introduced an element of 'rationing' to place an upper limit on the number of teachers selected per school.

In practice, this did not happen, for two key reasons. First, in some schools, there was a lot of enthusiasm for STIR, and it proved hard to turn down interested teachers. Second, in other schools, sufficient take-up of and retaining in the program (and in STIR schools more generally) proved lower than expected. Ultimately, rationing was not a substantial barrier between a given teacher being interested in joining STIR and her being invited to do so.

Teachers could face numerous personal and systemic incentives and barriers for joining STIR once invited. Some of these are detailed in Link 2 in Appendix Table A6. Each teacher weighs their understanding of the benefits and costs of joining STIR to make their decision to apply to join. Personal benefits may include enthusiasm about improving teaching practice, interest in achieving the Roehampton Changemaker Certificate, or interest in meeting teachers from other schools and being part of a community of practice. Personal barriers can include the time and transport money lost to participating as well as shyness and fear of failure or critique. Systemic benefits could accrue to teachers if the relevant leadership (HTs, BEOs) were enthusiastic about STIR. Other stakeholders may view STIR as a distraction from achieving the planned syllabus and curriculum and implicitly or explicitly discourage participation; particularly in private schools in Delhi, some school leaders worried about cross-school teacher poaching at network meetings. The view from education stakeholders may, thus, play a role in whether an invited teacher opts to join STIR (as indicated in Link 1).

For all facets of participating actively in STIR (from attending a first network meeting through completing the programming), an important role is played by senior education stakeholders including HTs and BEOs (Link 3). Through their overt support of STIR's programming (such as talking positively about STIR, showcasing teachers who participate with STIR, and/or attending network meetings) or more subtle support, these stakeholders alter the feasibility and desirability of active participation in the STIR program.

Becoming a participating teacher

Actively participating

The core of being an actively participating teacher in STIR, after joining, is to 'engage in (network) meetings,' 'complete portfolios,' and 'access other activities' (shown as boxes in Figure 5 in the main report). Whether and to what extent a teacher who joins STIR (expresses interest, applies and is accepted, and attends at least one meeting) becomes an active participant depends on many factors (Links 4, 5, and 6).

Engaging in meetings requires attendance and contribution for the duration of the meeting (between 45 minutes and two hours). After joining STIR, a teacher must make a recurring monthly decision about whether to attend and participate in these meetings (Link 4). Logistically, barriers to attending meetings can include the time, transport costs (borne by teachers), organizational burden of arranging travel to and from meetings, and safety concerns about transport. Meeting time can compete with weekend, evening, and/or classroom time, depending on the context. Concerns about travel safety, particularly for younger females, may be compounded if colleagues from her school are not also attending or otherwise cannot travel together. Teacher sex and age, then, may influence how regularly and comfortably teachers attend network meetings. Teacher interest in STIR and its benefits (linked to initial motivation to improve as a teacher) will, in part, determine how much effort a teacher exerts in trying to overcome individual and systemic barriers to engagement.

Still in Link 4, once at a meeting, teachers also likely need to contribute to maximize their benefit. Not all attendees may feel confident speaking up at meetings, especially when they must share their weak points or challenges faced. Confidence may be built (or lost) depending on whether teachers feel their contributions are valued. Teacher sex, age, qualification (training/degree), and years of experience may all play a role in whether a teacher decides to contribute during network meetings.

Actively participating also requires teachers to reflect on their teaching practice, in part by completing their portfolios. However, a teacher may face barriers to reflecting on the questions in the portfolios and filling in the answers (Link 5). First, teachers may not value or derive personal benefit from completing the portfolios. They may instead view it as another form of writing and paperwork (of which plenty already exists in school systems). Teachers also may see the portfolios as a requirement for achieving the Roehampton Changemaker Certificate rather than an instrument with intrinsic value for improving as a teacher. Second, even if a teacher does find value in the portfolio, they may not be able to adequately plan or allocate time to this activity. Third, even when they find both value in and sufficient time to engage with the portfolio, a teacher may not feel she has the skills or experience to engage in self-reflective practice, which is new to many of the teachers participating in STIR. During our theory-of-

change building activities with STIR frontline staff, we heard about requests from teachers to have ELs fill the portfolios for them.

Finally, there are a suite of 'other activities' that, for some teachers, will be part of actively participating in STIR. These are linked with the experimental design and research questions, described in more detail under "Program variations" in the main report and in Appendix A2. These activities take place outside of the network meetings and, therefore, teachers may face barriers and enablers similar to those detailed for network meeting attendance (Link 6).

From participating in STIR to making changes in classroom practice and culture

To move from the ideas and encouragement received in the monthly network meetings to making changes in the classroom, a teacher may move through several intermediate phases: generally being motivated to change; having a sense of potential and self-efficacy to activate a specific change; intending to change; and ultimately making changes in classroom practice and/or culture.

Getting motivated

Engaging in network meetings can be a consequence of but also a source of motivation to teach (better) (Link 7). Different activities during meetings are designed to remind teachers of the importance of student learning and the critical role of teachers in facilitating this learning. This can help to build a teacher's sense of intrinsic motivation and professional purpose; this leads, ideally, to a sense of commitment to one's profession (Pink 2010).[3]

Meeting content also encourages teachers to adopt a growth mindset, thereby seeing themselves as capable of becoming better teachers, regardless of their current skill and practice (Dweck 2010).[4] It also helps teachers view their students as capable of learning. This sense of potential can also be motivating, and may be reinforced by completing the reflective portfolios and receiving feedback on them (Link 8).

Motivation — especially extrinsic motivation — may be further enhanced when teachers access the 'other activities' offered through variations to STIR programming (Link 9). The specific ways through which motivation may be enhanced are detailed under "Intrinsic and extrinsic motivators: core and core-plus models," in Appendix A2.

Support from students' families likely influences teachers' motivation; this support may be changed during the "influence cycle," when teachers must specifically reach out to parents of five of their underperforming students (Link 29). Teachers' level of motivation to improve their teaching practice to help their students learn, to the extent that this is visible and interesting to the families and community surrounding the school, may influence how the community views

[3] Note that a sense of professional purpose as an educator need not have drawn current teachers to teaching in the first place. This may be particularly true in the government school system, where teachers may view themselves as civil servants first and teachers only second.

[4] A growth mindset, contrasted with a fixed mindset, is one oriented toward constant improvement and a sense that such improvement is possible (Dweck 2010). The idea extends further to teachers seeing their students as capable of improving rather than having a fixed level of intelligence.

the value of education and particular types of teachers (Link 10). One of the 'other activities' focused on "local recognition" (see Appendix A2 for more details) involved explicit efforts to make the surrounding community aware of teachers' efforts through posters and events.

Education stakeholders such as HTs and relevant government officials can also influence teachers' motivation in important ways, some of which are under STIR's influence and some of which are not — but we feel these are still worth mentioning to help calibrate expectations around what STIR's programming can/not feasibly achieve (Link 11). These factors include the types of contracts and job security teachers receive, the volume of additional responsibilities given to a teacher by the school administration, the extent to which teachers feel recognized by their superiors and key stakeholders, the extent to which teachers feel they can be creative in the workplace, and the extent to which they learn new skills on the job. Some of STIR's regular programming as well as 'other activities' tested during Year 1 of the program touch on these issues (as detailed in Appendix A2). For example, all of STIR's work to help teachers develop a growth mindset and a sense of self-efficacy to make specific changes in the classroom are linked to enlarging the space in which teachers can be creative. In the "government and policy exposure" activity bundle provided U.P. (see Appendix A2), teachers are supposed to have an opportunity to meet with and be recognized by government officials, while the "career and personal development" bundle provided teachers an opportunity to work on their English and other professional skills.

Toward intention

To alter classroom practice, as per the working theory of change, a teacher must translate a general sense of motivation to teach well in an interest and intention to make a specific change in classroom practice or culture (Link 12). It is possible that elevated motivation to help students learn and succeed could also be channeled into other activities, such as providing out-of-classroom remedial activities and tuitions/tutoring, which may/not link with STIR's intended outcomes. STIR focuses on in-school changes. Motivation to help students learn could also founder on poor student attendance or attention and/or on a lack of teaching materials — and therefore fail to be translated into an intention to change classroom practice (Link 34). STIR partially helps to overcome the latter concern (of limited teaching and learning materials) by encouraging teachers to make use of local materials to serve as educational inputs, rather than relying on what the school can and does provide.

Teachers may also move toward an intention to change classroom practice and culture in specific ways without experiencing changed motivation. Perhaps their motivation was already high, which lead to them participating actively with STIR in the first place. Or, through other pressures from colleagues or other education stakeholders, they may intend to make the changes learned about in meetings and through portfolios without changes in motivation (Links 13 and 14).

Once intent on making specific alterations to the classroom environment or to teaching practice, a teacher must make daily decisions to actualize those changes. A teacher needs to recall STIR-introduced ideas and approaches from meetings and portfolios, feel that they have the ability and permission to make these changes, and then work to introduce them in the classroom. The reflective portfolio, which requires teachers to delineate plans for effecting specific classroom changes can sharpen the intent to make a change.

Toward changed practice in classrooms, schools, and communities

There may be many barriers and possible facilitators between intending to make specific changes in classroom practice and doing so (Link 15).

First and foremost, a teacher must be present at school and in the classroom to change it. This is not always feasible. While we expect more motivated teachers to attend more frequently than less motivated counterparts, personal and family conflicts can get in the way of teachers attending school. In addition, especially in government schools, teachers have a variety of out-of-school tasks related to being civil servants.[5] Such obligations put systemic pressures on the teacher to omit time in the classroom for these duties.

Second, teachers need ideas to help animate their intention and they need time and the process skills to plan how to put ideas into action. Network meetings and reflective portfolios *may* provide these necessary inputs. Teachers further need to feel a sense of self-efficacy to operationalize their ideas, including to make the necessary changes in their classrooms and, as needed, to negotiate and justify those changes with their HTs and other stakeholders (Bandura 1977). Note that this implies two distinct strands of self-efficacy: one regarding teaching skill to make specific changes in the classroom and one with regard to explaining and negotiating these changes with stakeholders. A lack of agency or autonomy in the classroom can hinder a teacher's motivation or ability to translate intention into action.[6] So, too, will a sense that new ideas need to be executed perfectly the first time, rather than a recognition that practices be attempted, adapted, and tried again as needed (as proposed in the Learning Improvement Cycle, which features more prominently in STIR's updated programming).

Teachers may more directly decide to change classroom practice when STIR program staff (such as ELs) visit their schools and classrooms (Link 46).

As teachers begin to try new ideas and practices, it may stimulate two feedback loops. For one, if these changes are visible to students' parents and the broader surrounding community — and if they fit with the community's idea of good changing practice — they may influence larger attitudes about the value of education and schooling (Link 17). For two, as teachers see that change in their classrooms is possible — perhaps regardless of what student learning outcomes result — it may help spark additional interest in participating with STIR and in being motivated to exert effort to improve classroom practice (Link 18). Teachers may also gain additional motivation if their changes are recognized, whether by their own families, their colleagues and peers, their superiors, and/or their students and their families (as implied in Links 29 and 42).

[5] This may include attending workshops or helping with government duties, such as helping with local elections.

[6] Data from our April 2016 process evaluation suggest that teachers, especially in private schools, often need permission to make changes in their classrooms. More than 80% of teachers in Delhi private schools and roughly 35% of teachers in U.P. government schools answered 'yes' when asked "If you want to change practices in your classroom, do you need to take anyone's permission or opinion?"

Teachers can also translate their intention to change classroom and school practices into influencing the practice of other teachers in the school, whether through informal means over chai or through the formal mechanism of In-School Innovation Teams (ISITs). The extent to which this happens depends, at least in part, on the opportunities for sharing innovative ideas within schools and being part of school decision-making, as well as whether a teacher feels sufficiently confident to intentionally influence the practice of others, which may vary by age, qualification, experience, and sex (Link 16). STIR teachers can also influence the practice of others as they change their own practice and lead by example, tempered by the extent to which teachers get to see each other in action (Link 19). Similarly, STIR teachers may change their own practices as they gain ideas and confidence from seeing the innovations of others (Link 20).

Quantity and quality of classroom practice

One of the major assumptions underpinning STIR's programming is that teachers will make changes in their classrooms and schools that lead to *beneficial* changes in the quantity and/or quality of effective time for instruction or classroom management. Of course, not all changes in classroom practice among STIR teachers (Link 21) or others (Link 22) will necessarily lead to a useful reallocation of classroom time nor to useful changes in the quality of classroom practice and culture.

For quantity of practice, for example, teachers may change their practices within a given amount of teaching minutes but not actually alter the proportion of their classroom time devoted to instruction. STIR's programming does not provide direct guidance on how teachers should allocate their time between teaching and classroom management nor does it explicitly discourage off-task time. While we might *prima facie* expect more motivated teachers to devote more time to teaching than to classroom management or being off-task, this may not always be the case.

For quality of practice, some aspects of what teachers can do are restricted or misguided by their own teaching capacity, content mastery, pedagogical strategies and beliefs, and teaching skill. Some practice may be changed but not become objectively better or lead to the measured learning outcomes. Moreover, some teacher efforts at change may simply not be oriented toward what STIR considers to be good classroom practice or culture, especially since part of the goal of micro-innovating in the first year is simply to prove to teachers that they can effect change of any kind.[7]

Even when planned changes are aligned with STIR's view of high-quality classroom practice, teachers may simply not be successful at their (early) attempts to bring about change. Teachers require the growth mindset, process skills, problem-solving skills, and resilience to learn from challenges and barriers and then to adapt and try again — as most new practices will not work perfectly on the first attempt. Portfolios may help teachers to reflect on challenges and devise new strategies (as per the "learning improvement cycle") but only to the extent that teachers have

[7] As an extreme but illuminating example from a September 2015 process evaluation, one teacher's proudest micro-innovation was to have his students wear blindfolds so that they would concentrate more attentively to his lectures. It is not clear if this change would count as a positive change in classroom practice or culture in general and certainly would not show up positively in the indicators of classroom practice we collect.

the time, interest, and skill to engage thoughtfully with their workbooks. Collaboration (and the collaborative skills to work) with other teachers (in network meetings or within schools, whether in ISITs or through other channels) may offer means of reflecting, problem-solving, and building up the interest and gumption to try again.

The distribution of time use in the classroom may reinforce the existing quality of classroom practice (Link 23) and vice versa (Link 24). For example, spending more time on instruction, in general, may allow teachers to feel they can allow time for students to ask questions, or to use group work, or to share a joke — all of which may improve classroom culture. Helping teachers to harness the synergies between the quantity and quality of classroom practice may be an important part of STIR's programming moving forward.

The extent to which teachers can effect changes in their classrooms will also be influenced by education stakeholders, such as their openness to change and their degree of focus on the curriculum and syllabus (Link 25). In turn, if teachers do make changes in the quantity and quality of their classroom practice, to the extent that these changes are made visible to education stakeholders, it may alter stakeholders' views about the sanctity of the syllabus and the value of innovation (Link 50).

From changes in classroom practice and culture to changes in learning outcomes

Student learning is the ultimate school-level goal for STIR's programming. It is also a key measurement point for the randomized evaluation). A fundamental assumption in STIR's theory of change (Link 26) is that the alterations teachers make to the quantity and quality of their practice will improve student learning outcomes (for the purposes of this evaluation, specifically in Hindi and math).

The role of students, families, and the surrounding community

No amount of classroom and school innovation will bring about changes in learning outcomes if students are not in the classroom, able and interested to pay attention. This highlights the important role of students, their families/caregivers, and their surrounding communities in achieving the goal of student learning — these are included in the second row from the top of Figure 2 in the main report. Being surrounded by people that value (formal) education and the local school as the provider can reinforce habits of attending school; being surrounded by people that value agricultural or other work at the expense of school attendance or predict low returns to education will have the opposite effect (Link 27).

As teachers change their classroom practice (including, in the case of STIR, reaching out to the families of five under-performing children), this may change the way students and families feel about the quality of local schooling, the accessibility of schooling, and their motivation to take (formal) education seriously (Link 28).

Whether increased student and family interest in schooling translates into improved school attendance will depend on many factors, such as family pressure for a child to earn income or formal and informal costs associated with going to school (maintaining a uniform and so on) (Link 30). Whether motivation translates into increased student effort and engagement in and

outside the classroom will depend on whether the student can pay attention in school (receives breakfast, for example, and other health and nutrition inputs), has time to do homework and prepare for classes, and feels comfortable and accepted in the classroom (Link 31).

Student effort, engagement, attendance may be mutually reinforcing. A student able exert more effort on studies may be more interested in attending school (Link 32) and vice versa (Link 33), though the synergies may not be automatic. As students increasingly attend class and pay attention — and especially if teachers attribute these changes to their own efforts — teachers may gain more enthusiasm to participate actively with STIR and to feel more motivated about teaching in general (Link 34). However, it is important to recognize that student attendance and attention may be largely out of a teacher's control, with implications for how much their changed classroom practice can translate into average learning outcomes across all students. As students change their attendance and effort in school, to the extent that this is visible to the community, may lead to updated attitudes about education (Link 35). 'Other activities' such as "local recognition" may facilitate these links by making teacher efforts more visible to the community.

Both student attendance and student effort have the potential to influence student learning outcomes (Links 36 and 37). As student effort and/or learning outcomes change, so too might teachers' motivation to participate in STIR and to try to innovate in their classrooms and schools (Link 38). Changes in learning outcomes among students (especially if these are communicated back to the larger community) might also shape the attitude of the community about the value of education and teachers (Link 39).

The surrounding community can also play an influencing role for STIR in other ways. If the community, including students' and teachers' families, value teachers and the effort required to teach well, this can reinforce the interest of teachers in joining STIR (Link 40) and their ability to navigate the logistics and time requirements of participating actively in STIR (Link 41). To the extent that teachers feel valued and respected by the local community, they may also feel more motivated to become better teachers; alternatively, in a community where education and teachers are not valued, teachers may be unmotivated to invest time and energy in improving their teaching practice (Link 42).

Making STIR's programming happen

The role of STIR implementers

Front-line STIR implementers (namely the Education Leaders (ELs)) are central to interacting with teachers and ensuring that STIR activities happen and have the maximum potential to excite teachers and to impart new skills. ELs balance teaching new skills as per STIR's meeting curriculum while facilitating collaboration among network teachers. A good, respected EL can help motivate teachers to join and stay active in STIR through a variety of channels, tempered by the EL's skills and effort (Links 43 – 45; 47, 48).[8]

ELs undertake an array of tasks to help STIR's programming run as effectively as possible. This includes finding space for each network meeting, contacting and organizing schedules among

[8] Anecdotal evidence from the Delhi-based STIR team suggested that handsome, guitar-playing ELs maintained higher attendance rates at their network meetings.

network teachers, and then imparting material and facilitating conversation and participation during the meetings (Link 43). ELs further review the teachers' reflective portfolios and, when required, assist in completing the portfolios (Link 44). ELs (and sometimes Program Managers) also may visit schools and classrooms, which can include maintaining good relations with school gatekeepers as well as observing a teacher's classroom practice (Link 45). (Recall that this can also directly alter practice, as in Link 46). ELs also take on the role of coaches—fielding questions and concerns from participating teachers (Link 47). Finally, ELs need to organize the additional activities discussed under "Program variations," in Appendix A2 (Link 48). Together, these activities help to create STIR programming that is compelling to teachers such that teachers can and want to participate actively.

The role of senior education gatekeepers and stakeholders

Senior education gatekeepers and stakeholders play an important role in enabling ELs and PMs to effectively carry out STIR activities. For example, stakeholder assistance and buy-in is essential to securing time and a location for network meetings, which often are held in a different location each month (Link 52). Similarly, gatekeepers such as Head Teachers need to allow ELs into schools for visits; therefore, their view of STIR can influence their decision to help or hinder STIR's work in the school (Link 53 as well as Links 3, 11, 25). Finally, stakeholders and gatekeepers may have to play a role in the additional activities, such as allowing teachers to visit other local schools to gain exposure to new practices (Link 54). Stakeholders focused strictly on the syllabus and curriculum may not allow this sort of flexibility. The broader school system may, in addition, shape whether teachers' and schools' efforts at improving practice lead to changes that last sufficiently long to affect learning outcomes.

Whether stakeholders and gatekeepers view STIR positively will depend in part on the relationship they have with ELs and PMs. It may also be shaped by STIR-led opportunities for interaction between teachers and Block Education Officers (Link 54). Finally, if stakeholders are aware of changes in student learning outcomes that they attribute to STIR, it may change their valuation of STIR's programming (Link 51).

Appendix A2: Elaboration of core-plus programmatic flavors

For the purposes of this evaluation, there are five different flavors of core-plus activities; that were explored for year 1 of the program. Note that the explicit statement of aims for each of these flavors happened after the different activity bundles had been launched. Many of the bundles grew out of practices already being used by some ELs and PMs; others came from focus groups led by STIR with teachers. In Table A1, we summarize which core-plus flavors took place in which evaluation contexts, as some were Delhi- or U.P.-specific.

Table A1: Mapping flavors of core-plus extrinsic motivation packages to study context

Core-plus extrinsic motivation package		Delhi private schools	U.P. government schools
C+a:	local recognition	X	X
C+b:	government and policy exposure		X
C+c:	Head Teacher recognition	X	
C+d:	teacher exposure	X	X
C+e:	career and personal development	X	

C+a: Local recognition (both contexts)

Aim

To provide active STIR teachers recognition for their teaching efforts and best practices from their (1) school (students and colleagues), (2) community and (3) family as a source of extrinsic motivation.

Logic and assumptions

The underlying logic is that recognition can improve support, valuation and visibility of teaching both in a teachers' home and in the school. This can feed into motivation, depending on the teachers' own receptivity to recognition. In addition, specific activities such as a poster could provide a cue-to-action for teachers to, on a daily basis, move from intending to change classroom practice to actually making the effort to do so. Further, by highlighting and valuing the activities of active STIR teachers, other teachers in the school may notice these new ideas and be motivated to make changes in their own classrooms.

Included activities

Hanging a poster in the school (and sometimes in the wider village in U.P.) that highlights how a teacher is impacting students; receiving postcards from other network teachers; stickers on doors of STIR teachers in a school; sending a letter of appreciation to the teacher's family that

highlights the importance of family support to teachers; and organizing a 'family day' celebration.

C+b: Government and policy engagement (U.P. only)

Aim

To provide teachers a chance to interact with BEOs, Basic Shiksha Adhikaris (BSAs), DIET principals, and district-level officials, with the understanding that teachers will receive recognition (and therefore additional motivation) for their efforts from a person of authority. Motivation may come both from direct recognition but also indirectly, as teachers identify themselves as part of larger system and see themselves as having a role and voice in discussions within that system.

Logic and assumptions

The government system does not have a defined recognition structure for their teachers and teachers in turn often feel ignored. The engagements provided through this treatment provide opportunities for selected teachers to interact with their officials beyond monitoring purposes and make them feel part of local policy making in the sphere of education. Note that STIR did not intend to use this engagement as an explicit platform to improve local official support for their programming. Nevertheless, these engagements between local officials and active teachers may also facilitate a wider change in system pressures on practice and encourage enthusiasm for innovation that will influence all teachers.

Included activity

Arranged meetings between teachers (6 to 7 teachers per meeting, on a rotating basis), ELs and local government officials in a 'block level policy forum.' Teachers are offered an opportunity to present a story from their classroom or school, including successes, challenges and learnings. BEOs and other officials have time to comment and make suggestions. The meetings are intended to close with the development of a combined plan of action for represented schools.

C+c: Head Teacher (principal) recognition and development (Delhi only)

Aim

To provide Head Teachers with an opportunity to develop themselves as principals, with an indirect intent of exciting them about STIR's programming.

Logic and assumptions

The underlying logic is that Head Teachers who have a chance to develop themselves and feel excited by being part of a movement will in turn be more supportive toward teachers, including those involved with STIR. This support can improve the motivation of teachers. It can also facilitate teachers' ability to participate in STIR programming and, more broadly, to make changes in their classrooms and schools (regardless of whether they participate with STIR).

Included activities

Delivering a School Development Toolkit to assist Head Teachers in making assessments of their schools; facilitating conversations between HTs just beginning with STIR and those that have been involved in the program longer (and are therefore outside the randomized evaluation samples); and empowering Head Teachers to support teachers as they take on challenges. Principals also receive an 'Empowered Head Teacher' certificate and mentorship from a STIR Head Teacher further along in the Changemaker Journey.

C+d: Teacher exposure (both contexts)

Aim

To help teachers learn from experiences beyond their own schools which would help improve their motivation.

Logic and assumptions

Seeing innovations in action can provide teachers with new ideas to implement and possibly improved confidence to try these ideas by seeing them modelled by other teachers. In addition, teachers will be reminded that other teachers, beyond their own schools, are working on improving classroom culture and practice, thus improving their sense of being part of a movement. And, finally, a chance to go see another school is fun and rewarding in and of itself.

Included activity

Providing an opportunity to see other school environments. It includes: a trip with other teachers to another school and then structured reflection on what they saw; and engagement with an expert on growth mindset; and a 'teachers' report.'

C+e: Career and personal development (Delhi only)

Aim

To help teachers gain specific skills outside network meetings, which can help them in their careers as teachers and beyond, recognizing that skill mastery can be motivating.

Logic and assumptions

Facilitating the development of new, valued skills. Teachers in private schools highly value career progression and most of them look forward to moving to a larger private school. English speaking skills, classroom management and lesson planning are some of the key skills they want to pick up. Providing training in the aforementioned skills would not only be helpful for teachers in the immediate term but would contribute to their long-term career growth. Moreover, it reinforces the idea that being involved in STIR can provide access to such opportunities.

Included activities

Participating in events outside of network meetings to allow time for: (1) learning about best classroom practices through videos in a peer-to-peer setting; (2) talking with a 'growth mindset' expert; and training in spoken English.

Appendix A3: Micro-innovation details from April 2016 process evaluation

In a process evaluation conducted in April 2016, we asked teachers participating in STIR to describe the micro-innovations they had tried and aim they sought by doing them. We report the results below to give *suggestive* (rather than fully representative) ideas to the reader about the range of micro-innovations being deployed. In Table A2 and Table A3, we provided the coded responses of the aim of the micro-innovations described by teachers (in Delhi and U.P, respectively).

Table A2: Aim of micro-innovations described by teachers in Delhi private schools

Aim of micro-innovations in Delhi private schools	Count	Percentage
Attendance	8	11%
Unspecified learning	8	11%
Homework	5	7%
Not done	4	5%
Reading and writing	4	5%
Discipline (such as wearing proper uniform, coming regularly to class, being well behaved in class etc.)	3	4%
Student participation and interest	3	4%
Off topic answer	2	3%
Guardian engagement	1	1%
Blank/ Don't know	23	30%
Unclear/ Unsure	15	20%
Total	**76**	**100%**

Table A3: Aim of micro-innovations described by teachers in U.P. government schools

Aim of micro-innovations in U.P. government schools	Count	Percentage
Not done	15	14%
Attendance	13	12%
Numeracy	8	7%
Reading and writing	5	5%
Guardian engagement	4	4%
Student participation and interest	3	3%
Discipline (such as wearing proper uniform, coming regularly to class, being well behaved in class etc.)	2	2%
Encouragement/ motivation	2	2%
Off topic answer	2	2%
Environment	1	1%
Focus on girls	1	1%
Unclear/ Unsure	18	17%
Unspecified learning	18	17%

Blank/ Don't know	17	16%
Total	**109**	**100%**

In Table A4 and Table A5, we show the descriptions of micro-innovation activities from teachers, which, again, should only be taken as suggestive of the range of activities tried by all teachers actively participating in STIR.

Table A4: Description of micro-innovation activities by teachers in Delhi private schools

Aid	Count	Percentage
Chart display/ performance tracker	10	25%
Stars/ stickers	9	23%
Unsure	5	13%
Group Activity	3	8%
Appreciation (claps, reward etc.)	2	5%
Engaging students (questions, quiz etc.)	2	5%
TLM (*e.g.*, Flash cards)	2	5%
Word jumble	2	5%
Birthday announcement	1	3%
Meeting Parents	1	3%
One on one teaching/ remedial groups	1	3%
Tests or grading	1	3%
Writing	1	3%

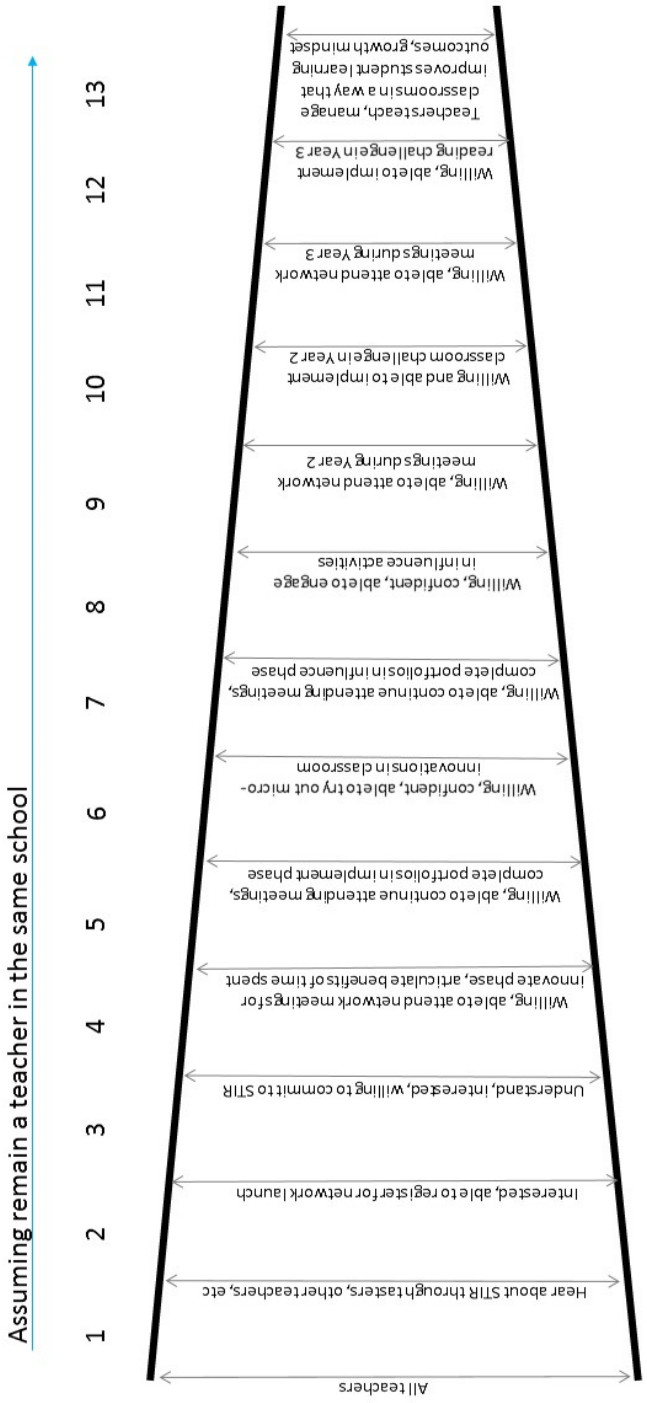
Table A5: Description of micro-innovation activities by teachers in U.P. government schools

Aid	Count	Percentage
Group Activity	9	11%
Appreciation (claps, reward etc.)	7	8%
One on one teaching/ remedial groups	6	7%
TLM (*e.g.*,: Flash cards)	6	7%
Engaging students (questions, quiz etc.)	5	6%
Encouragement/ motivation	4	5%
Meeting Parents	4	5%
Play way method	4	5%
Localizing	3	4%
Story telling	2	2%
Word jumble	2	2%
Writing	2	2%
Bubble gum	1	1%
Chart display/ performance tracker	1	1%

Garbage collection	1	1%
Magic box	1	1%
Tests or grading	1	1%
Unsure	25	30%

Appendix A4: Funnel of Attrition

Figure Aa: Funnel of attrition used with STIR in workshops



Appendix A5: Details of randomly assigning STIR's programming

Delhi

Defining the sample

In Delhi, the STIR team undertook a large search exercise for private schools in East Delhi with a maximum monthly fee of US$ 17.00.[9] The team initially reached out to around 500 schools. From these, STIR identified 200 private interested in working with STIR; STIR then formally invited them to participate in their program, starting with the 'taster' session. 180 of these schools agreed to participate. These 180 schools provided the full sample, for which we randomized the assignment of STIR's interventions at the school level.

Random assignment of treatment in Delhi

The 180 schools were then divided into 7 (roughly) equally sized strata based on geography, such that schools physically closer to one another were more likely to be in the same stratum. Strata had between 22 and 25 schools. Each stratum was assigned to a single STIR Education Leader, the key front-line implementer of STIR's program in Delhi. This is visualized in Figure Ab, in which the circles represent schools filled with teachers.

Figure Ab: Sample and strata in Delhi private schools

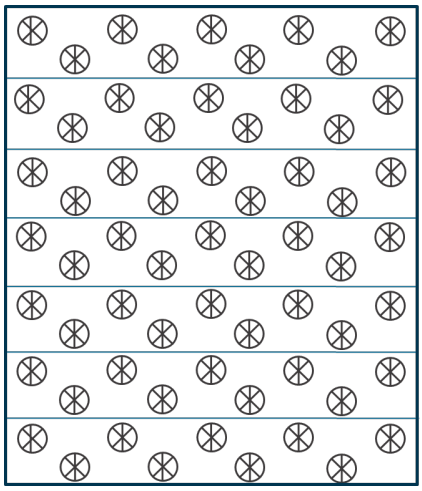
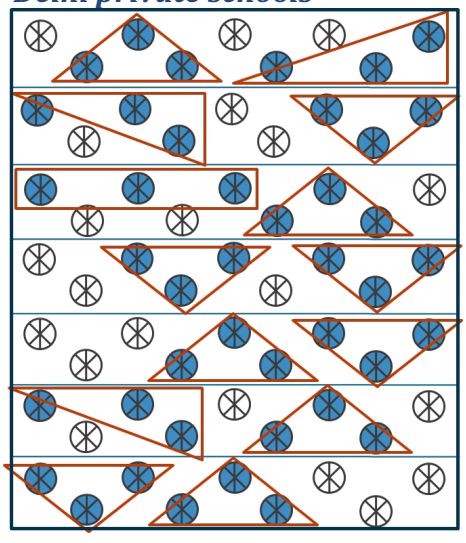


Figure Ac: First randomization and clustering of schools in Delhi private schools

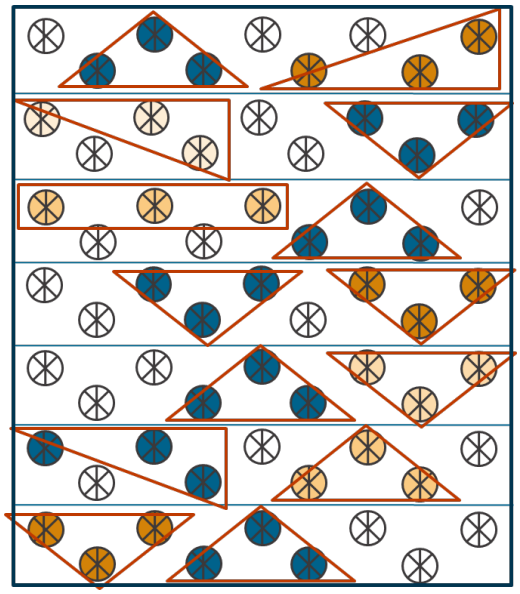


Within each stratum, one-third of schools were randomly assigned to comparison and two-thirds of schools to treatment (filled-in blue circles represent treated schools in Figure Ac). Then, within each stratum, the schools assigned to treatment were divided into 4 clusters based on geography; these clusters are program units for STIR.

Finally, within each stratum, two treatment clusters were randomly assigned to STIR core treatment (shown as blue circles in Figure Ad) while the remaining two clusters to the four STIR core-plus flavors (shown in shades of orange in Figure Ad) using sampling without replacement approach (*i.e.,* within each stratum, there are two flavors of the core-plus treatment).

[9] There is no universal definition of what counts as an 'affordable private school.' "APS are loosely defined as privately owned schools serving low income communities" (Tooley, Dixon, and Gomathi 2007) (Tooley & Dixon, 2007).

Figure Ad: Final randomization in Delhi private schools



Uttar Pradesh (U.P.)

Defining the sample

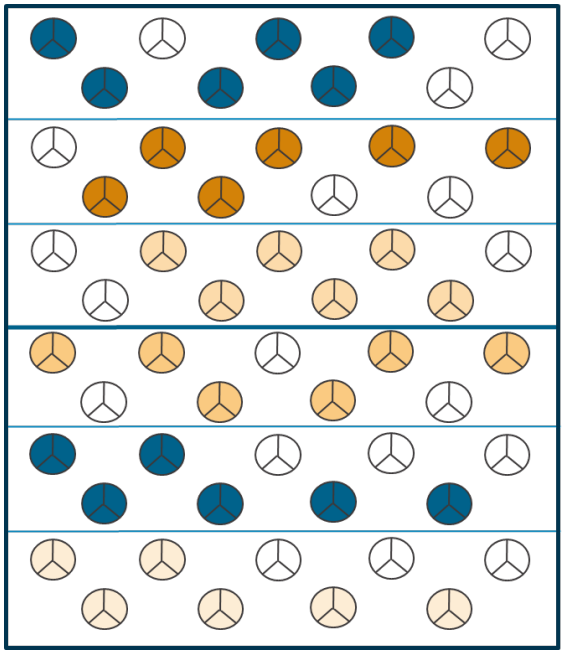
In Uttar Pradesh, districts are organized into government administrative units called "clusters" (hereafter government clusters); these form the strata within which we randomize. Within our two districts of interest, (Rae Bareli and Varanasi), we dropped government clusters with less than 15 schools from consideration. From among the remaining government clusters, we randomly selected 16 clusters. This forms our sample for assigning treatment.

Random assignment of treatment in U.P.

Within each government cluster, we randomly assigned one-third of schools to comparison and two-thirds to treatment.[10] In a second randomization, all treatment schools were randomized to either the core programming or one of the three flavors of core-plus programming available in U.P. Thus, all treatment schools in a government cluster received the same treatment. This is shown in Figure Ae, in which circles of schools (populated with wedges/teachers); core programming is shown in blue and core-plus programming in shades of orange.

Figure Ae: Details of random assignment in U.P.

[10] For a few schools, we didn't randomize at the individual school level. In some cases, two schools shared the same building or grounds (mostly the case where PS and U.P.S schools of the same village are very close to one another). Thus, we assured that schools with close proximity or sharing the same buildings had the same 'treatment status' to minimize the risk of contamination. In practice, around 30 schools in all were randomized at this level. (For the purpose of power, with an ICC of 0.2, it's approximately a 4% loss in precision.) We did the sample size calculation with an assumption of 0.2 as the ICC. Given this, the issue of randomizing to schools-on-same-grounds rather than individual schools results in a 4% loss of precision in those cases.



Appendix A6: Teacher motivation tool development report by NYU

Teacher Report Adaptation for Use in the Indian Context:
Description, Analyses, and Recommendations
Edward Seidman, Mahjabeen Raza, and Sharon Kim
New York University

There are several objectives of this report:

1. Describe the adaptation process of the Teacher Report for India.
2. Describe the data collection and sampling procedures
3. Present analytic plan and analyses: items and factors
4. Recommendations:
    a. For STIR's future use of TR, and selected sub-scales, in India
    b. For ID Insights use of TR factors to employ in RCT impact analyses

Description of adaptation process

In late 2016, the NYU and STIR India teams collaborated to adapt and contextualize the Teacher Report (TR) to the Indian context for data collection in Delhi and Uttar Pradesh to support the teacher measurement component in a larger RCT evaluation being conducted in India. Earlier in 2016, NYU had developed this teacher report for evaluating STIR programming in Uganda. Building on the findings in Uganda, a set of 46 items were retained in the TR and subsequently adapted for use in India. The TR was then translated into Hindi through an iterative process between NYU and STIR India; this translated version was used for the next step of adaptation. Given the vastly different contexts of the two countries, the first step of adaptation work was to conduct focus groups in India. The TIPPS team drafted a protocol for enacting focus groups. A final protocol for focus groups was established in concert with the STIR India team. The target sample consisted of a mix of teachers that were both affiliated and unaffiliated with STIR. Ms. Tanushree Sarkar, Monitoring Evaluation Associate STIR India was trained by the NYU team to conduct the focus groups and evaluate the information she gathered.

At this stage of development, participating teachers were asked to review and provide feedback on the TR items and share thoughts on any pertinent areas of note that may have been missed. Participants were also asked to complete the TR and share their thoughts on the length and ease of completing the survey, as well as the level of detail in the items. Upon completion of the focus groups, Ms. Sarkar shared the completed surveys, notes, and recording with the NYU team. The NYU team then reviewed those materials and discussed them with Ms. Sarkar. Four critical findings were:

1. Teachers in both groups found the language of the survey somewhat difficult to understand, albeit for different reasons. This was a substantiation of a prior concern – in earlier discussions, the STIR India team had highlighted the fact that the schools and teachers in both locations were markedly different in setting, cultural expression, and in Hindi language fluency. Where Delhi is an urban metropolitan area, teachers are acclimated to using a mix of Hindi, English, and Hinglish. In Uttar Pradesh, teachers are acclimated to using more formal Hindi (with little to no English or Hinglish).
2. Teachers were unfamiliar with Likert scales and found the task of deciphering meaning and proper use of the scale challenging.

3. Some items – notably those tapping growth or fixed mindset – were disruptive and confusing to teachers. Teachers indicated that the compound nature of the items made it difficult to understand and would skip them to revisit after completing other items in the survey. Since these problematic items were scattered in the measure, there was concern that this might lead to a higher number of incomplete surveys.
4. Further, teachers highlighted the nuance with which the ideas of growth and intelligence are understood in the Indian culture. Contrary to the idea of "innate" intelligence or success, teachers indicated that they attributed these characteristics as blessings from God or nature's gift.

Given these observations, the following adaptations were made to the TR:

1. The Hindi used in the survey was significantly revised. Through multiple discussions and iterations, the NYU and STIR India teams worked together to ensure that items were linguistically and culturally attuned. In lieu of transliteration (which would require sophisticated Hindi terminology to capture concepts around teacher mindset), the teams adapted items to ensure their content validity. For instance, keeping with teachers' recommendations on the mindset items, words such as "innate" were translated conceptually to "God given" or "blessed with" to attune the survey to the way Indian teachers could relate to and understand the concept of natural ability.

Both a verbal and written example of the Likert scale was included. The verbal example was included in the enumerator script and Ms. Sarkar incorporated this component in the training she conducted with enumerators. An example of a filled-out Likert scale item was also included on the first page of the survey.

2. All Growth/Fixed Mindset items were incorporated into a separate 8-item section and relegated to the end of the TRS to mitigate teachers' confusion over these items. This change was intended to allow teachers to focus on items they found easy to understand and prompt them on the difficult ones at the end; thus, if respondents did not answer mindset items due to complexity, it would be after they had completed the majority of the survey.

3. The resulting TR retained all 46-items (as in Uganda), however, the items were linguistically and contextually adapted to the Indian context, and re-ordered into two separate sections, as indicated in the previous paragraph. (See Appendix A for copies of the instrument in Hindi and English).

Data Collection

The TR was collected from the endline evaluation of the ongoing RCT in Delhi only. After correcting a number of data coding issues, the data set included 1072 completed protocols. So as not to allow the different treatment (Intrinsic, 38%, Extrinsic, 30%) and control (33%) conditions to differentially affect our psychometric analyses of the TR, we randomly sampled 50% from each of the three conditions to create two roughly equal samples. These samples were then used exclusively to conduct factor and reliability analyses to assess the robustness of the resulting factor solutions.

Analyses: Items and Factors

We examined means, standard deviations, and skew of each item (see Table 1. Item-level Descriptive Statistics). Ultimately, none of the items were excluded from the analysis: some amount of skewness in several items was anticipated. Given the nature of the questions being posed, it was expected that teachers' might have an inclination to respond in a socially desirable manner. There was still sufficient variance across most of the items to pursue factor analytic and reliability analyses. Across all 46 items, there was an insignificant amount (less than 5%) of missing data.

Next, we conducted an Exploratory Factor Analysis (EFA) on the data from sub-sample 1 (N=532). The EFA ensured that the measure was used to capture the impact of the Indian context on the TR factor structure. We then conducted a Confirmatory Factor Analysis (CFA) on sub-sample 2 (N=538) to verify the factor structure that emerged in the EFA. Lastly, we employed Cronbach alphas to examine the internal consistency of each factor.

As part of the EFA in sub-sample 1, the Cattell's Scree test suggested closer examination of the 3-5 factor solutions. Upon examination and evaluation of item loadings on each factor as well as model fit, it was determined that the 5-factor solution provided the best fit (RMSEA=0.043; CFI=0.907; TLI=0.902) and the greatest conceptual clarity. Using this 5-factor EFA solution, we tested the robustness of its fit by conducting a CFA on sub-sample 2. While the model indices were not as strong as the EFA model fit – the 5-factor CFA still revealed good fit statistics (RMSEA=0.059; CFI=0.95; TLI=0.937). Figure 1 (below) presents the measurement model, indicating the relationships between TR factors and associated items. The CFA model too, had good conceptual clarity. Table 2 provides a comparison of the factor model from both exploratory and confirmatory analyses. What follows is a description of each factor (see Table 2).

Factor 1 is labelled *Emotional Exhaustion/Burnout* (n=6 items) and is similar to the factor uncovered in the Uganda data set. But again, the internal consistency coefficients on both sub-samples remain marginal (alphas = 0.54, 0.55, respectively).

Interpretive note: In essence, this sub-construct is remarkably similar in content and reliability to that which emerged in Uganda.

Factor 2 is tentatively labelled *Positive Professional Outlook* (n=8) and did not appear as such in the Uganda data. The internal consistency in both sub-samples was maximized by dropping one (item #4) of the 9 loaded items (alphas = 0.84, 0.85).

Interpretive note: This sub-construct bears little resemblance to any uncovered in the Uganda study, but nevertheless quite relevant. Conceptually, though not empirically, it does appear to be a "cousin" of the following factor teacher efficacy.

Factor 3 is labelled *Teacher Efficacy* (n=21) and is similar, but broader, than the items encompassed in the Self-efficacy/Intrinsic motivation factor found in the Uganda data set. The internal consistency in both sub-samples is excellent (alphas = 0.93, 0.93), which are substantially higher than in Uganda. Beyond teacher efficacy, a small number of items suggest being valued by colleagues, supervisors and family.

Interpretive note: Eleven of the seventeen-items that loaded on the Ugandan Efficacious Mindset are repeated in this sub-construct, with several of the Growth Mindset items loading on a separate factor as originally intended.

Factor 4 contains only four items, and dropping one (#32) maximizes the alpha (n=3; 0.49, 0.52); this item also did not fit as well conceptually. The remaining 3 items tap "feedback from colleagues."

Interpretive note: While conceptually meaningful, the small number of items and the lack of adequate reliability does not support its use in subsequent analyses.

Factor 5 is labelled *Teacher Growth Mindset* and consists of five of the eight items originally constructed to tap growth mindset with the referent being other teachers (n=5; 0.73, 0.77). One of the original 6 items (#34) from the EFA lacked conceptual integration with the others and dropping it maximized the alphas.

Interpretive note: Though these items were designed for the Uganda study, they did not manifest themselves as a unique sub-construct there. In Delhi, they do appear as a distinct sub-construct and with reasonable reliability given that it only consists of 5 items. As noted above, this difference may lie in the manner in which the content of the items were conceptually adapted to the nuances of the Indian context.

Recommendations

Summary of Findings

Four conceptually meaningful factors were revealed, three with good to excellent levels of reliability. *Teacher efficacy* and *Teacher growth mindset* are at the heart of STIR's theory of change and align with the ingredients of the "special sauce." A *Positive professional outlook,* as suggested above, *a conceptual* "cousin of efficacy, would also appear to be of interest to STIR because it is tapping teacher outlook – a potentially important consideration in understanding what makes an effective teacher. While *Emotional exhaustion/Burnout* is not a central concept to STIR, it encompasses the concepts of dissatisfaction and burnout; a large extant literature points to both concepts as central to teacher effectiveness and motivation. However, the reliability of this factor is too low to recommend it for impact analyses.

For STIR's Use of TR in India

Can these four sub-constructs be shortened for future use in India?

Based on the aforementioned recommendations, we are now left with a 40-item instrument.

It may be possible to shave an item or two off of *Positive professional outlook*, creating improved conceptual clarity and not jeopardize its good reliability.

In a similar vein, it may be possible to reduce the number of items on the *Teacher efficacy* sub-construct without sacrificing reliability in a meaningful way.

Any such reductions can be explored at the request of STIR.

Use of Sub-scales in Planned Impact Analyses by ID Insight

Given the recommendations of the previous sub-section, the sub-constructs of *Teacher efficacy* and *Teacher growth mindset* should clearly be utilized. We think there is also a good case for *Positive professional outlook*. But of course, this decision is STIR's alone.
Technically, there are two different ways ID Insight might calculate these sub-construct scores for use in their planned impact analyses.
Use of *absolute weights* to calculate a factor score for each teacher on each of the sub-constructs employed, or
Use of *unit weights* to calculate a factor score for each teacher on each of the sub-constructs employed.

Our recommendation is to leverage the simplicity and conceptual clarity of unit weights. Essentially, it means that for each teacher on each sub-construct, all the actual scores of the specifically-loaded items would be added together and divided by the total number of items. This is consistent with our employment of alpha coefficients and our logic in creating distinct sub-constructs.

Table 1. Item-level Descriptive Statistics (N=1072)

	Item Statement:	***Mean***	***Std. Dev.***	***Skewness***
1	Teaching is mentally draining.	2.90	1.48	0.27
2	With the help of my colleagues, we can solve student issues.	4.97	1.32	-1.85
3	I feel used up at the end of the school day.	3.34	1.47	-0.04
4	My pay as a teacher is insufficient to support my family	3.54	1.78	-0.05
5	I feel fatigued when I get up in the morning and have to face another day at school.	1.86	1.06	1.92
6	I have the ability to get parents involved in their children's education.	4.75	1.29	-1.45
7	I ask my colleagues for feedback.	4.40	1.47	-1.01
8	With the help of my colleagues, we can identify innovative practices.	5.17	1.11	-2.22
9	As a teacher, I'm given more responsibilities than I can manage.	2.79	1.59	0.72
10	Some teachers at my school want to transfer to another school.	2.19	1.29	1.28
11	I do not get paid on time.	1.88	1.27	1.97
12	I can make my classroom a safe space for students, both emotionally and physically.	5.13	1.13	-2.13
13	No matter how much natural ability you may have, you can always find important ways to improve	5.19	1.03	-2.25
14	As a teacher, I am contributing positively to the lives of my students.	5.26	1.03	-2.41

15	I feel energized when my class greets me each morning.	5.33	1.03	-2.50
16	If I had to choose again, I would still want to be a teacher.	5.20	1.16	-2.09
17	My supervisors treat me with respect.	5.40	0.96	-2.73
18	My colleagues at school make work a fun place to be.	4.97	1.15	-1.75
19	My supervisor praises me for my efforts in the school.	5.09	1.02	-1.84
20	Parents value my work as a teacher.	5.27	0.91	-2.16
21	I plan lessons with a colleague.	4.26	1.51	-0.79
22	I feel confident about my abilities as a teacher.	5.38	0.90	-2.78
23	If a student does not remember information in a previous lesson, I would know how to help them remember.	5.27	0.89	-2.16
24	When a student gets a better grade than he or she usually gets, it is because I found a better way.	4.79	1.11	-1.26
25	If a student in my class is undisciplined, I know some techniques to direct him or her.	5.11	0.87	-1.85
26	Every teacher can continue to improve their practice throughout their career.	5.25	0.95	-2.19
27	I can get through to even the most difficult or unmotivated students.	4.81	1.07	-1.15
28	I can motivate students who show low interest in school.	5.14	0.93	-1.84
29	I can influence some of the decisions that are made in the school.	4.66	1.15	-1.10
30	I can get students to work in groups or pairs.	5.01	1.09	-1.82
31	I ask my supervisor for feedback.	4.45	1.49	-1.02
32	I can help students overcome some difficult home and community conditions.	4.29	1.37	-0.75
33	Teachers in my school work closely with supervisors.	5.20	1.01	-2.12
34	I spend too much time traveling to my school.	2.20	1.39	1.30
35	My fellow teachers can be counted on to influence decisions of the school.	4.76	1.13	-1.47
36	When I get new material, I am sure I am able to learn it.	5.33	0.93	-2.47
37	My family is proud that I am a teacher.	5.50	0.87	-3.01
38	Sometimes I share materials with colleagues.	5.02	1.06	-1.83
39	My colleagues praise me for coming up with new ways to teach a lesson.	5.07	0.98	-1.70
40	When I set a goal, no matter how difficult, I will eventually achieve it.	5.23	0.86	-1.75
41	I can learn new things, but I cannot really change my basic intelligence.	5.07	0.96	-1.53
42	The kind of teacher someone is, is something very basic about them, and can't be changed very much.	4.08	1.51	-0.62
43	Every teacher can significantly improve their teaching ability.	3.78	1.66	-0.23
44	Some teachers don't really benefit from professional learning because they have a natural ability.	3.48	1.59	-0.06
45	Teachers can change the way they teach in the classroom,	3.38	1.55	0.09

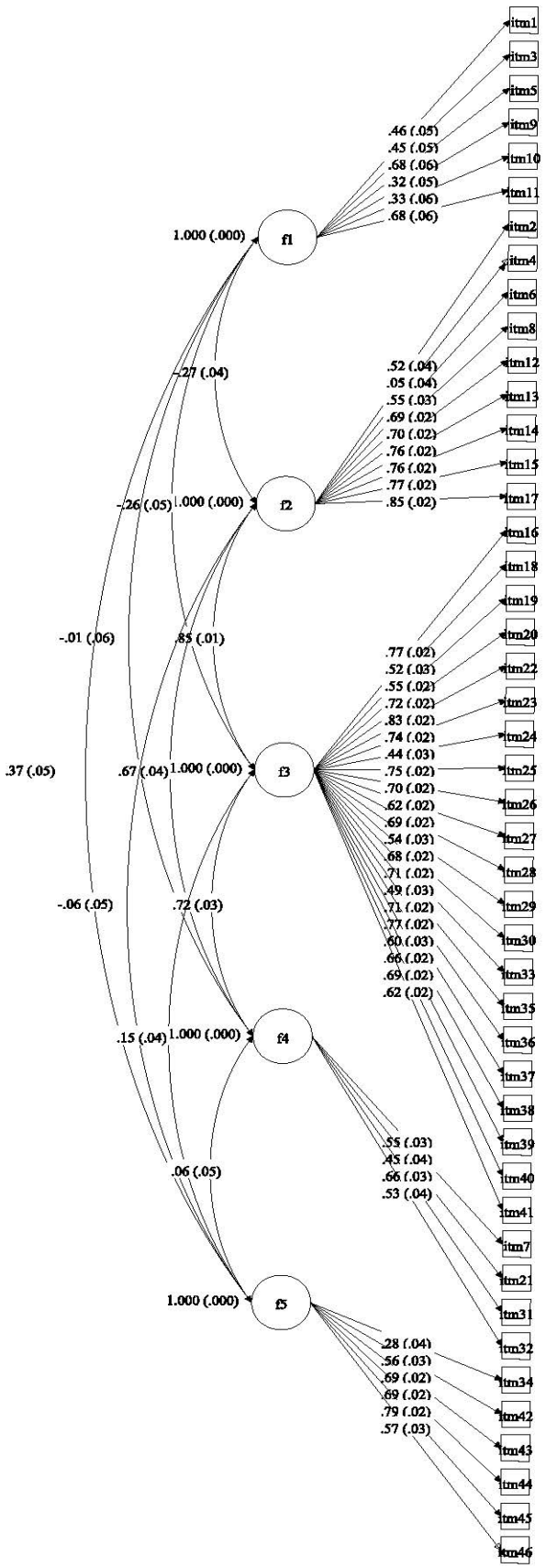
	but they can't really change their true ability.			
46	Some teachers will be ineffective no matter how hard they try to improve.	2.70	1.52	0.68

Table 2. Comparison of item loadings in EFA and CFA

#	**Factor 1**	***EFA***	***CFA***
1	Teaching is mentally draining.	0.53	0.46
3	I feel used up at the end of the school day.	0.38	0.45
5	I feel fatigued when I get up in the morning and have to face another day at school.	0.62	0.68
9	As a teacher, I'm given more responsibilities than I can manage.	0.41	0.32
10	Some teachers at my school want to transfer to another school.	0.47	0.33
11	I do not get paid on time.	0.57	0.68
	Factor 2	***EFA***	***CFA***
2	With the help of my colleagues, we can solve student issues.	0.40	0.52
6	I have the ability to get parents involved in their children's education.	0.38	0.55
8	With the help of my colleagues, we can identify innovative practices.	0.53	0.69
12	I can make my classroom a safe space for students, both emotionally and physically.	0.61	0.70
13	No matter how much natural ability you may have, you can always find important ways to improve	0.66	0.76
14	As a teacher, I am contributing positively to the lives of my students.	0.68	0.76
15	I feel energized when my class greets me each morning.	0.71	0.77
17	My supervisors treat me with respect.	0.42	0.85
	Factor 3	***EFA***	***CFA***
16	If I had to choose again, I would still want to be a teacher.	0.50	0.77
18	My colleagues at school make work a fun place to be.	0.30	0.52
19	My supervisor praises me for my efforts in the school.	0.41	0.55
20	Parents value my work as a teacher.	0.52	0.72
22	I feel confident about my abilities as a teacher.	0.57	0.83
23	If a student does not remember information in a previous lesson, I would know how to help them remember.	0.61	0.74
24	When a student gets a better graded than he or she usually gets, it is because I found a better way.	0.53	0.44
25	If a student in my class is undisciplined, I know some techniques to direct him or her.	0.61	0.75
26	Every teacher can continue to improve their practice throughout their career.	0.48	0.70

27	I can get through to even the most difficult or unmotivated students.	0.79	0.62
28	I can motivate students who show low interest in school.	0.69	0.69
29	I can influence some of the decisions that are made in the school.	0.56	0.54
30	I can get students to work in groups or pairs.	0.44	0.68
33	Teachers in my school work closely with supervisors.	0.43	0.71
35	My fellow teachers can be counted on to influence decisions of the school.	0.43	0.49
36	When I get new material, I am sure I am able to learn it.	0.72	0.71
37	My family is proud that I am a teacher.	0.671	0.77
38	Sometimes I share materials with colleagues.	0.37	0.60
39	My colleagues praise me for coming up with new ways to teach a lesson.	0.63	0.66
40	When I set a goal, no matter how difficult, I will eventually achieve it.	0.611	0.69
41	I can learn new things, but I cannot really change my basic intelligence.	0.74	0.62
	Factor 4	***EFA***	***CFA***
7	I ask my colleagues for feedback.	0.67	0.55
21	I plan lessons with a colleague.	0.29	0.45
31	I ask my supervisor for feedback.	0.70	0.66
	Factor 5	***EFA***	***CFA***
42	The kind of teacher someone is, is something very basic about them, and can't be changed very much.	0.50	0.56
43	Every teacher can significantly improve their teaching ability.	0.70	0.69
44	Some teachers don't really benefit from professional learning because they have a natural ability.	0.59	0.69
45	Teachers can change the way they teach in the classroom, but they can't really change their true ability.	0.68	0.79
46	Some teachers will be ineffective no matter how hard they try to improve.	0.55	0.57

Figure 1. CFA Model of the TR (Delhi)



Appendix A7: Classroom observation tool — development and iteration

Literature

Instructional time and appropriate classroom management time are key inputs into a production function of student learning outcomes (Glewwe and Kremer 2006). The former can be increased through adding hours to the school day or recovering minutes from time off-task during any given school period (Ganimian and Murnane 2014).

A key way of assessing time use during a class period comes from the work of Jane Stallings (Stallings 1977; World Bank 2015).[11] The full Stallings Classroom Snapshot is a well-established structured observation tool that captures how classroom inputs — including time — are employed to improve learning. The tool has been used in low- and middle-income countries in Latin America and Sub-Saharan Africa to describe classroom practice (Bruns and Luque 2015; Bruns, De Gregoria, and Taut 2016; Bold et al. 2016). Through repeated use of the Stallings snapshot by researchers (though not in India), a rough benchmark has been set for quality teaching: 85% of class time spent teaching and 15% of time spent on classroom management (with 0% of time spent off-task) (Bruns, De Gregoria, and Taut 2016).

Much of STIR's focus is on changing the tone and environment in the classroom in a way that is ultimately more conducive for learning. To capture this, we drew on a series of child-friendliness indicators developed by the ASER Centre to highlight classroom features in India, based on easily observable aspects of India's National Curriculum Framework's guidance on good teacher practice (NCERT 2005; S. Bhattacharjea, Wadhwa, and Banerji 2011; Suman Bhattacharjea 2017). These represent a subset of the inclusive, child-centered, and physically and emotionally healthy conception of child-friendly schools deployed by UNICEF (UNICEF 2006). To our knowledge, most indicators related to child-friendliness have been used descriptively, to provide reports on and comparisons between different schools, rather than as outcomes in an impact evaluation.

We developed a classroom practice observation tool that drew on key components of the Stallings snapshot.

Overview of tool

The classroom observation tool used was adapted from the Stallings Classroom Snapshot a tool developed by Jane Stallings in 1977 (Stallings 1977; World Bank 2015). The snapshot captures how classroom inputs are employed to improve learning. This includes how a teacher spends own time and what physical resources and materials are used in the classrooms.

We used the snapshot to assess teachers' behavior and practices within the classroom and the changes that may arise because of STIR's program. Enumerators 'sit-in' in classrooms and code student and teacher activities, four times for an interval of five minutes each. The

[11] Only two measures of time use in classrooms have gained traction to date in low- and middle-income countries — the Classroom Assessment Scoring System (CLASS) and the Stallings classroom observation instrument (Bruns, De Gregoria, and Taut 2016).

Stallings tool is a well-established and widely used observational tool used to gauge classrooms and has especially been extensively used in classrooms in Latin America.

Adaptation to the original Stallings

Certain changes were made to the Stallings tool in its original form to make it more suitable for our use.

- The number of observations was reduced to four to make it more suitable to Indian classrooms and easier to administer.
- A section was added to capture and quantify the flow of verbal interactions within the classroom. This was like the Flanders tool (Flanders 1961; Amidon 1966).[12] It involved rapidly capturing 'who is speaking?' and 'what is being said?' This was recorded thirty times over a two-and-a-half-minute window, in the middle of the observation period.
- A section was also added to capture the content level being taught in the class.
- Finally, the a set of child-friendliness indicators developed by ASER was added to one of the observations (S. Bhattacharjea, Wadhwa, and Banerji 2011).

Baseline instruments

While the Stallings classroom snapshot tool includes multiple components of classroom activities, our main focus is on teacher time-use. Time use is divided into three main, pre-coded categories: teaching, managing the classroom, or off-task.[13]

To capture this, an enumerator is placed, as unobtrusively as possible, in a classroom to take note of what is happening in the classroom at specific points in time. At these points in time (signaled through the SurveyCTO software used for data collection), enumerators record what the teacher is doing at that moment, disregarding what has been happening any time prior (SurveyCTO (version 2.02) 2016).[14]

---

[12] The verbal interaction tool recorded the speaker and content of speech for several minutes inside the classroom. This data was then used to compute metrics such as teachers' reactions to students' questions, the relative frequency teachers praising or criticizing students, etc.

[13] Teaching is defined to include time spent by a teacher in lecture, answering academic matter related questions, correction of students' academic work and engaging with all or a group of students with regards to what is being taught in the classrooms. Classroom management is defined to include non-academic clarifications of doubts and questions, general classroom discipline and behavior, and other defined administrative responsibilities such as attendance, checking uniforms etc. Finally, if a teacher is spending time on any activity other than teaching and classroom management, then the teacher is off-task.

[14] We also made some additions to the observation tool not strictly related to time use. First, we added to our tool a section to capture the content level being taught in the class. Second, to help contextualize our findings, we also included a one-time capture of basic classroom amenities, such as whether there were desks for the students and teacher. These results are presented in Appendix A27. Finally, we included a section on classroom interactions between teachers and students, which drew on the work of Flanders (Flanders 1961; Amidon 1966). However, we ultimately found these data difficult to collect and to meaningfully analyze, with no clear sense of what types of changes we 'should' expect over time in an improving classroom. We thus dropped this section from the classroom practice observation tool for midline. We discuss this in more detail in Appendix A13.

Due to the realities of school and classroom entry and timing in India, we made two central changes to how the snapshot is used in our classroom practice observation tool. (More details on the development of the tool are in Appendix A13).

First, the original tool calls for ten recording points during a single classroom session. Given the length of class times in India, however, we reduced the number of points to four; every five minutes, enumerators selected from a coded menu of descriptions of what was occurring. Note that since the time-use categories are mutually exclusive and since they capture particular points in time rather than duration of activities, the result can be 0%, 25%, 50%, 75%, or 100% for any time-use category, since each observation has four 'rounds' of measurement.

Second, in other settings, enumerators go to a classroom in advance of class starting and begin recording from the time the class is scheduled to start. As such, if the teacher is five minutes late getting to class or starting a specific lesson, those 5 minutes would be recorded as "off-task." However, we rely on teachers for the entry to their classrooms and the timing of specific lessons is often slightly fluid.[15] Thus, we only begin recording once the teacher enters the classroom; our results are internally consistent but should not be taken as directly comparable to the results of other researchers deploying Stallings from the beginning of scheduled class timings.

Outcomes from the classroom observation tool

We broadly classified outcomes from the classroom observation tool into two 'families'. These are the 'time use' and the 'child friendliness' families.

Time use: There are two indicators of time use that we look at as outcomes:

- **Teaching time:** Defined as percentage of times teachers were *observed* as teaching. Note: For each teacher this could be either 0% or 25% or 50% or 75% or 100% since each observation has four 'rounds' of observation.
- **Off-task time:** Defined as the percentage of times teachers were *observed* as off-task. As with teaching time, this could be either 0% or 25% or 50% or 75% or 100% since each observation has four 'rounds' of observation.

Amongst these, teaching time is considered as the primary indicator. Both indicators are coded as response to the question *'What is the teacher doing?'* (Q1 in section A, C, E, and G in the classroom observation tool).

Child friendliness: There are six indicators of child friendliness that were used as outcomes:

- **Teacher smiled/joked:** Defined as the percentage of times teachers were *observed* as smiling or joking with students at least once during the observation window. (Q4 in section B, D, F, and H in the classroom observation tool).

[15] In several instructive episodes, teachers told us that they were not necessarily planning to teach a particular class (in favor of a cup of chai) but since we were there, they would teach after all.

- **Students asked at least one question:** Defined as the percentage of times teachers were *observed* as being asked at least one questions by students during the observation window. (Q5 in section B, D, F, and H in the classroom observation tool).
- **Teachers praised or displayed students' work:** Defined as the percentage of times teachers were *observed* as praising or displaying work done by students within the classroom at least once during the observation window. (Q6 in section B, D, F, and H in the classroom observation tool).
- **Teachers made use of local information, materials:** Defined as the percentage of times teachers were *observed* as having used local information or materials while teaching at least once during the observation window. (Q7 in section B, D, F, and H in the classroom observation tool).
- **Teachers used learning aides:** Defined as the percentage of times teachers were *observed* using learning aids while teaching at least once during the observation window. (Q8 in section B, D, F, and H in the classroom observation tool).
- **Students worked in small groups/ pairs:** Defined as the percentage of times teachers were *observed* as encouraging students to work in small groups or pairs at least once during the observation window. (Q9 in section B, D, F, and H in the classroom observation tool).

While STIR views these as 'good' classroom practices, note that none of these behaviors were necessarily stressed during the first year of STIR's programming as evaluated here. It is possible that different micro-innovations incorporated these elements but in Year 1, these aspects of classroom culture and practice were not specifically endorsed by STIR.

Changes from baseline to midline

The observation tool was changed between baseline and midline based on learnings from baseline and to make it more useful for STIR in the following ways:

- **Dropping the Flanders' section:** The Flanders section was based on work by Flanders to study classroom interactions (Flanders 1961). This section has been used to capture communication within classrooms but presented a few complications:
    - Flanders works best in classrooms where there is a two-way communication channel. In Indian classrooms, especially in schools where the evaluation took place, most communication is still led by teachers. The analysis of the data was complex, and made it tough to make statements interesting/ useful for STIR.
    - The Flanders tool was also tough to administer from an enumerator standpoint due to challenges in classifying and coding the classroom communication in the given categories and the strict timing component.
- **ASER matrix:** At baseline, we only captured child-friendly behaviors once during a single 40-minute classroom observation. At midline, we decided to look for these indicators four times through the classroom observation. This was to allow us a clearer and more confident picture of what was happening in the classroom, given STIR's interest in the baseline descriptive results

using this indicators. We also added two further indicators based on our understanding of STIR's program – 1) **Teachers praised or displayed students' work** (Defined as the percentage of times teachers were *observed* as praising, showing-off, or displaying work done by students within the classroom at least once during the observation window) and 2) **Teacher refers to students by name** (Defined as the percentage of times teachers were *observed* as always referring to students by their name during the observation window). The classroom practice observation tool used at midline is included in Appendix A8.

- **Changing question on student activities:** The Stallings tool had a question on student activities, namely the number of students engaged in different activities. However, given the lack of agency students have within a classroom the interpretation of the results became difficult. Instead, we replaces the more detailed observation of student activities with a simpler and more meaningful (from STIR's point of view) set of questions. We used a two-part question to gauge if students in the classroom are following instructions given by teachers (on task) or not.
- **Adding a new section on classroom information:** A new section was added which enumerators filled out before the 'main observation'. This included classroom-level information, including number of students; number of girls and boys; if outside noise affects the classroom, etc. We felt this would be useful in providing context to our analysis of how classrooms are functioning.

Lessons learnt

The Stallings tool is relevant for STIR as it is useful in understanding teacher activities in the classroom. The ASER matrix includes themes which teachers focus on as part of their first year journey with STIR via micro innovations and otherwise. These include emphasis on questions asked by students, smiling, displaying students' work etc.

A limitation of the Stallings tool in its original form and even in the adapted form we use, is that it does not allow us to distinguish between good and bad observations; *e.g.*, "drills" could be productive or could reflect recitation without learning. Further it does not capture quality of instructions and makes it tough for us to make normative statements.

Appendix A8: Classroom observation tool (endline version)

Observation date: [ ][ ] – [ ][ ] – [ ][ ]
Classroom entry time: [ ][ ] : [ ][ ]

Begin observation 1 (5 minutes after classroom entry time): [ ][ ] : [ ][ ]
Record observation 1 (8 minutes after classroom entry time): [ ][ ] : [ ][ ]
Record observation 2 (11 minutes after classroom entry time): [ ][ ] : [ ][ ]
Record observation 3 (14 minutes after classroom entry time): [ ][ ] : [ ][ ]
Record observation 4 (17 minutes after classroom entry time): [ ][ ] : [ ][ ]
Record observation 5 (20 minutes after classroom entry time): [ ][ ] : [ ][ ]
Record observation 6 (23 minutes after classroom entry time): [ ][ ] : [ ][ ]
Record observation 7 (26 minutes after classroom entry time): [ ][ ] : [ ][ ]

Before beginning with the observation, please note the following:

1. Number of students –
2. Number of girls visible in the class –
3. Number of boys visible in the class –
4. How many teachers are in present the classroom?
   a. None
   b. 1
   c. 2
   d. 3 or above
5. What best describes the classroom?
   a. Open/ outdoor class
   b. Roofed but open from the sides
   c. Covered with walls
6. How would you describe the way the students are seated?
   a. In rows
   b. In groups
   c. No particular arrangement
7. Majority of the students are on:
   a. Bare floor
   b. Mats
   c. Seats with tables
   d. Seats without tables
   e. Not seated
8. Please note the following as yes, no or unclear:

Statement	Yes	No	Unclear
Are children wearing uniform?			
Does outside noise effect communication?			
Does the classroom have a blackboard or whiteboard?			
Is there a chair and/or a table for the teacher?			
Are there posters, etc., on the walls or otherwise on display (other than student work)?			
Is student work (posters, drawings, etc.) on display in the classroom?			

STOP. Please wait on this page until it is time to begin observation 1.

IDinsight
DATA. DECISIONS. DEVELOPMENT.

OBSERVATION 1 RECORD TIME…… [ ][ ] : [ ][ ]

A. Classroom Snapshot (1): *Answer questions about this moment (now).*

1 What is the teacher doing?

A Teaching students (discussing academic material)
B Classroom management (discipline, attendance, or other non-academic interaction)
C Out of classroom or off-task

2 What are students supposed to be doing?

Listening to, watching the teacher or repeating what the teacher says	
Working or discussing in pairs, groups or as a class	
Working quietly (individually)	
Sitting or standing quietly for non-academic purposes (such as uniform distribution etc.)	
No particular instructions on what they are supposed to be doing	
Unclear	

3a Based on the instructions given by teachers mentioned above which of the following most accurately describes the students?

A Students are engaged in whatever they are supposed to do
B Students are not engaged in whatever they are supposed to do
C Unclear (only when unclear option is selected in question 2; loop ends here for this)

3b (*For options A and B in 3a*) To what extent are the students engaged or not engaged?

A Somewhat
B Very much

B. Classroom overview (1): *Please answer this question based on the past five minutes only.*

4 Did the teacher smile, laugh or joke with at least some students?

A Yes
B No
C Don't know

5 Did the students ask the teacher at least one question?

A Yes
B No
C Don't know

6 Did the teacher praise at least one child or shared/showcased the work of one child in front of the rest of the class?

A Yes
B No
C Don't know

7 Did the teacher use local information to make academic content relevant?

This includes use of objects, events, places or people with which students are familiar.

A Yes
B No
C Don't know

8 Did the teacher use any learning aides (posters, chalkboard, supplies) other than the textbook?

A Yes
B No
C Don't know

9 Did the teacher ask children to work in small groups or pairs?

A Yes
B No
C Don't know

10 Did the teacher always refer to her students by their name?

A Yes
B No
C Don't know

11 What topics were covered during this class?

Math

Single-digit numbers	
Double-digit numbers	
Addition	
Subtraction	
Multiplication	
Division	
Fractions	

Other math	
Hindi	
Letters	
Words	
Sentences	
Stories	
Vocabulary	
Other	
Other	
Other subject	

STOP.

Please wait on this page until it is time to begin observation 2.

OBSERVATION 2 RECORD TIME…… [ ][ ] : [ ][ ]

C. Classroom Snapshot (2): *Answer questions about this moment (now).*

1 What is the teacher doing?

A Teaching students (discussing academic material)
B Classroom management (discipline, attendance, or other non-academic interaction)
C Out of classroom or off-task

2 What are students supposed to be doing?

Listening to, watching the teacher or repeating what the teacher says	
Working or discussing in pairs, groups or as a class	
Working quietly (individually)	
Sitting or standing quietly for non-academic purposes (such as uniform distribution etc.)	
No particular instructions on what they are supposed to be doing	
Unclear	

3a Based on the instructions given by teachers mentioned above which of the following most accurately describes the

students?

A Students are engaged in whatever they are supposed to do
B Students are not engaged in whatever they are supposed to do
C Unclear (only when unclear option is selected in question 2; loop ends here for this)

3b (*For options A and B in 3a*) To what extent are the students engaged or not engaged?

A Somewhat
B Very much

D. Classroom overview (2): *Please answer this question based on the past five minutes only.*

4 Did the teacher smile, laugh or joke with at least some students?

A Yes
B No
C Don't know

5 Did the students ask the teacher at least one question?

A Yes
B No
C Don't know

6 Did the teacher praise at least one child or shared/showcased the work of one child in front of the rest of the class?

A Yes
B No
C Don't know

7 Did the teacher use local information to make academic content relevant?

This includes use of objects, events, places or people with which students are familiar.

A Yes
B No
C Don't know

8 Did the teacher use any learning aides (posters, chalkboard, supplies) other than the textbook?

A Yes
B No

C Don't know

9 Did the teacher ask children to work in small groups or pairs? ☐

A Yes
B No
C Don't know

10 Did the teacher always refer to her students by their name? ☐

A Yes
B No
C Don't know

11 What topics were covered during this class?

Math

Single-digit numbers	☐
Double-digit numbers	☐
Addition	☐
Subtraction	☐
Multiplication	☐
Division	☐
Fractions	☐
Other math	☐

Hindi

Letters	☐
Words	☐
Sentences	☐
Stories	☐
Vocabulary	☐
Other	☐

Other

Other subject	☐
	☐

STOP.
Please wait on this page until it is time to begin observation 3.

OBSERVATION 3 RECORD TIME…… [ ][ ] : [ ][ ]

E. Classroom Snapshot (3): *Answer questions about this moment (now).*

1 What is the teacher doing? ☐

A Teaching students (discussing academic material)

B Classroom management (discipline, attendance, or other non-academic interaction)
C Out of classroom or off-task

2 What are students supposed to be doing?

Listening to, watching the teacher or repeating what the teacher says	
Working or discussing in pairs, groups or as a class	
Working quietly (individually)	
Sitting or standing quietly for non-academic purposes (such as uniform distribution etc.)	
No particular instructions on what they are supposed to be doing	
Unclear	

3a Based on the instructions given by teachers mentioned above which of the following most accurately describes the students?

A Students are engaged in whatever they are supposed to do
B Students are not engaged in whatever they are supposed to do
C Unclear (only when unclear option is selected in question 2; loop ends here for this)

3b (*For options A and B in 3a*) To what extent are the students engaged or not engaged?

A Somewhat
B Very much

F. Classroom overview (3): *Please answer this question based on the past five minutes only.*

4 Did the teacher smile, laugh or joke with at least some students?

A Yes
B No
C Don't know

5 Did the students ask the teacher at least one question?

A Yes
B No
C Don't know

6 Did the teacher praise at least one child or shared/showcased the work of one child in front of the rest of the class?

A Yes
B No
C Don't know

7 Did the teacher use local information to make academic content relevant?

This includes use of objects, events, places or people with which students are familiar.

A Yes
B No
C Don't know

8 Did the teacher use any learning aides (posters, chalkboard, supplies) other than the textbook?

A Yes
B No
C Don't know

9 Did the teacher ask children to work in small groups or pairs?

A Yes
B No
C Don't know

10 Did the teacher always refer to her students by their name?

A Yes
B No
C Don't know

11 What topics were covered during this class?

Math

Single-digit numbers	
Double-digit numbers	
Addition	
Subtraction	
Multiplication	
Division	
Fractions	
Other math	

Hindi

Letters	
Words	

Sentences	
Stories	
Vocabulary	
Other	

Other

Other subject	

STOP.

Please wait on this page until it is time to begin observation 4.

OBSERVATION 4 RECORD TIME…… [ ][ ] : [ ][ ]

G. Classroom Snapshot (4): *Answer questions about this moment (now).*

1 What is the teacher doing?

A Teaching students (discussing academic material)
B Classroom management (discipline, attendance, or other non-academic interaction)
C Out of classroom or off-task

2 What are students supposed to be doing?

Listening to, watching the teacher or repeating what the teacher says	
Working or discussing in pairs, groups or as a class	
Working quietly (individually)	
Sitting or standing quietly for non-academic purposes (such as uniform distribution etc.)	
No particular instructions on what they are supposed to be doing	
Unclear	

3a Based on the instructions given by teachers mentioned above which of the following most accurately describes the students?

A Students are engaged in whatever they are supposed to do
B Students are not engaged in whatever they are supposed to do

C Unclear (only when unclear option is selected in question 2; loop ends here for this)

3b (*For options A and B in 3a*) To what extent are the students engaged or not engaged?

A Somewhat
B Very much

H. Classroom overview (4): *Please answer this question based on the past five minutes only.*

4 Did the teacher smile, laugh or joke with at least some students?

A Yes
B No
C Don't know

5 Did the students ask the teacher at least one question?

A Yes
B No
C Don't know

6 Did the teacher praise at least one child or shared/showcased the work of one child in front of the rest of the class?

A Yes
B No
C Don't know

7 Did the teacher use local information to make academic content relevant?

This includes use of objects, events, places or people with which students are familiar.

A Yes
B No
C Don't know

8 Did the teacher use any learning aides (posters, chalkboard, supplies) other than the textbook?

A Yes
B No
C Don't know

9 Did the teacher ask children to work in small groups or pairs?

A Yes
B No

C Don't know

10 Did the teacher always refer to her students by their name? ☐

A Yes
B No
C Don't know

11 What topics were covered during this class?

Math

Single-digit numbers	☐
Double-digit numbers	☐
Addition	☐
Subtraction	☐
Multiplication	☐
Division	☐
Fractions	☐
Other math	☐

Hindi

Letters	☐
Words	☐
Sentences	☐
Stories	☐
Vocabulary	☐
Other	☐

Other

Other subject	☐
	☐

STOP.
Please wait on this page until it is time to begin observation 5.

OBSERVATION 5 RECORD TIME…… [ ][ ] : [ ][ ]

A. Classroom Snapshot (5): *Answer questions about this moment (now).*

1 What is the teacher doing? ☐

A Teaching students (discussing academic material)
B Classroom management (discipline, attendance, or other non-academic interaction)
C Out of classroom or off-task

2 What are students supposed to be doing?

Listening to, watching the teacher or repeating what the	☐

teacher says	
Working or discussing in pairs, groups or as a class	
Working quietly (individually)	
Sitting or standing quietly for non-academic purposes (such as uniform distribution etc.)	
No particular instructions on what they are supposed to be doing	
Unclear	

3a Based on the instructions given by teachers mentioned above which of the following most accurately describes the students?

A Students are engaged in whatever they are supposed to do
B Students are not engaged in whatever they are supposed to do
C Unclear (only when unclear option is selected in question 2; loop ends here for this)

3b (*For options A and B in 3a*) To what extent are the students engaged or not engaged?

A Somewhat
B Very much

B. Classroom overview (5): *Please answer this question based on the past five minutes only.*

4 Did the teacher smile, laugh or joke with at least some students?

A Yes
B No
C Don't know

5 Did the students ask the teacher at least one question?

A Yes
B No
C Don't know

6 Did the teacher praise at least one child or shared/showcased the work of one child in front of the

rest of the class?

A Yes
B No
C Don't know

7 Did the teacher use local information to make academic content relevant?

This includes use of objects, events, places or people with which students are familiar.

A Yes
B No
C Don't know

8 Did the teacher use any learning aides (posters, chalkboard, supplies) other than the textbook?

A Yes
B No
C Don't know

9 Did the teacher ask children to work in small groups or pairs?

A Yes
B No
C Don't know

10 Did the teacher always refer to her students by their name?

A Yes
B No
C Don't know

11 What topics were covered during this class?

Math

Single-digit numbers	
Double-digit numbers	
Addition	
Subtraction	
Multiplication	
Division	
Fractions	
Other math	

Hindi

Letters	
Words	
Sentences	
Stories	

Vocabulary	
Other	

Other

Other subject	

STOP.
Please wait on this page until it is time to begin observation 6.

OBSERVATION 6 RECORD TIME…… [ ][ ] : [ ][ ]

A. Classroom Snapshot (6): *Answer questions about this moment (now).*

1 What is the teacher doing?

A Teaching students (discussing academic material)
B Classroom management (discipline, attendance, or other non-academic interaction)
C Out of classroom or off-task

2 What are students supposed to be doing?

Listening to, watching the teacher or repeating what the teacher says	
Working or discussing in pairs, groups or as a class	
Working quietly (individually)	
Sitting or standing quietly for non-academic purposes (such as uniform distribution etc.)	
No particular instructions on what they are supposed to be doing	
Unclear	

3a Based on the instructions given by teachers mentioned above which of the following most accurately describes the students?

A Students are engaged in whatever they are supposed to do
B Students are not engaged in whatever they are supposed to do
C Unclear (only when unclear option is selected in question 2; loop ends here for this)

3b (*For options A and B in 3a*) To what extent are the students engaged or not engaged?

A Somewhat
B Very much

B. Classroom overview (6): *Please answer this question based on the past five minutes only.*

4 Did the teacher smile, laugh or joke with at least some students?

A Yes
B No
C Don't know

5 Did the students ask the teacher at least one question?

A Yes
B No
C Don't know

6 Did the teacher praise at least one child or shared/showcased the work of one child in front of the rest of the class?

A Yes
B No
C Don't know

7 Did the teacher use local information to make academic content relevant?

This includes use of objects, events, places or people with which students are familiar.

A Yes
B No
C Don't know

8 Did the teacher use any learning aides (posters, chalkboard, supplies) other than the textbook?

A Yes
B No
C Don't know

9 Did the teacher ask children to work in small groups or pairs?

A Yes
B No
C Don't know

10 Did the teacher always refer to her students by their name? [ ]

A Yes
B No
C Don't know

11 What topics were covered during this class?

Math

Single-digit numbers	[ ]
Double-digit numbers	[ ]
Addition	[ ]
Subtraction	[ ]
Multiplication	[ ]
Division	[ ]
Fractions	[ ]
Other math	[ ]

Hindi

Letters	[ ]
Words	[ ]
Sentences	[ ]
Stories	[ ]
Vocabulary	[ ]
Other	[ ]

Other

Other subject	[ ]
	[ ]

STOP.

Please wait on this page until it is time to begin observation 7.

OBSERVATION 7 RECORD TIME…… [ ][ ] : [ ][ ]

A. Classroom Snapshot (7): *Answer questions about this moment (now).*

1 What is the teacher doing? [ ]

A Teaching students (discussing academic material)
B Classroom management (discipline, attendance, or other non-academic interaction)
C Out of classroom or off-task

2 What are students supposed to be doing?

Listening to, watching the teacher or repeating what the teacher says	[ ]
Working or discussing in pairs, groups or as a class	[ ]

Working quietly (individually)	
Sitting or standing quietly for non-academic purposes (such as uniform distribution etc.)	
No particular instructions on what they are supposed to be doing	
Unclear	

3a Based on the instructions given by teachers mentioned above which of the following most accurately describes the students?

A Students are engaged in whatever they are supposed to do
B Students are not engaged in whatever they are supposed to do
C Unclear (only when unclear option is selected in question 2; loop ends here for this)

3b (*For options A and B in 3a*) To what extent are the students engaged or not engaged?

A Somewhat
B Very much

B. Classroom overview (7): *Please answer this question based on the past five minutes only.*

4 Did the teacher smile, laugh or joke with at least some students?

A Yes
B No
C Don't know

5 Did the students ask the teacher at least one question?

A Yes
B No
C Don't know

6 Did the teacher praise at least one child or shared/showcased the work of one child in front of the rest of the class?

A Yes

B No
C Don't know

7 Did the teacher use local information to make academic content relevant?
This includes use of objects, events, places or people with which students are familiar.
A Yes
B No
C Don't know

8 Did the teacher use any learning aides (posters, chalkboard, supplies) other than the textbook?
A Yes
B No
C Don't know

9 Did the teacher ask children to work in small groups or pairs?
A Yes
B No
C Don't know

10 Did the teacher always refer to her students by their name?
A Yes
B No
C Don't know

11 What topics were covered during this class?

Math	
Single-digit numbers	
Double-digit numbers	
Addition	
Subtraction	
Multiplication	
Division	
Fractions	
Other math	
Hindi	
Letters	
Words	
Sentences	
Stories	
Vocabulary	
Other	

Other

Other subject	

END OF OBSERVATION IN THIS CLASSROOM.
You have completed the observation part of the survey in this classroom.
Please find a suitable time and ask the teacher the following.

12 Which book(s) were being used for teaching today?

Book1	
Book2	
Book3	
Book4	
Book5	
Book6	
Book7	
Book8	

For each book what was the chapter *number* being covered in the class today?

Book1	
Book2	
Book3	
Book4	
Book5	
Book6	
Book7	
Book8	

(*If possible*) For each book what was the page *number* being covered in the class today?

Book1	
Book2	
Book3	
Book4	
Book5	
Book6	
Book7	
Book8	

Appendix A9: Student learning tool

Literature

One common and popular way of understanding student competency in core subjects has been developed by Pratham and the ASER Centre. The ASER student testing tool (across different languages and competencies) is widely used in India (in the Annual State of Education Reports; Bhattacharjea S 2013; Banerjee S, Mandal K S, and De P; Shotland M, Berry J and Banerji R 2014) and is now also used in Sub-Saharan Africa. The popularity of the ASER tool comes from what it captures — basic reading and arithmetic which are considered as foundational skills; and the advantage of the tools being simple, quick, cost-effective, and easy to train examiners to administer (Wagner 2003). The yearly ASER survey (using the ASER tool) is widely regarded as providing information on Indian education by allowing for rapid assessments (Muralitharan 2013)[16]. For the purposes of our evaluations, the ASER tool helped capture 'learning' well by assessing H Hindi reading (the local language in northern India) and math levels.

`

Hindi baseline instruments

To measure student competency in Hindi, we rely on a modified version of the ASER test for learning levels ("Annual Status of Education - Rural" 2005). Our modifications were to extend the range of skills (learning levels) covered by the tests as our sample includes 1st to 8th grades. Specifically, we included two additional levels of "story" that were slightly more difficult than the other stories to avoid potential ceiling effects. Difficulty was assessed based on a combination of a few different criteria – total number of words, total number of sentences, number of words with four letters, number of words with more than four letters, words with half letters, complexity of specific words. The added 'levels' were piloted extensively before baseline.

We illustrate how the tool is administered using the Hindi assessment as an example. A student's Hindi level is defined as the highest question he/ she answers correctly. Students in different grades start from different places in the student testing tool. Details can be seen in Table A7, below.

Table* A7*: Starting point to administer the Hindi assessment tool in Delhi and U.P.

Grade	Starting point Delhi	Starting point U.P.
First	Letters	Letters
Second	Words	Letters
Third	Paragraph	Letters
Fourth	Story 1	Letters
Fifth	Story 2	Story 1
Sixth	Story 2	Story 1
Seventh	Story 2	Story 2
Eighth	Story 2	Story 2

[16] Note however Muralitharan (2013) does call for the use of more 'advanced' tool in case a deeper diagnosis *vis-à-vis* the whole range of the syllabus is required.

The testing tool is ordinal in nature, *i.e*, if the student is at paragraph level, that would imply that they can do everything from letters through paragraph but not Story 1 and above. From the first question a student is given (as per grade level, shown in Table A7), the student can either move up to a higher question (if they answer correctly) or move down to a lower question (if they answer incorrectly). To illustrate, a fourth-grade student in Delhi would start the Hindi test from level the Story 2 level. If the student gets this question correct, they will be asked a question about Story 3. If not, the student will be asked about Story 1; if they cannot read Story 1, they will be asked to try the paragraph, and so on.

For letters and words, students need to read three out of five correctly to progress to the next question (see Appendix A10 for the Hindi testing tool). For paragraph and stories, students are allowed a maximum of three errors.[17]

In developing an expectation about how much Year 1 of STIR's programming might affect Hindi learning outcomes, it is important to remember while STIR aims to improve student learning outcomes overall, (a) STIR does not specifically target teachers focused on Hindi and (b) the changes teachers make in their classroom in the first year do not necessarily target pedagogical strategies to improve Hindi acquisition.

Update for endline measurement

Based on our baseline results, we added an additional level story level to avoid potential ceiling effects in our results. The instrument we used at endline are in Appendix A10.

Math baseline instruments

To measure student competency in math, we rely on a modified version of the ASER test for learning levels ("Annual Status of Education - Rural" 2005). Our modifications were to extend the range of skills (learning levels) covered by the tests as our sample includes 1st to 8th grades. Specifically, we included fractions beyond being able to do division.

Update for endline measurement

There were no updates made from baseline to endline

[17] All errors in one word are considered one and reading the same word incorrectly over and over is also considered one. For all questions, if the student is unsuccessful in the first attempt (based on the conditions mentioned above) they are allowed a second chance.

Appendix A10: Student learning tool — Hindi and Math (Sample A & Sample B) (endline version)

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Student Testing Tool

STIR Impact Evaluation

Sample A

8th July 2017

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Oral Tool - IDinsight testing – Sample A **हिंदी**

सर्वेक्षक ध्यान दें:

- *कक्षा 1 के छात्रों के साथ अक्षर के स्तर से शुरू करें।*
- *कक्षा 2 के छात्रों के साथ **शब्द** के स्तर से शुरू करें।*
- *कक्षा 3 के छात्रों के साथ **अनुच्छेद** के स्तर से शुरू करें।*
- *कक्षा 4 के छात्रों के साथ **कहानी 1** के स्तर से शुरू करें।*
- *कक्षा 5 के छात्रों के साथ **कहानी 2** के स्तर से शुरू करें।*
- *अक्षर और शब्द को उत्तीर्ण करने के लिए हर भाग में कम से कम 4 जवाब सही होने चाहिए।*
- *यदि अक्षर और शब्द स्तर के दौरान छात्र 10 सेकेंड तक जवाब ना दें तो उन्हे जवाब देने को प्रोत्साहित करें। यदि इसके बाद भी वे और 10 सेकेंड के लिए जवाब ना दें तो अगले अक्षर या शब्द पे जाएँ।*
- *छात्र के द्वारा दिए गए सही जवाबों की संख्या को ODK app में भरें और सर्वेक्षक अपने साथ में ODK फॉर्म की मैनुअल प्रति जरूर रखें।*

A1

IDinsight
DATA. DECISIONS. DEVELOPMENT.

छात्र को निम्नलिखित अक्षर और शब्द पढ़ने को कहें। किसी भी अक्षर या शब्द को छात्र से पढ़वाने के लिए उस अक्षर या शब्द के नीचे अपनी उंगली रखकर तब छात्र को पढ़ने को कहें। अगर पढ़ने के दौरान छात्र कोई अक्षर या शब्द गलत पढ़ता है तो छात्र को वो अक्षर या शब्द दोबारा पढ़ने के लिए कहें लेकिन तीसरा मौक़ा ना दें।

Q1: अक्षर		
ग		य
	ड	
		ज्ञ
छ		

A2

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Q2: शब्द
दूध हरा चोर काफी शेर

A3

IDinsight
DATA. DECISIONS. DEVELOPMENT.

छात्र से निम्नलिखित अनुच्छेद पढ़ने को कहें और ध्यान दे की बच्चा अनुच्छेद पढ़ सकता है की नहीं। खास ध्यान दें की छात्र अनुच्छेद को एक प्रवाह में पढ़ें न कि एक एक शब्द अलग से पढ़ें। अगर वह कोई शब्द या वाक्य सही से नहीं पढ़ पाता है तो उससे दोबारा पढ़ने को कहें। छात्र को अनुच्छेद पढ़ते समय अधिकतम 3 गलतियाँ माफ़ हैं। किसी भी एक शब्द में की गयी सभी गलतियों को केवल एक ही गलती गिनें। अनुच्छेद में एक ही शब्द में बार बार ग़लती को एक ही माना जाएगा ।3 से ज्यादा गलतियाँ करने पर छात्र से अगला भाग ना करवाएं। अगर छात्र पढ़ते हुए 10 सेकेंड से ज़्यादा तक कुछ नही बोल पाते या अनुच्छेद का प्रयास नही कर सकते हैं तो उन्हे बढ़ावा दें। यदि तब भी वे अगले 10 सेकेंड में ना कर पाए तो उनका टेस्ट यहीं ख़तम करें।

Q3: अनुच्छेद
सुबह हो गयी है । **सूरज निकल आया है ।** **चिड़ियाँ चहक रही हैं ।** **सब लोग उठ गये हैं ।**

A4

IDinsight
DATA. DECISIONS. DEVELOPMENT.

*छात्र से निम्नलिखित कहानियाँ स्वयं पढ़ने को कहें। जैसे ये दिए गए हैं उस ही क्रम में छात्रों से इसे पढ़ने को कहे। पहले कहानी एक करवाया जाना चाहिए। बच्चे को यह भी बता दें कि कहानी पढ़ने के बाद उसे कहानी के आधारित प्रश्नो के उत्तर देने हैं। खास ध्यान दें की छात्र कहानी को एक प्रवाह में पढ़ें ना की एक एक शब्द अलग से पढ़ें। अगर वह कोई शब्द या वाक्य सही से नहीं पढ़ पाता है तो उससे दोबारा पढ़ने को कहें। छात्र को कहानी पढ़ते समय अधिकतम 3 गलतियाँ माफ़ हैं। किसी भी एक शब्द में की गयी सभी गलतियों को केवल एक ही गलती गिनें। एक कहानी में एक ही शब्द में बार बार ग़लती को एक ही माना जाएगा। अगर वे कुल 3 से ज्यादा गलतियाँ करें तो भी छात्र को भाग के अंत में दिए प्रश्न पूछें। बच्चे से प्रश्न के उत्तर मौखिक रूप में **पूरे वाक्य** में देने को कहें। यदि छात्र(Q4) कहानी एक में तीन से ज़्यादा ग़लतियाँ करें तो उनका टेस्ट वहीं रोक देना चाहिए। यदि छात्र कहानी एक में तीन या तीन से कम ग़लती करें तभी उन्हे (कहानी दो) दिया जाना चाहिए। । ऐसा ही कहानी 2 और कहानी 3 में भी होगा। अगर छात्र पढ़ते हुए 10 सेकेंड से ज़्यादा तक कुछ नही बोल पाते हैं या कहानी का प्रयास नही कर सकते हैं तो उन्हे बढ़ावा दें। यदि तब भी वे अगले 10 सेकेंड में ना कर पाए तो उनका टेस्ट यहीं ख़तम करें। **सिर्फ़ इस स्तिति में** छात्र को सवाल नही दिए जाएँगे और ODK में इसका जवाब अधूरा (incomplete) दर्ज करें। ।*

A5

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Q4: कहानी-1
रविवार का दिन था। सुबह हो चुकी थी। लोग अपने–अपने काम पर जा रहे थे। सुरेश पिताजी के साथ खेत में जा रहा था। मीना भी सुरेश के साथ खेत जाने के लिए ज़िद करने लगी। सुरेश ने मीना को भी साथ में ले लिया। खेत में दोनों ने खूब खेला तथा गन्ना खाया। शाम को पिताजी के साथ दोनों घर आ गए और पढ़ाई करने लगे। 1) सुरेश कहाँ जा रहा था ? 2) सुरेश और मीना ने शाम को क्या किया?

Source: ASER Hindi reading tool

A6

IDinsight
DATA. DECISIONS. DEVELOPMENT.

नोट: नीचे दी गई कहानी और प्रश्न उन्ही बच्चों को दे जो Q4. में दी गई कहानी को धाराप्रवाह और 3 या 3 से कम ग़लतियों के साथ पढ़ पाए थे।

Q5: कहानी-2

वर्षों पहले राम नगर के पास जहाँगीर पुरी में एक राजा राज्य करता था। राजा बहुत ही दयावान था। उसके राज्य में सभी लोग बहुत सुखीपूर्वक मिल–जुल कर रहते थे। अपनी प्रजा के बारे में जानने के लिए राजा भेष बदल कर राज्य में रात में विचरण करता था।

एक दिन राजा ने रात में एक घर में कुछ अजनबी लोगों को जाते देखा। राजा ने छिपकर उनका पीछा किया तथा उनकी बातें सुनने लगा। वो लोग किसी दूसरे राज्य से आये थे और जहाँगीर पुरी पर हमला करने की योजना बना रहे थे। राजा ने चतुराई से काम लेते हुए अपने सैनिकों को बुलाकर तुरंत उन लोगों को कैद करवा लिया। अगले दिन अपने दरबार में पूरी प्रजा के सामने राजा ने उनको सजा दी।

1) राजा किस प्रदेश में राज करता था?
2) राजा अपनी प्रजा के बारे में जानने के लिए क्या करता था?

Source: ASER Hindi Reading Tool

नोट: नीचे दी गई कहानी और प्रश्न उन्ही बच्चों को दे जो Q5. में दी गई कहानी को धाराप्रवाह और 3 या 3 से कम ग़लतियों के साथ पढ़ पाए थे।

A7

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Q6: कहानी-3

एक बार महात्मा गाँधी अपने कुछ अनुयायियों के साथ भ्रमण कर रहे थे। अचानक एक व्यक्ति उनके निकट से गुजरा और अपशब्द कहने लगा। गांधी जी मुस्कुराते हुए आगे बढ़ गए। गांधी जी को मुस्कुराते देख कर वह व्यक्ति और ज्यादा भला बुरा कहने लगा। गांधी जी फिर भी मुस्कुराते रहे और चुप–चाप आगे बढ़ते रहे।

गांधी जी का एक साथी, जिनका नाम मुकुन्द था, ने गांधी जी से पूछा कि वह व्यक्ति आपको इतना भला बुरा कह रहा था और आपने उसे कुछ भी नहीं कहा ? गांधी जी हँसते हुए बोले कि आप यहीं ठहरिये मै आपकी जिज्ञासा का समाधान थोड़ी देर में करता हूँ। गांधी जी अपने घर के अन्दर गए और जब वापस आये तो उनके हाथ में कुछ मैले वस्त्र थे। गांधी जी ने उन वस्त्रों को मुकुन्द जी को पहनने के लिए कहा। मुकुन्द जी ने उन कपड़ों की तरफ देखा और कहा कि ये तो बहुत मैले हैं इन्हें मैं नहीं पहन सकता हूँ। गांधी जी ने कहा कि जिस प्रकार तुम मैले वस्त्र धारण नहीं कर सकते हो ठीक उसी प्रकार मैं किसी के द्वारा कहे गए मैले शब्दों को धारण नहीं करता हूँ। मैं मैले शब्दों की तरफ ध्यान ही नहीं देता हूँ इसलिए मुझ पर उनका कोई प्रभाव नहीं पड़ता है|

1) जब गाँधी जी को अपशब्द कहे गये तो उन्होंने क्या किया?
2) गाँधी जी ने अपशब्दों की तुलना किस चीज़ से की?

A8

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Q7: कहानी-4
प्रत्येक वर्ष की भांति इस वर्ष भी विद्यालय में वार्षिक उत्सव की तैयारी जोर शोर से जारी है। इस कार्यक्रम में होने वाले नाटक में पदमिनी को जादूगरी का किरदार निभाना है। अपने प्रदर्शन को लेकर पदमिनी अत्यंत उत्साहित है। पदमिनी ने चेन्नई वाले मौसा जी से जादू की बारीकियां सीखी हैं। चेन्नई वाले मौसा जी दक्षिण भारत के बहुत ही मशहूर जादूगर हैं। नाटक के अलावा इस कार्यक्रम में, विज्ञान प्रश्नोत्तरी, सामूहिक वाद-विवाद, गीत प्रस्तुति, रंगोली तथा खेल प्रतियोगिता का भी आयोजन किया जाता है जिसमे सभी कक्षा के विद्यार्थियों को भाग लेने के लिए प्रोत्साहित किया जाता है। खेल प्रतियोगिता में मुख्यरूप से कबड्डी, क्रिकेट और बैडमिंटन का खेल करवाया जाता है। पदमिनी के छोटे भाई का नाम उत्कर्ष है और गत वर्ष बैडमिंटन प्रतियोगिता उत्कर्ष ने जीता था। उत्कर्ष इस बार भी जीतने के लिए कठिन अभ्यास कर रहा है। इस कार्यक्रम में भाग लेने वाले सभी छात्रों को पुरस्कृत किया जाता है। कार्यक्रम को सफल बनाने के लिए प्रधानाचार्य तथा अन्य सभी अध्यापकगण कठिन परिश्रम करते हैं। अपनी इस विशेषता के लिए यह विद्यालय अपने क्षेत्र में विख्यात है।

A9

IDinsight
DATA. DECISIONS. DEVELOPMENT.

1) पदमिनी ने जादूगरी किससे सीखी?
2) खेलों की **प्रतियोगिता** में कौनसे तीन खेल खेले जाते हैं?

A10

IDinsight
DATA. DECISIONS. DEVELOPMENT.

गणित

सर्वेक्षक ध्यान दें:

- *कक्षा 1 के छात्रों के साथ L1 के स्तर से शुरू करें।*
- *कक्षा 2 के छात्रों के साथ L2 के स्तर से शुरू करें।*
- *कक्षा 3 के छात्रों के साथ **जोड़** के स्तर से शुरू करें।*
- *कक्षा 4 के छात्रों के साथ **घटाने** के स्तर से शुरू करें।*
- *कक्षा 5 के छात्रों के साथ **गुणा** के स्तर से शुरू करें।*
- *L1 और L2 में छात्र के द्वारा दिए गए सही जवाबों की संख्या को ODK app में भरें और सर्वेक्षक अपने साथ में ODK फॉर्म की मैनुअल प्रति जरूर रखें। बाकी स्तरों में छात्र के द्वारा दिया गया जवाब ODK में भरे. ODK खुद ही (जवाब सही या ग़लत होने पर निर्धारित) अगला सवाल पेश करेगा।*
- *कोई भी छात्र अगर एक भाग उत्तीर्ण कर पाता है तभी अगले भाग पर जाएँ, अन्यथा छात्र का टेस्ट समाप्त कर दें और उसे धन्यवाद देकर जाने को कहें।*
- *L1 और L2 को उत्तीर्ण करने के लिए छात्र के कम से कम 4 जवाब सही होने चाहिए।*
- *अगर छात्र किसी भी सवाल में 10 सेकेंड से ज़्यादा तक कुछ नही बोल पाते हैं या सवाल का प्रयास नही कर सकते हैं तो उन्हे बढ़ावा दें। यदि तब भी वे अगले 10 सेकेंड में ना कर पाए तो उनको अगला सवाल पेश करें। यदि यह सवाल खंड का दूसरा सवाल है और उन्होने पहला सवाल ग़लत किया है या उसका जवाब नही दिया है तो उनका टेस्ट समाप्त करें।*

A11

IDinsight
DATA. DECISIONS. DEVELOPMENT.

छात्र को निम्नलिखित अंक पढ़ने को कहें/ किसी भी अंक या संख्या को छात्र से पढ़वाने के लिए उस अंक या संख्या के नीचे अपनी उंगली रखकर तब छात्र को पढ़ने को कहें/ अगर पढ़ने के दौरान छात्र कोई अंक या संख्या गलत पढ़ता है तो छात्र को वो अंक या संख्या दोबारा पढ़ने के लिए कहें लेकिन तीसरा मौका ना दें/

L1 (1-9)
2
9
8
6
4

A12

IDinsight
DATA. DECISIONS. DEVELOPMENT.

L2 (11-99)
21 33 67 43 98

A13

IDinsight
DATA. DECISIONS. DEVELOPMENT.

भाग 1 : जोड़

बच्चे को जोड़ के दोनों सवालों में से कोई भी एक सवाल लगाने को कहें| अगर वह यह सवाल सही से कर लेता है तो बच्चे को घटाना वाला भाग को करने को कहें| अगर बच्चा यह जोड़ गलत करता है तो उसे दूसरा वाला जोड़ करने को कहें| बच्चा अगर यह जोड़ भी गलत करता है तो उसे गणित में L2 स्तर का छात्र मानें लेकिन, बच्चा यदि दूसरा जोड़ सही कर लेता है तो उसे घटाना करने को कहें|

48 + 37 ____ ____	73 + 18 ____ ____

93 - 35 ____ ____	42 - 37 ____ ____

A14

IDinsight
DATA. DECISIONS. DEVELOPMENT.

भाग 2 : घटाना

*बच्चे को घटाने के दोनों सवालों में से कोई भी एक सवाल लगाने को कहें। अगर वह यह सवाल सही से कर लेता है तो बच्चे को घटाने के स्तर का छात्र मानें। अगर बच्चा यह सवाल गलत करता है तो उसे दूसरा वाला सवाल करने को कहें। बच्चा अगर यह सवाल भी गलत करता है तो उसे गणित में जोड़ के स्तर का छात्र मानें लेकिन, बच्चा यदि दूसरा सवाल सही कर लेता है तो उसे घटाना स्तर का छात्र ही मानें। यदि बच्चा कोई भी एक घटाने का सवाल सही करता है तो ही उसे **गुणा** का सवाल दे*

भाग 3 : गुणा

*बच्चे को **गुणा** के दोनों सवालों में से कोई भी एक सवाल लगाने को कहें। अगर वह यह सवाल सही से कर लेता है तो बच्चे को **गुणा** के स्तर का छात्र मानें। अगर बच्चा यह सवाल गलत करता है तो उसे दूसरा वाला सवाल करने को कहें। बच्चा अगर यह सवाल भी गलत करता है तो उसे गणित में घटाना के स्तर का छात्र मानें लेकिन, बच्चा यदि दूसरा सवाल सही कर लेता है तो उसे **गुणा** स्तर का छात्र ही मानें। यदि बच्चा कोई भी एक गुणक का सवाल सही करता है तो ही उसे भाग का सवाल दें।*

39 × 55	46 × 18

IDinsight
DATA. DECISIONS. DEVELOPMENT.

भाग 4 : भाग

बच्चे को भाग के दोनों सवालों में से कोई भी एक सवाल लगाने को कहें। अगर वह यह सवाल सही से कर लेता है तो बच्चे को भाग के स्तर का छात्र मानें। अगर बच्चा यह सवाल गलत करता है तो उसे दूसरा वाला सवाल करने को कहें। बच्चा अगर यह सवाल भी गलत करता है तो उसे गणित में गुणक के स्तर का छात्र मानें लेकिन, बच्चा यदि दूसरा सवाल सही कर लेता है तो उसे भाग स्तर का छात्र ही मानें। यदि बच्चा कोई भी एक भाग का सवाल सही करता है तो ही उसे भिन्न का सवाल दें।

7) 825 (	8) 637 (

भाग 5 : *भिन्न*

बच्चे को भिन्न के दोनों सवालों में से कोई भी एक सवाल लगाने को कहें। अगर वह यह सवाल सही से कर लेता है तो बच्चे को भिन्न के स्तर का छात्र मानें। अगर बच्चा यह सवाल गलत करता है तो उसे दूसरा वाला सवाल करने को कहें। बच्चा अगर यह सवाल भी गलत करता है तो उसे गणित में भाग के स्तर का छात्र मानें लेकिन, बच्चा यदि दूसरा सवाल सही कर लेता है तो उसे भिन्न स्तर का छात्र ही मानें।

$\frac{7}{9} \times \frac{16}{12}$	$\frac{3}{5} \times \frac{8}{6}$

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Student Testing Tool

STIR Impact Evaluation

Sample B

27th January 2017

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Oral Tool - IDinsight testing – Sample B **हिंदी**

सर्वेक्षक ध्यान दें:

- *कक्षा 1 के छात्रों के साथ अक्षर के स्तर से शुरू करें।*
- *कक्षा 2 के छात्रों के साथ **शब्द** के स्तर से शुरू करें।*
- *कक्षा 3 के छात्रों के साथ **अनुच्छेद** के स्तर से शुरू करें।*
- *कक्षा 4 के छात्रों के साथ **कहानी 1** के स्तर से शुरू करें।*
- *कक्षा 5 के छात्रों के साथ **कहानी 2** के स्तर से शुरू करें।*
- *अक्षर और शब्द को उत्तीर्ण करने के लिए हर भाग में कम से कम 4 जवाब सही होने चाहिए।*
- *यदि अक्षर और शब्द स्तर के दौरान छात्र 10 सेकंड तक जवाब ना दें तो उन्हे जवाब देने को प्रोत्साहित करें। यदि इसके बाद भी वे और 10 सेकंड के लिए जवाब ना दें तो अगले अक्षर या शब्द पे जाएँ।*
- *छात्र के द्वारा दिए गए सही जवाबों की संख्या को ODK app में भरें और सर्वेक्षक अपने साथ में ODK फॉर्म की मैनुअल प्रति जरुर रखें।*

B1

IDinsight
DATA. DECISIONS. DEVELOPMENT.

छात्र को निम्नलिखित अक्षर और शब्द पढ़ने को कहें| किसी भी अक्षर या शब्द को छात्र से पढ़वाने के लिए उस अक्षर या शब्द के नीचे अपनी उंगली रखकर तब छात्र को पढ़ने को कहें| अगर पढ़ने के दौरान छात्र कोई अक्षर या शब्द गलत पढ़ता है तो छात्र को वो अक्षर या शब्द दोबारा पढ़ने के लिए कहें लेकिन तीसरा मौक़ा ना दें|

Q1: अक्षर

फ

ब

र

ध

म

B2

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Q2: शब्द

चोरी

पीला

राजा

घड़ी

भाषा

B3

IDinsight
DATA. DECISIONS. DEVELOPMENT.

छात्र से निम्नलिखित अनुच्छेद पढ़ने को कहें और ध्यान दे की बच्चा अनुच्छेद पढ़ सकता है की नहीं| खास ध्यान दें की छात्र अनुच्छेद को एक प्रवाह में पढ़ें ना की एक एक शब्द अलग से पढ़ें| अगर वह कोई शब्द या वाक्य सही से नहीं पढ़ पाता है तो उससे दोबारा पढ़ने को कहें| छात्र को अनुच्छेद पढ़ते समय अधिकतम 3 गलतियाँ माफ़ हैं| किसी भी एक शब्द में की गयी सभी गलतियों को केवल एक ही गलती गिनें| अनुच्छेद में एक ही शब्द में बार बार ग़लती को एक ही माना जाएगा |3 से ज़्यादा गलतियाँ करने पर छात्र से अगला भाग ना करवाएं| अगर छात्र पढ़ते हुए 10 सेकंड से ज़्यादा तक कुछ नही बोल पाते या अनुच्छेद का प्रयास नही कर सकते हैं तो उन्हे बढ़ावा दें| यदि तब भी वे अगले 10 सेकंड में ना कर पाए तो उनका टेस्ट यहीं ख़तम करें|

Q3: अनुच्छेद
राम सो रहा था | **कुछ देर में सुबह हो गयी |** **फिर राम उठ गया |** **उठकर वह स्कूल चला गया |**

B4

IDinsight
DATA. DECISIONS. DEVELOPMENT.

*छात्र से निम्नलिखित कहानियाँ स्वयं पढ़ने को कहें| जैसे ये दिए गए हैं उस ही क्रम में छात्रों से इसे पढ़ने को कहे| पहले कहानी एक करवाया जाना चाहिए| बच्चे को यह भी बता दें कि कहानी पढ़ने के बाद उसे कहानी के आधारित प्रश्नो के उत्तर देने हैं| खास ध्यान दें की छात्र कहानी को एक प्रवाह में पढ़ें ना की एक एक शब्द अलग से पढ़ें | अगर वह कोई शब्द या वाक्य सही से नहीं पढ़ पाता है तो उससे दोबारा पढ़ने को कहें| छात्र को कहानी पढ़ते समय अधिकतम 3 गलतियाँ माफ़ हैं| किसी भी एक शब्द में की गयी सभी गलतियों को केवल एक ही गलती गिनें| एक कहानी में एक ही शब्द में बार बार ग़लती को एक ही माना जाएगा| अगर वे कुल 3 से ज़्यादा गलतियाँ करें तो भी छात्र को भाग के अंत में दिए प्रश्न पूछें | बच्चे से प्रश्न के उत्तर मौखिक रूप में **पूरे वाक्य** में देने को कहें| यदि छात्र(Q4) कहानी एक में तीन से ज़्यादा ग़लतियाँ करें तो उनका टेस्ट वहीं रोक देना चाहिए| यदि छात्र कहानी एक में तीन या तीन से कम ग़लती करें तभी उन्हे (कहानी दो) दिया जाना चाहिए| | ऐसा ही कहानी 2 और कहानी 3 में भी होगा| अगर छात्र पढ़ते हुए 10 सेकंड से ज़्यादा तक कुछ नही बोल पाते हैं या कहानी का प्रयास नही कर सकते हैं तो उन्हे बढ़ावा दें| यदि तब भी वे अगले 10 सेकंड में ना कर पाए तो उनका टेस्ट यहीं ख़तम करें| **सिर्फ़ इस स्तिति में** छात्र को सवाल नही दिए जाएँगे और ODK में इसका जवाब अधूरा (incomplete) दर्ज करें| |*

B5

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Q4: कहानी-1
एक बहुत बड़ा जंगल था	जंगल में बहुत से जानवर रहते थे	जंगल का राजा भोलू नाम का शेर था	भोलू जंगल के सभी जानवरों की देखभाल किया करता था	एक दिन भोलू दलदल में फंस गया	भोलू को बाहर निकालने के लिए सभी जानवर दलदल के पास पहुँच गए	हाथी दादा ने अपनी सूंड से भोलू को बाहर निकाल दिया	सभी जानवर बहुत खुश थे | 1) जंगल का राजा कौन था? 2) भोलू को दलदल से बाहर किसने निकाला?

Source: ASER Hindi reading tool

B6

IDinsight
DATA. DECISIONS. DEVELOPMENT.

नोट: नीचे दी गई कहानी और प्रश्न उन्ही बच्चों को दे जो Q4. में दी गई कहानी को धाराप्रवाह और 3 या 3 से कम ग़लतियों के साथ पढ़ पाए |

Q5: कहानी-2
महेश के गाँव में प्रत्येक वर्ष एकादशी के दिन बहुत बड़ा मेला लगता है	मेला देखने के लिए बहुत दूर-दूर से लोग आते हैं	मेले में बहुत भीड़ होती है और बहुत सी दुकाने होती हैं	पिछली बार दीपू और माला भी अपने चाचा के साथ मेला देखने गए थे	मेले में माला ने अपने चाचा के साथ झूले का मज़ा लिया था	दीपू को झूले से डर लगता था इसलिए वो झूले पर नहीं बैठा था	मेले में मिठाई की बहुत सी दुकानें थीं	दोनों ने खूब सारी मिठाईयां खाई	महेश ने दीपू और माला को पूरा मेला घुमाकर दिखाया	इस बार दीपू और माला के साथ उनके स्कूल के कई दोस्त भी मेला देखने की योजना बना रहे हैं | 1) महेश के गाँव में किस दिन मेला लगता था? 2) दिपू झूले पर क्यों नही बैठा?

Source: ASER Hindi Reading Tool

B7

IDinsight
DATA. DECISIONS. DEVELOPMENT.

नोट: नीचे दी गई कहानी और प्रश्न उन्ही बच्चों को दे जो Q5. में दी गई कहानी को धाराप्रवाह और 3 या 3 से कम ग़लतियों के साथ पढ़ पाए |

Q6: कहानी-3
एक दिन चिंटू सियार भोजन की तलाश करते-करते रास्ता भटक गया और शहर पहुँच गया	शहर के कुत्तों से जान बचाकर भागते हुए चिंटू एक कुम्हार के घर में घुसा और फिसलकर नील वाले बर्तन में गिर गया	चिंटू के शरीर का रंग नीला हो गया था और वो विचित्र दिखने लगा था	बड़ी मुश्किल से जान बचाकर चिंटू रातोंरात जंगल में वापस आया | सुबह विचित्र जानवर को देखकर जंगल में अफरा-तफरी मच गई और जानवर भयभीत होकर इधर - उधर भागने लगे	चिंटू का खुरापाती दिमाग तुरंत सक्रिय हो उठा	उसने सभी को एकत्रित किया और कहा कि आज से वह इस जंगल का राजा है और उसे जंगल की देवी ने भेजा है	सभी जानवरों ने भय के कारण उसे अपना राजा मान लिया	एक दिन जंगल के सभी सियार इकट्ठा होकर गाना गा रहे थे	गाना सुनते-सुनते चिंटू भी मगन होकर उनके साथ गाना गाने लगा और भूल गया कि लोग उसे उसकी आवाज से पहचान जायेंगे	जब जंगल के जानवरों ने चिंटू को गाते हुए सुना तो उन्होंने तुरंत पहचान लिया कि ये चिंटू है और इसकी सूचना जंगल के असली राजा शेर को दी	शेर ने चिंटू से खूब पूछताछ की तब चिंटू ने पूरी कहानी बताई और माफ़ी मांगी कि वो

IDinsight
DATA. DECISIONS. DEVELOPMENT.

दुबारा ऐसी गलती नहीं करेगा |

1) चिंटू शहर कैसे पहुँचा?
2) चिंटू ने ऐसा क्या किया जिससे उसकी असलियत बाकी जानवरों के सामने आ गयी?

IDinsight
DATA. DECISIONS. DEVELOPMENT.

नोट: नीचे दी गई कहानी और प्रश्न उन्ही बच्चों को दे जो Q6. में दी गई कहानी को धाराप्रवाह और 3 या 3 से कम ग़लतियों के साथ पढ़ पाए|

Q7: कहानी-4
प्रत्येक वर्ष की भांति इस वर्ष भी विद्यालय में वार्षिक उत्सव की तैयारी जोर शोर से जारी है	इस कार्यक्रम में होने वाले नाटक में पदमिनी को जादूगरी का किरदार निभाना है	अपने प्रदर्शन को लेकर पदमिनी अत्यंत उत्साहित है	पदमिनी ने चेन्नई वाले मौसा जी से जादू की बारीकियां सीखी हैं	चेन्नई वाले मौसा जी दक्षिण भारत के बहुत ही मशहूर जादूगर हैं	नाटक के अलावा इस कार्यक्रम में, विज्ञान प्रश्नोतरी, सामूहिक वाद-विवाद, गीत प्रस्तुति, रंगोली तथा खेल प्रतियोगिता का भी आयोजन किया जाता है जिसमे सभी कक्षा के विद्यार्थियों को भाग लेने के लिए प्रोत्साहित किया जाता है	खेल प्रतियोगिता में मुख्यरूप से कबड्डी, क्रिकेट और बैडमिंटन का खेल करवाया जाता है	पदमिनी के छोटे भाई का नाम उत्कर्ष है और गत वर्ष बैडमिंटन प्रतियोगिता उत्कर्ष ने जीता था	उत्कर्ष इस बार भी जीतने के लिए कठिन अभ्यास कर रहा है	इस कार्यक्रम में भाग लेने वाले सभी छात्रों को पुरस्कृत किया जाता है	कार्यक्रम को सफल बनाने के लिए प्रधानाचार्य तथा अन्य सभी अध्यापकगण कठिन परिश्रम करते हैं	अपनी इस विशेषता के लिए यह विद्यालय अपने क्षेत्र में विख्यात है	

B10

IDinsight
DATA. DECISIONS. DEVELOPMENT.

1) पदमिनी ने जादूगरी किससे सीखी?

2) खेलों की **प्रतियोगिता** में कौनसे तीन खेल खेले जाते हैं?

B11

IDinsight
DATA. DECISIONS. DEVELOPMENT.

IDinsight
DATA. DECISIONS. DEVELOPMENT.

गणित

सर्वेक्षक ध्यान दें:

- *कक्षा 1 के छात्रों के साथ **L1** के स्तर से शुरू करें।*
- *कक्षा 2 के छात्रों के साथ **L2** के स्तर से शुरू करें।*
- *कक्षा 3 के छात्रों के साथ **जोड़** के स्तर से शुरू करें।*
- *कक्षा 4 के छात्रों के साथ **घटाने** के स्तर से शुरू करें।*
- *कक्षा 5 के छात्रों के साथ **गुणा** के स्तर से शुरू करें।*
- *L1 और L2 में छात्र के द्वारा दिए गए सही जवाबों की संख्या को ODK app में भरें और सर्वेक्षक अपने साथ में ODK फॉर्म की मैनुअल प्रति जरूर रखें। बाकी स्तरों में छात्र के द्वारा दिया गया जवाब ODK में भरे. ODK खुद ही (जवाब सही या ग़लत होने पर निर्धारित) अगला सवाल पेश करेगा।*
- *कोई भी छात्र अगर एक भाग उत्तीर्ण कर पाता है तभी अगले भाग पर जाएँ, अन्यथा छात्र का टेस्ट समाप्त कर दें और उसे धन्यवाद देकर जाने को कहें।*
- *L1 और L2 को उत्तीर्ण करने के लिए छात्र के कम से कम 4 जवाब सही होने चाहिए।*
- *अगर छात्र किसी भी सवाल में 10 सेकेंड से ज़्यादा तक कुछ नही बोल पाते हैं या सवाल का प्रयास नही कर सकते हैं तो उन्हे बढ़ावा दें। यदि तब भी वे अगले 10 सेकेंड में ना कर पाए तो उनको अगला सवाल पेश करें। यदि यह सवाल खंड का दूसरा सवाल है और उन्होने पहला सवाल ग़लत किया है या उसका जवाब नही दिया है तो उनका टेस्ट समाप्त करें।*

B12

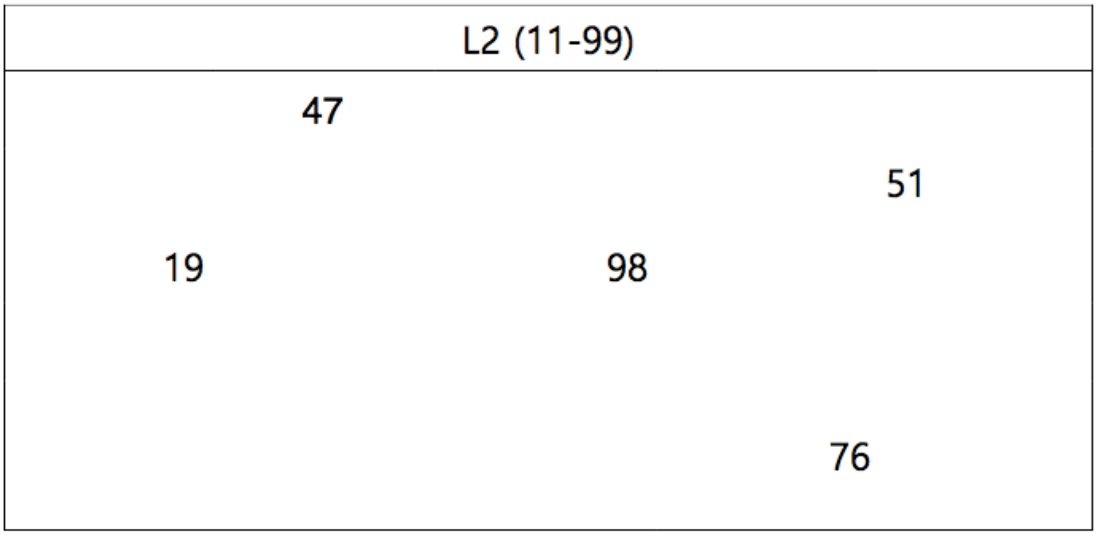
IDinsight
DATA. DECISIONS. DEVELOPMENT.

छात्र को निम्नलिखित अंक पढ़ने को कहें/ किसी भी अंक या संख्या को छात्र से पढ़वाने के लिए उस अंक या संख्या के नीचे अपनी उंगली रखकर तब छात्र को पढ़ने को कहें/ अगर पढ़ने के दौरान छात्र कोई अंक या संख्या गलत पढ़ता है तो छात्र को वो अंक या संख्या दोबारा पढ़ने के लिए कहें लेकिन तीसरा मौका ना दें/

L1 (1-9)
1 5 3 7 6

B13

IDinsight
DATA. DECISIONS. DEVELOPMENT.



B14

IDinsight
DATA. DECISIONS. DEVELOPMENT.

भाग 1 : जोड़

बच्चे को जोड़ के दोनों सवालों में से कोई भी एक सवाल लगाने को कहें अगर वह यह सवाल सही से कर लेता है तो बच्चे को घटाना वाला भाग को करने को कहें अगर बच्चा यह जोड़ गलत करता है तो उसे दूसरा वाला जोड़ करने को कहें बच्चा अगर यह जोड़ भी गलत करता है तो उसे गणित में L2 स्तर का छात्र मानें लेकिन, बच्चा यदि दूसरा जोड़ सही कर लेता है तो उसे घटाना करने को कहें

66 + 17 ____ ____	25 + 76 ____ ____

भाग 2 : घटाना

बच्चे को घटाने के दोनों सवालों में से कोई भी एक सवाल लगाने को कहें। अगर वह यह सवाल सही से कर लेता है तो बच्चे को घटाने के स्तर का छात्र मानें। अगर बच्चा यह सवाल गलत करता है तो उसे दूसरा वाला सवाल करने को कहें। बच्चा अगर यह सवाल भी गलत करता है तो उसे गणित में जोड़ के स्तर का छात्र मानें लेकिन, बच्चा यदि दूसरा सवाल सही कर लेता है तो उसे घटाना स्तर का छात्र ही मानें। यदि बच्चा कोई भी एक घटाने का सवाल सही करता है तो ही उसे गुणा का सवाल दें।

76 - 37 ____ ____	81 - 39 ____ ____

B15

IDinsight
DATA. DECISIONS. DEVELOPMENT.

भाग 3 : गुणा

*बच्चे को **गुणा** के दोनों सवालों में से कोई भी एक सवाल लगाने को कहें। अगर वह यह सवाल सही से कर लेता है तो बच्चे को **गुणा** के स्तर का छात्र मानें। अगर बच्चा यह सवाल गलत करता है तो उसे दूसरा वाला सवाल करने को कहें। बच्चा अगर यह सवाल भी गलत करता है तो उसे गणित में घटाना के स्तर का छात्र मानें लेकिन, बच्चा यदि दूसरा सवाल सही कर लेता है तो उसे **गुणा** स्तर का छात्र ही मानें। यदि बच्चा कोई भी एक गुणक का सवाल सही करता है तो ही उसे भाग का सवाल दें।*

35 × 39 = _____	47 × 59 = _____

भाग 4 : भाग

बच्चे को भाग के दोनों सवालों में से कोई भी एक सवाल लगाने को कहें। अगर वह यह सवाल सही से कर लेता है तो बच्चे को भाग के स्तर का छात्र मानें। अगर बच्चा यह सवाल गलत करता है तो उसे दूसरा वाला सवाल करने को कहें। बच्चा अगर यह सवाल भी गलत करता है तो उसे गणित में गुणक के स्तर का छात्र मानें लेकिन, बच्चा यदि दूसरा सवाल सही कर लेता है तो उसे भाग स्तर का छात्र ही मानें। यदि बच्चा कोई भी एक भाग का सवाल सही करता है तो ही उसे भिन्न का सवाल दें।

8) 738 (	6) 934 (

IDinsight
DATA. DECISIONS. DEVELOPMENT.

भाग 5 : *भिन्न*

बच्चे को भिन्न के दोनों सवालों में से कोई भी एक सवाल लगाने को कहें। अगर वह यह सवाल सही से कर लेता है तो बच्चे को भिन्न के स्तर का छात्र मानें। अगर बच्चा यह सवाल गलत करता है तो उसे दूसरा वाला सवाल करने को कहें। बच्चा अगर यह सवाल भी गलत करता है तो उसे गणित में भाग के स्तर का छात्र मानें लेकिन, बच्चा यदि दूसरा सवाल सही कर लेता है तो उसे भिन्न स्तर का छात्र ही मानें।

$\frac{5}{8} \times \frac{13}{14}$	$\frac{4}{6} \times \frac{8}{9}$

Appendix A11: Teacher motivation questionnaire (endline version)

Table A7: Translated items from the teacher motivation questionnaire

STATA Variable name	Statement/Question
A1	Teaching is mentally draining.
A2	With the help of my colleagues, we can solve student issues.
A3	I feel used up at the end of the school day.
A4	My pay as a teacher is insufficient to support my family
A5	I feel fatigued when I get up in the morning and have to face another day at school.
A6	I have the ability to get parents involved in their children's education.
A7	I ask my colleagues for feedback.
A8	With the help of my colleagues, we can identify innovative practices.
A9	As a teacher, I'm given more responsibilities than I can manage.
A10	Some teachers at my school want to transfer to another school.
A11	I do not get paid on time.
A12	I can make my classroom a safe space for students, both emotionally and physically.
A13	As a teacher, I am contributing positively to the lives of my students.
A14	I feel energized when my class greets me each morning.
A15	If I had to choose again, I would still want to be a teacher.
A16	My supervisors treat me with respect.
A17	My colleagues at school make work a fun place to be.
A18	My supervisor praises me for my efforts in the school.
A19	Parents value my work as a teacher.
A20	I plan lessons with a colleague.
A21	I feel confident about my abilities as a teacher.
A22	If a student does not remember information in a previous lesson, I would know how to help them remember.
A23	When a student gets a better graded than he or she usually gets, it is because I found a better way.
A24	If a student in my class is undisciplined, I know some techniques to direct him or her.
A25	Every teacher can continue to improve their practice throughout their career.
A26	I can get through to even the most difficult or unmotivated students.
A27	I can motivate students who show low interest in school.
A28	I can influence some of the decisions that are made in the school.
A29	I can get students to work in groups or pairs.
A30	I ask my supervisor for feedback.
A31	I can help students overcome some difficult home and community conditions.
A32	Teachers in my school work closely with supervisors.

A33	I spend too much time traveling to my school.
A34	My fellow teachers can be counted on to influence decisions of the school.
A35	When I get new material, I am sure I am able to learn it.
A36	My family is proud that I am a teacher.
A37	Sometimes I share materials with colleagues.
A38	My colleagues praise me for coming up with new ways to teach a lesson.
B1	Every teacher can significantly improve his or her teaching ability.
B2	No matter how much natural ability you may have, you can always find important ways to improve.
B3	When I set a goal, no matter how difficult, I will eventually achieve it.
B4	I can learn new things, but I cannot really change my basic intelligence.
B5	The kind of teacher someone is, is something very basic about them, and can't be changed very much.
B6	Some teachers don't really benefit from professional learning because they have a natural ability.
B7	Teachers can change the way they teach in the classroom, but they can't really change their true ability.
B8	Some teachers will be ineffective no matter how hard they try to improve.

Appendix A12: Baseline and endline sampling

Programmatic and evaluation timeline

Table A8 below presents timeline for how our evaluations relate to both STIR's programming and the academic year in India. STIR's network meetings happen roughly once a month but not in precisely the same week for all networks. The placement of network meetings in the timeline in Table A8 should be taken as indicative.

Table A8: Timeline of academic, programmatic, and evaluation events for first year of randomized evaluation (2015-2016)

	Academic year	STIR programming	Evaluation
February 2015	In-session		Baseline TM
	In-session		Baseline TM
March 2015	In-session		Baseline TM
	In-session		Baseline TM
April 2015	**Begin academic year 1**		Baseline TM
	In-session	Taster session	Baseline TM
May 2015	In-session		*[randomization & sampling]*
	Holiday		
June 2015	Holiday		Baseline CP & SL
	Holiday		Baseline CP & SL
July 2015	Holiday		Baseline CP & SL
	In-session		Baseline CP & SL
August 2015	In-session	Network meeting (Year 1) (Innovate)	Baseline CP & SL
	In-session		Baseline CP & SL
September 2015	In-session	Network meeting (Year 1) (Innovate)	Baseline CP & SL
	In-session		Baseline CP & SL
October 2015	In-session	Network meeting (Year 1) (Implement)	Baseline CP & SL
	In-session		Baseline CP & SL
November 2015	In-session	Network meeting (Year 1) (Implement)	Baseline CP & SL
	In-session		Baseline CP & SL
December 2015	In-session	Network meeting (Year 1) (implement)	
	Holiday		
January 2016	Holiday		
	In-session	Network meeting (Year 1) (Influence)	
February 2016	In-session		
	In-session	Network meeting (Year 1) (Influence)	*[process evaluation]*
March 2016	In-session		*[process evaluation]*
	In-session	Network meeting (Year 1) (Influence)	*[process evaluation]*
April 2016	**Begin academic year 2**		Midline TM
	In-session	Network meeting (Year 2)	Midline TM
May 2016	In-session		Midline TM
	Holiday		
June 2016	Holiday		

	Holiday		
July 2016	Holiday		
	In-session		Midline TA, CP, & SL
August 2016	In-session	Network meeting (Year 2)	Midline TA, CP, & SL
	In-session		Midline TA, CP, & SL
September 2016	In-session	Network meeting (Year 2)	Midline TA, CP, & SL
	In-session (exams)		
October 2016	In-session	Network meeting (Year 2)	
	In-session		
November 2016	In-session	Network meeting (Year 2)	
	In-session		
December 2016	In-session	Network meeting (Year 2)	
	Holiday		
January 2017	Holiday		
	In-session		Delhi Endline TM, CP, & SL
February 2017	In-session		Delhi Endline TM, CP, & SL
	In-session		Delhi Endline TM, CP, & SL
March 2017	In-session		
	In-session		
April 2017	In-session		
	In-session		
May 2017	In-session		
	Holiday		
June 2017	Holiday		
	Holiday		
July 2017	Holiday		
	In-session	Network meeting (Year 2-Refresher sessions[18])	U.P. Endline TM, CP, & SL
August 2017	In-session		U.P. Endline TM, CP, & SL
	In-session		U.P. Endline TM, CP, & SL

Delhi private schools: sampling strategies and baseline and endline samples

In this section, we describe baseline and endline sampling for our different data collection needs. The endline target and actual samples are summarized in Table A9.

Table A9: Delhi endline targeted and actual samples

	Teachers for motivation questionnaire	Teachers for classroom practice observation	Students for learning outcomes
Target endline sample	All teachers in control and treatment schools in our sample	All teachers from Delhi CP original list (n= 811). If a school has fewer than 2 teachers left from	All 3367 students surveyed at baseline

[18] STIR felt that since the U.P. data collection was delayed due to state elections, it would be beneficial to conduct another session with the teachers before the data collection in treatment schools so as to make up for the long lag in between the last meeting and the survey period.

		this list, randomly select one or two teachers randomly from among those teachers who were present as on 1st July 2016	
Total number of units sampled at endline	1072	462	1846
Characteristics of the Estimation Sample	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 endline.

4.5.3.1 Delhi teacher motivation assessment (using motivation questionnaire)

Delhi TM baseline

Sampling strategy

Our sampling strategy for using the teacher motivation tool was to capture all teachers in all our treatment and comparison schools. The sample frame in Delhi consists of 180 private schools located in East Delhi; these are the 180 schools identified by STIR as fitting their criterion for being a private schools (fees ≤ US$ 17/month) and that expressed interest in STIR's programming and agreed to host a taster session[19]. As shown in Table A8, teacher motivation baseline took place from February to April 2015, before the randomization of STIR's programming.

Baseline sample

After a maximum of three visits per school, 1,249 teachers completed the teacher motivation questionnaire. Hereafter this list of 1,249 teachers is referred to as **Delhi TM baseline list**. The first column of Table 15 in Appendix A14 illustrates that the average teacher in our baseline sample is about 29 years old and has an average of 5.8 years of teaching experience[20]. 94% of teachers in the Delhi TM baseline list are female; 25% possess a bachelor degree or higher.

Delhi TM endline

Sampling strategy

The endline teacher motivation survey took place at the end of the second academic year between January and February 2017. As shown in Table A9, all teachers in our sample

[19] STIR held a taster session before rolling out the actual program (and network meetings) to gauge interest of teachers of the school in the program and give teachers an idea of what they could expect from the sessions, portfolios and other elements of the program.

[20] Standard deviations are reported in parentheses.

schools formed the endline target list. From the teachers surveyed at baseline, 734 teachers dropped out of our sample during the midline and endline and hence were not available for surveying at endline[21].

Endline sample

We surveyed 514 teachers at endline for whom we also have baseline data (48% of the target sample).

Delhi classroom practice assessment (time use and child-friendliness)

Delhi CP baseline

Sampling strategy

Prior to randomization, STIR conducted a taster session in all 180 interested private schools. From these sessions, STIR identified 811 teachers interested in STIR's programming. This became the target list for observing classroom practice at baseline, hereafter the **Delhi CP original list**. *Note that this sampling strategy means that we only look at teachers, in both treatment and comparison schools, who expressed a degree of interest in STIR's programming.* As mentioned in section 3, not all teachers who expressed interest ended up joining STIR's program. However, to the extent that teachers who showed initial interest in the program differ from those who did not (albeit on unobservable characteristics), the results should be generalized with a bit of caution. As shown in Table A8, classroom observations took place between July and October 2015, following the summer break of the 2015-16 school year.

Baseline sample

While we targeted 811 teachers for baseline classroom observations, only 333 teachers (41% of the target) were ultimately observed. These form the **Delhi CP baseline list**.[22]

The lower number of classroom observations than planned can be explained by a high number of school refusals and teacher drop-outs. First, without an over-arching authority over all private schools, surveying permissions (and interest in STIR) were regularly renegotiated on a school-by-school basis. Some Head Teachers and owners were particularly skitterish about data collection, given the proprietary nature of their schools as well as concerns about government regulatory check-ins.[23]

---

[21] While we do not know the exact reason for dropouts for each of these teachers; our experience has been that teachers leave schools either to join other private schools or drop out of teaching altogether. As part of the process evaluation in early 2016, we tried following up telephonically with teachers who had dropped out of our sample between both rounds of baseline. Among the 50 teachers we were able to talk with, nearly 40% of teachers were no longer working, roughly 20% had moved onto teaching in another private school, the same proportion had moved onto teaching private tuitions and a small proportion (less than 5%) were now teaching in government schools.

[22] There are 9 classroom observations that were only partially conducted and thus not used in the analysis sample.

[23] Private schools of the type with which STIR works in Delhi have come under threat of government ordered closure due to inability to meet some of the quality and infrastructure standards of the Right to Education Act. See, for example, a 2015 news item from *The Hindu* (PTI 2015). This may cause some schools to be more hesitant in sharing information with external parties.

Column 1 of Table 16 in Appendix A14 presents descriptive statistics of the teachers for which classroom observations have been conducted at baseline.

Delhi CP endline

Sampling strategy

Endline classroom observations took place from Jan to February 2017, after the winter break of the 2016-17 school year (see Table A8). The endline target was to conduct classroom observations for all 811 teachers on the **Delhi original list**, of which 333 teachers (41%) were observed at baseline.

In schools at which fewer than 2 teachers could be found from the **Delhi original list**, we would randomly sample additional teachers from that school, if they been present as of 1 July 2015. The number sampled per school depended on the number that had dropped out and the number in the **Delhi original list**. Say a school had 5 teachers and all but one dropped out, we would add four teachers. If a school had say only 1 in the **Delhi original list** and they dropped out we replaced only one.

Endline sample

Our total endline sample for the classroom practice is 462. Of the **Delhi CP original list** we followed up with 221, and these are the teachers we have baseline data for.

241 teachers were added since baseline during the following two rounds whenever the threshold of minimum two teachers per school was not met.

Data collection was hampered by objections of school owners and Head Teachers. We faced two main types of challenges during data collection:

- **School refusals:** In Delhi, STIR Education Leaders managed our school entry. In a few schools, school leadership consistently put off scheduling a precise date for surveying, both because of a lack of buy-in to the STIR program and suspicion around data collection more generally. Further, a few schools refused to have us survey altogether despite providing dates for survey. This impacted both classroom practice observation and student testing.
- **Refusals to add new teachers to our lists:** In some schools, Head Teachers allowed us to speak with teachers interviewed at baseline but did not allow us to add new teachers to our sample due to apprehension with data collection and disinterest in subsequent rounds of data collection. This affected only classroom practice observations.

Delhi student learning assessment (Hindi and math)

Delhi SL baseline

Sampling strategy

We sampled students at the same time as we conducted the classroom practice assessments, from July to October 2015 (see Table A8). Our sampling strategy was to link the students selected with the teachers for whom we observed classroom practice. We aimed to randomly

select 10 students from the 'main' class of each of the observed teachers.[24] The hypothetical sampling frame was thus 8110.

Baseline sample

As we only observed the classrooms of 342 teachers, we aimed to assess learning levels for 3420 students. Given student refusals[25], we ultimately tested 3367 students at baseline. Hereafter, this list of students is referred to as the **Delhi Student Learning baseline list**.

The average student in our baseline sample is 8 years old. 60% of students in our sample are male and 40 % are female.

Delhi SL endline

Sampling strategy

Student testing took place at the same time as classroom observations and teacher motivation surveys (January to February 2017). As with the teacher samples, our goal was to have a panel dataset, meaning that we would follow-up with the same students tested at midline. Our target sample was thus the 3367 students on the **Delhi Student Learning baseline list**. We did up to five revisits to schools in pursuit of these students.

Endline sample

We tracked and tested 1846 students at endline.

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

In this section, we describe baseline and endline sampling for our different data collection needs. The endline target and actual samples are summarized in Table A10.

Table A10: U.P. endline targeted and actual samples

	Teachers for motivation survey	Teachers for classroom observation	Students for testing learning levels
Target endline sample	All teachers in control and treatment schools in our sample	All 838 teachers surveyed at baseline. Wherever all teachers have dropped out one teacher to be added on the spot.	All 7386 students surveyed at baseline
Total number of units sampled at midline	1133	724[26]	3152

[24] 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 additional responsibility for administrative tasks such as taking attendance for a particular grade level or classroom.

[25] Refusals may not be directly by students but also by teachers or head-teachers 'on behalf.'

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

Characteristics of the Estimation Sample	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 present at midline

U.P teacher motivation assessment (using motivation questionnaire)

U.P. TM baseline

Sampling strategy

We conducted our baseline TM survey from February to April 2015, prior to STIR beginning to implement their programming (see Table A8). The sampling frame included all teachers in schools in the sampling frame: 270 government schools in Rae Bareli and Varanasi districts.[27] IDinsight enumerators visited all schools up to three times to reach all available teachers.

Baseline sample

1,145 teachers completed the motivation questionnaire. Hereafter, this list of 1145 teachers is referred to as **U.P. TM baseline list**. The average teacher is about 38.7 years old and has about 11.2 years of teaching experience. About 54% of teachers in our sample are female and more than 90% of teachers' education background is M. Phil or higher.

U.P. TM endline

Sampling strategy

The endline teacher motivation survey took place at the beginning of the third academic year between, July and August 2017. The goal was to survey as many teachers as possible by offering the survey to all the teachers in our sample schools.

Endline sample

582 teachers (51%) from the **U.P. TM baseline list** were surveyed successfully at endline.[28] 551 teachers were added on the spot by offering the form to all teachers. We do not have baseline data for these teachers.

U.P. classroom practice assessment (time-use and child-friendliness)

U.P. CP baseline

Sampling strategy

Classroom practice observations took place between July and October 2015, following the summer break of the 2015-16 school year. We aimed to randomly select an average of three

[27] Within these districts, IDinsight randomly selected 16 clusters, conditional on the cluster having least 15 schools per cluster. From these 16 clusters, all 270 schools in those schools were included in the study. We provide additional details about randomization in Appendix A5.

[28] Note that this contains the teachers that have been added to the list according to the procedure described above.

teachers from each of the 270 schools included in the sample, drawing from the 1,145 teachers on the U.P. TM baseline list.

Baseline sample

In total, classroom observations took place for 838 teachers. Hereafter, this list is referred to as **U.P. CP baseline list**. he average teacher is about 38.4 years old and has about 11 years of teaching experience. 54% of teachers in the baseline sample are female and about 84% have at least obtained an M.Phil.

U.P. CP endline

Sampling strategy

Endline classroom observations took place at the beginning of the third academic year between, July and August 2017. The goal was to conduct classroom observations for all teachers included in the **U.P. CP baseline list** (n=838). We included additional teachers in the survey in cases for which the number of teachers on the U.P CP baseline list still teaching in this school dropped to 0.[29,30] For adding teachers to the list, one teacher was selected at random from all the teachers teaching in the school as of 1st July 2016. A maximum of five revisits were done to ensure maximum opportunity to track the same teachers.

Endline sample

We successfully conducted endline classroom practice observations for 629 (86%) of the teachers from the **U.P. CP baseline list**. 95 new teachers were added on the spot since after baseline, resulting in a total of 724 teacher classroom observations.

U.P. student learning assessment (Hindi and math)

U.P. SL baseline

Sampling strategy

The student testing survey was conducted in parallel to the classroom observations (July-October 2015). For each of the teachers in the U.P CP baseline list, we aimed to assess a random sample of 10 students from the main course the teacher was teaching is selected for student testing, for a total of 8380 students.[31]

---

[29] Even though we knew from STIR that within schools some teachers had become more active in STIR than others, we did *not* make use of this information in selecting teachers within schools. Since our main estimate is ITT at the school level, in line with the final unit of assignment, we used simple random sampling within each school and so should capture, in aggregate, the average teacher offered STIR's program, including those who never joined and those who are the most enthusiastic participators.

[30] This sampling strategy has two advantages over other re-sampling methods:

1. This re-sampling method helps to target teachers and students that have been exposed to STIR for as long as possible. The supporting thought is that the timeline for the overall evaluation is relatively short and a longer exposure to STIR is expected to represent the effects of interest more accurately.

2. Given the rather sophisticated design of the evaluations, this re-sampling procedure allows for a clear interpretation and narrative surrounding the results.

[31] 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 additional responsibility for administrative tasks such as taking attendance for a particular grade level or classroom.

Baseline sample

At baseline, a total of 7386 students' math and Hindi proficiency was tested. Hereafter, this list of students is referred to as **U.P. SL baseline list**. In U.P., it is common for classes to have a total enrollment (or attendance across multiple days) of less than 10 and have two teachers teach the same group of students. Thus, even though the target was on average 10 students per teacher, in actuality, the total number of students surveyed per teacher often fell less than this number. The average student in our sample was 9 years old. Roughly 45 percent of students in our sample were male.

U.P. SL endline

Sampling strategy

Endline student learning was assessed jointly with classroom observations (July and August 2017). In pursuit of a panel dataset, we aimed to assess the 7386 students from the U.P. SL baseline list at endline. We initially planned to do up to three revisits per school but ultimately made between 4 and 6 revisits.

Endline sample

We completed student learning assessments with 3152 students (43% of the baseline sample)[32] in the endline survey.

[32] Again, as with Delhi, we have less verifiable information on why students drop out as compared to teacher level drop outs. Anecdotally, there seemed to be three main reasons: a) Students moving to private schools (especially after 5th grade) b) Students dropping out of school altogether and c) Students moving to UPSs of other villages.

Appendix A13: Attrition

The effects of attrition (Glennerster and Takavarasha, 2013) are summarized below to help guide the reader through the potential threat on any evaluation:

1. **Reduce comparability between treatment and comparison:** When people's data are lost through attrition, the comparability of the treatment and comparison group is undermined, if the rates of attrition or types of attrition differ between the treatment and comparison group. In both cases, we would end up with a biased estimate of impact.
2. **Attrition lowers statistical power:** Statistical power depends on sample size. Attrition reduces sample size and in turn reduces power. The experiment loses sensitivity, and the impact of the program must be higher in order to detect it.

Detailed understanding of attrition Assessing differential sample loss (attrition)

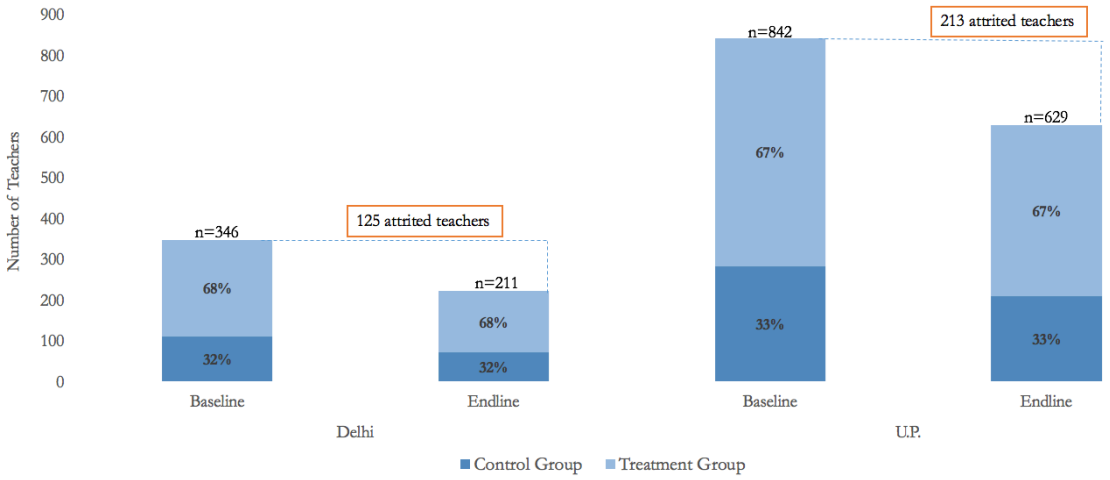
One underpinning assumption of randomized evaluations is that treatment and comparison groups were and remain, on average, similar on key characteristics. This similarity is the basis for the comparative claim of relative (to what would have happened absent the program) changes to teachers in the treated schools. If we see, for example, a greater number of teachers in the treated schools attriting from (leaving) our sample than is the case in comparison schools, this warrants concern about the validity of our causal comparison. This is because we worry that the program itself may be causing the overall differential attrition levels between treated and comparison groups.

Following from section 4.4.4, we provide here details of the tests we ran for attrition. As mentioned previously the evaluations have faced high rates of attrition at the teacher and the student level.

Teacher level

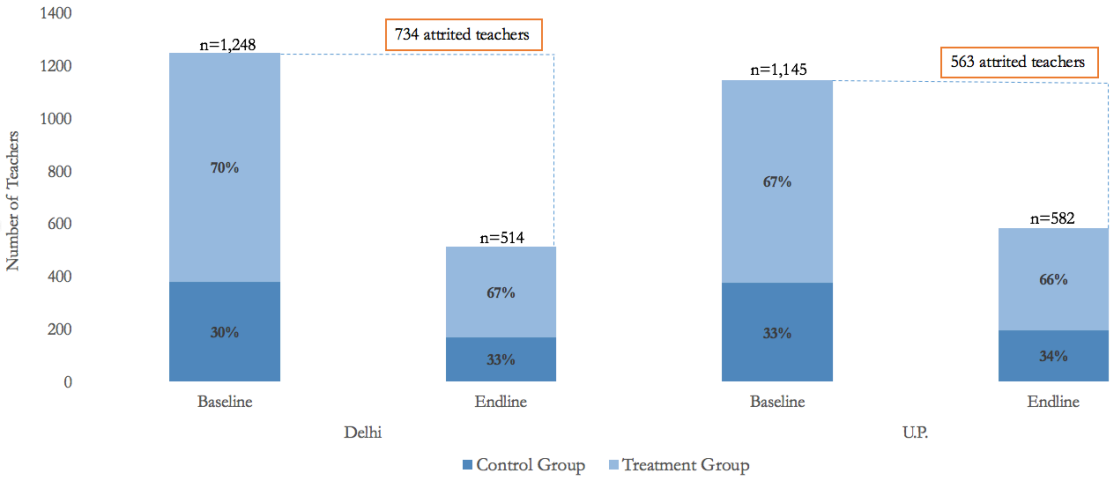
At the teacher level, we notice large number of teachers dropping out of our sample lists between baseline and endline. The figures below give a quick indication of number of teachers who dropped out of the sample lists (CO and PMB).

Figure Af: Attrition in teacher sample from baseline to endline (classroom observation survey)



Note: The endline numbers presented here are not the "final" number of teachers surveyed since new teachers were added at endline. New teachers were added in U.P. only if all teachers in a school from our list had dropped out and new teachers were added in Delhi only where less than two teachers from our list were available. The numbers here represent teachers from baseline lists who were followed up with. Teacher attrition is largely due to dropout/transfers from schools in our sample. There is no differential attrition of teachers across treatment arms.

Figure Ag: Attrition in teacher sample from baseline to endline (PMB survey)



Note: As with above, the endline numbers presented here are not the "final" number of teachers surveyed since new teachers were added at endline. There is no differential attrition of teachers across treatment arms.

It is possible to assess the concern of differential attrition between treatment and comparison samples empirically. First, we investigate overall differential attrition by comparing teacher dropout across treatment status (comparison, standard, exploratory) in the different study sites (U.P., Delhi) for the two survey rounds (Professional Mindsets & Behaviors (PMB), Classroom Observations (CO)). If differential attrition was absent, overall trends for attriting teachers are expected to be comparable across treatment and comparison groups. Thus, we compare average teacher dropout across treatment status. Statistical inference on differential attrition is based on a simple linear regression of the teacher-level dropout indicator (which equals 1 if the teacher dropped out at endline) on indicators of the treatment status, where the omitted category is given by the comparison group.[33] Technically speaking, we find evidence

[33] All attrition tests presented in this section cluster standard errors at the school level.

against similar attrition trends if we reject the null hypothesis from a joint test for both treatment indicators being equal to zero.

In Table A11, we report overall attrition rates from the PMB survey in Delhi. Overall, dropout is approximately 60%. Comparing dropout across the entire treatment group and across treatment arms, in a similar manner as described above, we find no evidence for differential attrition at the 5% level of significance.

Table A11: Assessment of differential attrition of PMB sample - endline in Delhi

	Overall	Treatment	Comparison	Standard	Exploratory	p-val (all STIR)	P-val (std. & exploratory)
Attrition	0.59	0.60	0.56	0.59	0.62		
	(0.492)	(0.490)	(0.498)	(0.492)	(0.487)	0.24	0.46
n	1249	868	381	509	359		

Note: This table reports means and standard deviations (in parentheses) of teacher attrition from baseline to endline. The final two columns contain results from a balance test. Technically, we run two OLS regressions of a binary dropout indicator on (1) an overall treatment indicator (All STIR) and (2) on two binary variables representing the two treatment arms (standard & exploratory). For the later regression, we use a joint test for whether any of the treatment arm coefficients is statistically different from 0. If the p-value is lower than 5%, we are able to reject the null hypothesis of similar means. * < 0.10, ** < 0.05, *** < 0.01.

In Table A12, we report overall attrition rates from the PMB survey in U.P. Overall, dropout is about 50%. Comparing dropout across the entire treatment group and across treatment arms, in a similar manner as described above, we find no evidence for differential attrition rates at the 5% level of significance.

Table A12: Assessment of differential attrition of PMB sample – endline in U.P.

	Overall	Treatment	Comparison	Standard	Exploratory	p-val (all STIR)	p-val (std. & exploratory)
Attrition	0.49	0.50	0.48	0.51	0.48		
	(0.500)	(0.500)	(0.500)	(0.500)	(0.5)	0.48	0.55
n	1145	770	375	382	388		

Note: This table reports means and standard deviations (in parentheses) of teacher attrition from baseline to endline. The final two columns contain results from a balance test. Technically, we run two OLS regressions of a binary dropout indicator on (1) an overall treatment indicator (All STIR) and (2) on two binary variables representing the two treatment arms (standard & exploratory). For the later regression, we use a joint test for whether any of the treatment arm coefficients is statistically different from 0. If the p-value is lower than 5%, we are able to reject the null hypothesis of similar means. * < 0.10, ** < 0.05, *** < 0.01.

In Table A13, we report overall attrition rates from the CO survey in Delhi for those teachers surveyed at baseline.[34] Overall, dropout is about 36%. Comparing dropout across the comparison group and entire treatment group, we see that overall dropout of teachers is statistically equivalent. However, when comparing the dropout across the standard and exploratory treatment arms against the dropout of the comparison group, we see a statistically

[34]As described above, the target list at endline is given by the CP Delhi baseline list. At endline some teachers from this list are part of the survey, even though not surveyed at baseline. Further, there are additions to the Delhi CO baseline list according to the decision rule. Since the decision rule focuses on random re-sampling of teachers, we do not expect these teachers to differ in terms of baseline characteristics from teachers already in the CP baseline list.

significant difference at the 10% level. We would encourage readers to read through the further analysis we provide ahead before thinking through implications of this finding on our evaluation results.

Table A13: Assessment of differential attrition at classroom practice - endline in Delhi

	Overall	Treatment	Comparison	Standard	Exploratory	p-val (all STIR)	p-val (std. & exploratory)
Attrition	0.36	0.36	0.36	0.44	0.26		
	(0.481)	(0.482)	(0.482)	(0.498)	(0.443)	0.98	0.08*
n	342	235	107	133	102		

Note: This table reports means and standard deviations (in parentheses) of teacher attrition from baseline to endline. The final two columns contain results from a balance test. Technically, we run two OLS regressions of a binary dropout indicator on (1) an overall treatment indicator (All STIR) and (2) on two binary variables representing the two treatment arms (standard & exploratory). For the later regression, we use a joint test for whether any of the treatment arm coefficients is statistically different from 0. If the p-value is lower than 5%, we are able to reject the null hypothesis of similar means. * < 0.10, ** < 0.05, *** < 0.01.

In Table A14, we report overall attrition rates from the CP survey in U.P. Overall, dropout is about 25%. Comparing dropout across the entire treatment group and across the two treatment arms, in a similar manner as described above, we find no evidence for differential attrition at the 5% level of significance.

Table A14: Assessment of differential attrition at classroom practice - endline in U.P.

	Overall	Treatment	Comparison	Standard	Exploratory	p-val (all STIR)	p-val (std. & exploratory)
Attrition	0.25	0.25	0.26	0.23	0.28		
	(0.435)	(0.434)	(0.437)	(0.420)	(0.447)	0.93	0.53
n	838	560	278	273	287		

Note: This table reports means and standard deviations (in parentheses) of teacher attrition from baseline to endline. The final two columns contain results from a balance test. Technically, we run two OLS regressions of a binary dropout indicator on (1) an overall treatment indicator (All STIR) and (2) on two binary variables representing the two treatment arms (standard & exploratory). For the later regression, we use a joint test for whether any of the treatment arm coefficients is statistically different from 0. If the p-value is lower than 5%, we are able to reject the null hypothesis of similar means. * < 0.10, ** < 0.05, *** < 0.01.

In addition to caring about the absolute numbers of treated and comparison schools and teachers that remain in the sample, we care about the composition of these samples. Some central characteristics are 'time-invariant' characteristics – ones that we do not expect to change with the passing of time or with the introduction of the program. This includes, for example, teacher sex and teacher experience (less experienced teachers cannot suddenly gain more years of experience than long-serving teachers).

For other characteristics, we *do* expect to see change over the course of time and with the introduction of the program. A key example of this is motivation level. What is important for the validity of our causal claims, then, is maintaining balance on initial baseline characteristics. For example, we want to see similar proportions of teachers with initially low baseline motivation remain in our sample, regardless of how their motivation levels changed between the baseline and endline measurements.

Column 1 of Table A15 reports summary statistics of baseline and time-invariant characteristics for those teachers that make up our endline PMB sample (note: new teachers added at endline are included in these results)[35]. Columns 2 to 5 report characteristics for the respective subsample indicated in the header. The final two columns present the results of statistical tests to assess the similarity of baseline characteristics between endline treatment and comparison groups.[36]

We find that teachers in the PMB survey in Delhi at endline do not significantly differ when comparing across the entire STIR treatment group and the control comparison group in their baseline characteristics (at the 5% level of significance). However, teacher qualification appears to be imbalanced when comparing the control group with either the standard or the exploratory treatment arms in isolation (at the 10% level of significance). We are confident that, despite this imbalance, the interpretation of the professional mindset and behavior results for the two treatment arms need not be caveated for the following reasons: (a) this imbalance existed in the baseline sample, implying that this is not the result of differential attrition; (b) there is only one imbalanced characteristic, which may be a result of a Type 1 error; and (c) we added teacher qualification as a control variable in our professional mindset and behavior regressions to account for differential qualification levels across the respective treatment arms and the comparison group.

[35] This analysis includes teachers added to the survey list at endline. Most baseline teacher socio-economic characteristics for this sample could be inferred using data from endline (e.g. age & gender). Baseline levels of the outcome variable were imputed for added teachers using simple mean imputation.

[36] The p-value in the second to last column comes from a regression of the baseline characteristic indicated in the respective row on a dummy variable representing whether a teacher was part of a treatment school or not. Standard errors are clustered at the level of the school.

The p-value in the last column comes from a joint hypotheses test that is based on a regression of the baseline characteristic indicated in the respective row on two treatment indicators – whether part of the standard or exploratory treatment arm (for those teachers in the endline sample) – with standard errors clustered at the level of the school.

Table A15: Balance of observed characteristics in Delhi PMB sample

	Endline Overall	Endline Treatment	Endline Comparison	Endline Standard	Endline Exploratory	p-val (all STIR)	p-val (std. & exploratory)
Baseline Outcomes							
Motivation (BL)	1.95	1.94	1.97	1.90	1.99	0.53	0.36
	(0.813)	(0.817)	(0.807)	(0.837)	(0.788)		
Socio-Demographics							
Age	28.19	28.14	28.31	27.87	28.48	0.86	0.86
	(10.331)	(10.345)	(10.317)	(10.363)	(10.328)		
Experience	5.43	5.54	5.21	5.37	5.77	0.42	0.52
	(5.534)	(5.707)	(5.159)	(5.789)	(5.600)		
Gender (Female)	0.94	0.93	0.94	0.93	0.94	0.75	0.90
	(0.244)	(0.247)	(0.238)	(0.253)	(0.239)		
Educ: >= Bachelor	0.11	0.12	0.11	0.15	0.07	0.75	0.06*
	(0.318)	(0.321)	(0.312)	(0.357)	(0.261)		
Educ: <=Pass 12th	0.89	0.88	0.89	0.85	0.93	0.75	0.06
	(0.318)	(0.321)	(0.312)	(0.357)	(0.261)		
n	1072	722	350	408	314		

Note: This table reports means and standard deviations (in parentheses) of baseline characteristics at endline. The final two columns contain results from a balance test of characteristics among the teachers in the endline sample. The column, "p-value (all STIR)" contains the p-value in an OLS regression of the characteristic on a dummy variable indicating whether the teacher was part of the treatment or comparison group (standard errors clustered at the school level). The last column presents results from an OLS regression of the given characteristic on a set of two treatment indicators and jointly tests whether any of these coefficients is statistically different from 0. If the p-value is lower than 5%, we are able to reject the null hypothesis of similar means. * < 0.10, ** < 0.05, *** < 0.01.

Table A16 presents the same analysis as above for the teachers that form our endline classroom observation sample (note: as before, new teachers added at endline were included in these results). We find that teachers in the CP sample in Delhi at endline do not differ across treatment groups in terms of their baseline characteristics at the 5% level of significance.

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Table A16: Balance of observed characteristics in Delhi classroom practice sample

	Endline Overall	Endline Treatment	Endline Comparison	Endline Standard	Endline Exploratory	p-val (all STIR)	p-val (std. & exploratory)
Baseline Outcomes							
Motivation	1.99	2.03	1.90	1.93	2.14	0.25	0.13
	(0.904)	(0.911)	(0.885)	(0.892)	(0.925)		
Teaching	0.69	0.69	0.69	0.67	0.70	0.76	0.49
	(0.240)	(0.227)	(0.266)	(0.232)	(0.222)		
Off Task	0.02	0.02	0.02	0.01	0.02	0.70	0.42
	(0.047)	(0.044)	(0.054)	(0.027)	(0.058)		
Smile, Laugh	0.24	0.22	0.27	0.26	0.19	0.30	0.21
	(0.309)	(0.288)	(0.349)	(0.293)	(0.281)		
At least 1 Qn	0.64	0.64	0.64	0.64	0.64	0.99	1.00
	(0.378)	(0.371)	(0.396)	(0.387)	(0.357)		
Local Info	0.25	0.25	0.25	0.20	0.31	0.94	0.22
	(0.363)	(0.349)	(0.393)	(0.315)	(0.373)		
Learning Aides	0.48	0.48	0.46	0.48	0.48	0.80	0.96
	(0.402)	(0.406)	(0.396)	(0.396)	(0.418)		
Group Work	0.01	0.01	0.00	0.01	0.01	0.34	0.59
	(0.054)	(0.060)	(0.040)	(0.066)	(0.054)		
Baseline Teacher Characteristics							
Gender (female)	0.95	0.95	0.94	0.95	0.94	0.82	0.88
	(0.222)	(0.219)	(0.230)	(0.208)	(0.232)		
Gender (missing)	0.33	0.32	0.34	0.32	0.32	0.70	0.93
	(0.470)	(0.468)	(0.475)	(0.469)	(0.468)		
Experience	5.66	5.65	5.69	5.52	5.81	0.95	0.92
	(5.601)	(5.662)	(5.482)	(6.061)	(5.135)		
Experience (missing)	0.34	0.33	0.35	0.33	0.33	0.73	0.94
	(0.473)	(0.471)	(0.478)	(0.473)	(0.471)		
Age	28.23	28.11	28.49	27.68	28.64	0.69	0.70
	(8.471)	(8.472)	(8.491)	(8.785)	(8.061)		
Age (missing)	0.33	0.33	0.34	0.32	0.33	0.80	0.95
	(0.471)	(0.470)	(0.475)	(0.469)	(0.473)		
Educ: <=Pass 12th	0.81	0.80	0.82	0.75	0.86	0.69	0.17
	(0.396)	(0.402)	(0.386)	(0.433)	(0.350)		
Educ: >= Bachelor	0.19	0.20	0.18	0.25	0.14	0.69	0.17
	(0.396)	(0.402)	(0.386)	(0.433)	(0.350)		
Qualification (missing)	0.33	0.32	0.34	0.32	0.32	0.70	0.93
	(0.470)	(0.468)	(0.475)	(0.469)	(0.468)		
N	462	318	144	177	141		

Note: This table reports means and standard deviations (in parentheses) of baseline characteristics at endline. The final two columns contain results from a balance test of characteristics among the teachers in the endline sample. The column, "p-value (all STIR)" contains the p-value in an OLS regression of the characteristic on a dummy variable indicating whether the teacher was part of the treatment or comparison group (standard errors clustered at the school level). The last column presents results from an OLS regression of the given characteristic on a set of two treatment indicators and jointly tests whether any of these coefficients is statistically different from 0. If the p-value is lower than 5%, we are able to reject the null hypothesis of similar means. * < 0.10, ** < 0.05, *** < 0.01.

In Table A17, we present baseline characteristics of the U.P. PMB sample. We find no evidence against similar baseline characteristics of teachers surveyed at endline.

Table A17: Balance of observed characteristics in U.P. PMB sample

	Endline Overall	Endline Treatment	Endline Comparison	Endline Standard	Endline Exploratory	p-val (all STIR)	p-val (std. & exploratory)
Baseline Outcomes							
Motivation	1.75	1.74	1.79	1.72	1.76	0.34	0.50
	(0.756)	(0.749)	(0.770)	(0.740)	(0.758)		
Baseline Teacher Characteristics							
Gender (Female)	0.56	0.55	0.58	0.54	0.56	0.53	0.77
	(0.497)	(0.498)	(0.495)	(0.499)	(0.497)		
Experience	9.86	9.83	9.94	9.90	9.76	0.76	0.93
	(5.082)	(5.123)	(5.001)	(4.889)	(5.353)		
Age	37.07	37.08	37.05	37.46	36.70	0.96	0.56
	(8.484)	(8.589)	(8.273)	(8.627)	(8.545)		
Educ: <= Bachelor	0.23	0.23	0.23	0.21	0.25	1.00	0.46
	(0.419)	(0.419)	(0.419)	(0.407)	(0.431)		
Educ: >= M.Phil.	0.77	0.77	0.77	0.79	0.75	1.00	0.46
	(0.419)	(0.419)	(0.419)	(0.407)	(0.431)		
Endline addition	0.49	0.50	0.46	0.52	0.48	0.27	0.27
	(0.500)	(0.500)	(0.499)	(0.500)	(0.500)		
N	1133	767	366	384	383		

Note: This table reports means and standard deviations (in parentheses) of baseline characteristics at endline. The final two columns contain results from a balance test of characteristics among the teachers in the endline sample. The column, "p-value (all STIR)" contains the p-value in an OLS regression of the characteristic on a dummy variable indicating whether the teacher was part of the treatment or comparison group (standard errors clustered at the school level). The last column presents results from an OLS regression of the given characteristic on a set of two treatment indicators and jointly tests whether any of these coefficients is statistically different from 0. If the p-value is lower than 5%, we are able to reject the null hypothesis of similar means. * < 0.10, ** < 0.05, *** < 0.01.

Finally, we look at teachers who were part of the endline classroom observation sample in U.P. In Table A18 we see that overall treatment group teachers surveyed at endline are comparable to comparison group teachers on baseline characteristics.[37] In contrast, though, we do find evidence of imbalance in our treatment arm sample. In particular, teachers in standard and exploratory schools seem to differ in their baseline time-use compared to comparison schools. Teachers in exploratory schools are observed to spend less time teaching and more time off-task. However, these differences are not the result of differential attrition as these differences existed in the baseline sample. To control for this imbalance, all time-use regressions at endline include baseline levels as a covariate.

[37] Note that this analysis, as above, includes teachers added to the survey list at endline. Most baseline teacher socio-economic characteristics for this sample could be identified using data from endline (e.g. age & gender). Baseline levels of the outcome variable were imputed for added teachers using simple mean imputation.

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Table A18: Balance of observed characteristics at classroom practice endline in U.P.

	Endline Overall	Endline Treatment	Endline Comparison	Endline Standard	Endline Exploratory	p-val (all STIR)	p-val (std. & exploratory)
Baseline Outcomes							
Motivation	1.76	1.72	1.85	1.67	1.78	0.17	0.21
	(1.010)	(1.014)	(0.999)	(1.012)	(1.015)		
Teaching	0.78	0.77	0.78	0.82	0.72	0.78	0.03**
	(0.330)	(0.331)	(0.326)	(0.296)	(0.357)		
Off Task	0.15	0.15	0.15	0.11	0.19	0.98	0.03**
	(0.293)	(0.294)	(0.292)	(0.243)	(0.334)		
Smile, Laugh	0.07	0.06	0.07	0.04	0.08	0.51	0.11
	(0.183)	(0.179)	(0.190)	(0.135)	(0.214)		
At least 1 Qn	0.24	0.26	0.20	0.30	0.21	0.24	0.16
	(0.353)	(0.361)	(0.334)	(0.389)	(0.326)		
Local Info	0.13	0.14	0.13	0.17	0.10	0.71	0.22
	(0.276)	(0.286)	(0.256)	(0.316)	(0.248)		
Learning Aides	0.44	0.47	0.39	0.51	0.43	0.10	0.07*
	(0.426)	(0.427)	(0.419)	(0.428)	(0.423)		
Group Work	0.06	0.05	0.07	0.05	0.05	0.47	0.76
	(0.200)	(0.185)	(0.226)	(0.188)	(0.182)		
Baseline Teacher Characteristics							
Gender (female)	0.58	0.57	0.59	0.58	0.56	0.76	0.91
	(0.495)	(0.496)	(0.494)	(0.495)	(0.497)		
Gender (missing)	0.03	0.03	0.02	0.03	0.03	0.53	0.82
	(0.169)	(0.177)	(0.152)	(0.176)	(0.179)		
Experience	10.44	10.47	10.38	10.35	10.59	0.89	0.96
	(6.735)	(6.834)	(6.545)	(6.549)	(7.130)		
Experience (missing)	0.05	0.05	0.05	0.04	0.07	0.78	0.79
	(0.221)	(0.225)	(0.212)	(0.199)	(0.249)		
Age	38.05	37.98	38.20	38.26	37.69	0.76	0.77
	(8.052)	(8.150)	(7.868)	(8.376)	(7.918)		
Age (missing)	0.05	0.05	0.04	0.04	0.06	0.78	0.78
	(0.211)	(0.216)	(0.202)	(0.188)	(0.240)		
Educ: <= Bachelor	0.41	0.42	0.39	0.40	0.45	0.47	0.46
	(0.492)	(0.494)	(0.489)	(0.490)	(0.498)		
Educ: >= M.Phil.	0.57	0.56	0.59	0.59	0.52	0.52	0.37
	(0.496)	(0.497)	(0.494)	(0.493)	(0.501)		
Qualification (missing)	0.06	0.06	0.06	0.05	0.07	0.84	0.75
	(0.233)	(0.230)	(0.240)	(0.209)	(0.249)		
N	644	431	213	219	212		

Note: This table reports means and standard deviations (in parentheses) of baseline characteristics at endline. The final two columns contain results from a balance test of characteristics among the teachers in the endline sample. The column, "p-value (all STIR)" contains the p-value in an OLS regression of the characteristic on a dummy variable indicating whether the teacher was part of the treatment or comparison group (standard errors clustered at the school level). The last column presents results from an OLS regression of the given characteristic on a set of two treatment indicators and jointly tests whether any of these coefficients is statistically different from 0. If the p-value is lower than 5%, we are able to reject the null hypothesis of similar means. * < 0.10, ** < 0.05, *** < 0.01.

In addition to testing endline balance on baseline characteristics, we run one final check to test for differential attrition. In what follows, we present results from regressions of teacher attrition on baseline characteristics and on time-invariant teacher characteristics, denoted X below, interacted by treatment assignment.

$$Attrited = \alpha + \beta_1 X + \beta_2 Treatment + \beta_3 Treatment * X + \varepsilon$$

These regressions perform a check of whether baseline characteristics of attritors are different across treatment and comparison groups. For example, we test whether attritor teachers in the treatment group are more likely to have their masters degrees than comparison teachers, which if true, would cast doubt on the assumption of non-differential attrition.

Table A18.1: Comparison of Attritor Characteristics Between Treatment Groups – PMB Sample

	Treatment Group (Delhi)	Comparison Group (Delhi)	p-val of difference (Delhi)	Treatment Group (U.P.)	Comparison Group (U.P.)	p-val of difference (U.P.)
Baseline Outcomes						
Motivation	0.001	-0.028	0.25	-0.002	-0.009	0.79
Baseline Teacher Characteristics						
Gender (female)	0.138	0.408	0.02**	-0.005	-0.035	0.65
Experience	-0.017	-0.022	0.38	0.004	0.006	0.49
Age	-0.012	-0.015	0.42	0.003	0.003	0.93
Educ: >= Bachelor	0.014	0.082	0.37	-0.148	-0.241	0.38
Educ: <= 12th Pass	-0.014	-0.082	0.37	0.148	0.241	0.38
N	868	380		770	375	

Note: This table reports coefficients from a linear probability model with the dependent variable indicating whether the teacher dropped out from baseline to endline and the independent variables include a series of baseline outcomes and time-invariant baseline teacher characteristics. The column, "p-value of difference" contains the p-value from a regression of the interaction between treatment status and each baseline characteristic (standard errors clustered at the school level). If the p-value is lower than 5%, we are able to reject the null hypothesis of similar means, indicating that the attritors have different characteristics between the treatment and control group. Note the regressions do not include sample added after baseline. * < 0.10, ** < 0.05, *** < 0.01.

Table A18.2: Comparison of Attritor Characteristics Between Treatment Groups – Classroom Practice Sample

	Treatment Group (Delhi)	Comparison Group (Delhi)	p-val of difference (Delhi)	Treatment Group (U.P.)	Comparison Group (U.P.)	p-val of difference (U.P.)
Baseline Outcomes						
Motivation	-0.006	0.008	0.82	0.010	-0.021	0.17
Teaching	-0.050	0.175	0.11	-0.07	-0.078	0.94
Off Task	0.785	-1.064	0.00***	0.114	0.082	0.77
Smile, Laugh	0.071	0.005	0.71	-0.056	0.12	0.30
At least 1 Qn	-0.164	0.073	0.13	0.003	0.096	0.38
Local Info	-0.144	-0.012	0.32	0.014	0.063	0.72
Learning Aides	-0.198	-0.069	0.35	-0.013	-0.012	0.99
Group Work	0.358	0.798	0.46	0.084	0.039	0.76
Baseline Teacher Characteristics						
Gender (female)	0.388	0.379	0.97	0.371	0.375	0.95
Experience	-0.017	-0.021	0.55	-0.008	-0.008	0.96
Age	-0.009	-0.012	0.51	0.000	-0.001	0.76
Educ: >= Bachelor	-0.436	-0.435	0.99	0.268	0.27	0.97
Educ: <= 12th Pass	0.436	0.435	0.99	-0.268	-0.27	0.97
N	235	111		560	282	

Note: This table reports coefficients from a linear probability model with the dependent variable indicating whether the teacher dropped out from baseline to endline and the independent variables include a series of baseline outcomes and time-invariant baseline teacher characteristics. The column, "p-value of difference" contains the p-value from a regression of the interaction between treatment status and baseline characteristic (standard errors clustered at the school level). If the p-value is lower than 5%, we are able to reject the null hypothesis of similar means, indicating that the attritors have different characteristics between the treatment and control group. Note the regressions do not include sample added after baseline. * < 0.10, ** < 0.05, *** < 0.01.

In both the PMB teacher sample and the classroom practice teacher sample, there does not exist strong evidence that the attrition is unbalanced across treatment groups. Largely, attritors share the same characteristics across treatment and the comparison.

Summary and implications for teacher level results:

Our main concern when looking at attrition is if our results are in any way driven by differential attrition (both across treatment arms and baseline characteristics) thus resulting in biased estimates. Our interpretations from the tests above is that there is no conclusive evidence suggesting that differential attrition is confounding our estimates of treatment effects.

- Overall there is no conclusive trend of differential attrition between treatment arms across survey rounds e.g.: there is no reasonable theory that could explain why there is such high dropout in standard schools in Delhi private schools in the classroom observation sample (table A10) while not for the PMB sample (table A8) given that for both surveys, baseline and endline happen in different school years. Furthermore, teachers in standard schools continue to show similar baseline characteristics as teachers in other schools. The significant results we see is much likely due to a chance difference or a 'false-positive' given the multiple hypotheses we run or an imbalance

that was present originally at baseline. Furthermore, there are no imbalances when comparing baseline characteristics between the aggregated treatment group and the comparison group, which should inspire confidence in the validity of the all-STIR treatment effect results.

- Our endline results come from multivariate regressions that control for baseline characteristics. This means that differential attrition (by baseline characteristics) would be problematic only were we to assume that treatment effects vary in baseline characteristics (heterogenous treatment effects). We ran these heterogeneity tests for baseline teaching and off-task in U.P. (which appear as significant in table A18), which showed that there is no heterogeneity in treatment by baseline time use measures.

Thus, for the teacher level parameters we feel attrition does not skew our results – it could have at most a very marginal influence that, even under the most favorable assumptions, would not overturn our key findings.

Student level

Similar to the tests at the teacher level, we ran tests at the student level to check for differential attrition across treatment arms and baseline characteristics.

In Table A19, we report overall attrition rates from the student testing survey in U.P. Overall, dropout is about 57%. Comparing dropout across the treatment arms, we find no evidence for differential attrition at the 5% level of significance.

Table A19: Assessment of differential attrition at student testing endline in U.P.

	Overall	Treatment	Comparison	Standard	Exploratory	p-val (all STIR)	p-val (std. & exploratory)
Attrition	0.57	0.56	0.59	0.57	0.56	0.51	0.78
	(0.495)	(0.496)	(0.493)	(0.495)	(0.497)		
N	7386	4829	2557	2335	2494	.	.

Note: This table reports means and standard deviations (in parentheses) of student attrition from baseline to endline. The final two columns contain results from a balance test. Technically, we run two OLS regressions of a binary dropout indicator on (1) an overall treatment indicator (All STIR) and (2) on two binary variables representing the two treatment arms (standard & exploratory). For the later regression, we use a joint test for whether any of the treatment arm coefficients is statistically different from 0. If the p-value is lower than 5%, we are able to reject the null hypothesis of similar means. * < 0.10, ** < 0.05, *** < 0.01.

Similarly looking at the attrition rates from the student testing survey in Delhi (Table A20) we find no evidence for differential attrition at the 5% level of significance.

Table A20: Assessment of differential attrition at student testing endline in Delhi

	Overall	Treatment	Comparison	Standard	Exploratory	p-val (all STIR)	p-val (std. & exploratory)
Attrition	0.45	0.45	0.44	0.48	0.40	0.86	0.44
	(0.497)	(0.497)	(0.497)	(0.500)	(0.491)		
N	3047	2063	984	1217	846	.	.

Note: This table reports means and standard deviations (in parentheses) of student attrition from baseline to endline. The final two columns contain results from a balance test. Technically, we run two OLS regressions of a binary dropout indicator on (1) an overall treatment indicator (All STIR) and (2) on two binary variables representing the two treatment arms (standard & exploratory). For the later regression, we use a joint test for whether any of the treatment arm coefficients is statistically different from 0. If the p-value is lower than 5%, we are able to reject the null hypothesis of similar means. * < 0.10, ** < 0.05, *** < 0.01.

Column 1 of Table A21 reports summary statistics of baseline characteristics for those students in U.P. that were surveyed at endline. Columns 2 to 5 report similar characteristics for the respective subsample indicated in the header. We observe that overall these characteristics seem comparable. The final two columns assess the similarity of baseline characteristics across the overall treatment and the two treatment arms respectively. We find that students in U.P. at endline do not differ in their baseline characteristics at the 5% level of significance.

Table A21: Balance of observed characteristics at student testing endline in U.P.

	Overall	Treatment	Comparison	Standard	Exploratory	p-val (all STIR)	p-val (std. & exploratory)
Hindi levels (BL)	1.81	1.74	1.95	1.76	1.73	0.23	0.48
	(1.825)	(1.812)	(1.843)	(1.860)	(1.768)		
Math levels (BL)	1.63	1.56	1.77	1.54	1.58	0.17	0.37
	(1.671)	(1.625)	(1.752)	(1.668)	(1.586)		
Baseline Student Characteristics							
Gender (Female)	0.55	0.55	0.55	0.56	0.55	0.75	0.82
	(0.497)	(0.497)	(0.498)	(0.496)	(0.498)		
Gender (Missing)	0.02	0.02	0.02	0.03	0.02	0.97	0.78
	(0.154)	(0.155)	(0.153)	(0.176)	(0.133)		
Age	8.44	8.35	8.60	8.33	8.37	0.23	0.48
	(1.759)	(1.708)	(1.848)	(1.794)	(1.626)		
Second grade	0.19	0.19	0.19	0.20	0.18	0.96	0.88
	(0.394)	(0.394)	(0.395)	(0.400)	(0.388)		
Third grade	0.23	0.25	0.20	0.22	0.27	0.12	0.09
	(0.421)	(0.430)	(0.399)	(0.412)	(0.445)		
Fourth grade	0.15	0.16	0.13	0.15	0.17	0.28	0.40
	(0.359)	(0.367)	(0.341)	(0.357)	(0.375)		
Fifth grade	0.10	0.10	0.10	0.10	0.11	0.73	0.88
	(0.302)	(0.306)	(0.296)	(0.299)	(0.311)		
Sixth grade	0.18	0.15	0.23	0.17	0.14	0.17	0.35
	(0.384)	(0.360)	(0.423)	(0.372)	(0.348)		
Seventh grade	0.00	0.00	0.00	0.00	0.00	0.16	0.37
	(0.025)	(0.031)	(0.000)	(0.032)	(0.030)		
Eighth grade	0.00	0.00	0.00	0.00	0.00	0.53	0.37
	(0.040)	(0.031)	(0.054)	(0.000)	(0.043)		
N	3152	2111	1041	1006	1105	.	.

Note: This table reports means and standard deviations (in parentheses) of baseline characteristics at endline. The final two columns contain results from a balance test of characteristics among the students in the endline sample. The column, "p-value (all STIR)" contains the p-value in an OLS regression of the characteristic on a dummy variable indicating whether the student was part of the treatment or comparison group (standard errors clustered at the school level). The last column presents results from an OLS regression of the given characteristic on a set of two treatment indicators and jointly tests whether any of these coefficients is statistically different from 0. If the p-value is lower than 5%, we are able to reject the null hypothesis of similar means. * < 0.10, ** < 0.05, *** < 0.01.

Similarly, Table A22 summarizes baseline characteristics for students in Delhi private schools who were surveyed at midline. We find that students in Delhi at midline do not differ in their baseline characteristics at the 5% level of significance.

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Table A22: Balance of observed characteristics at student testing endline in Delhi

	Overall	Treatment	Comparison	Standard	Exploratory	p-val (all STIR)	p-val (std. & exploratory)
Hindi levels (BL)	3.56	3.55	3.58	3.48	3.64	0.91	0.72
	(1.804)	(1.776)	(1.864)	(1.749)	(1.805)		
Math levels (BL)	3.31	3.24	3.45	3.28	3.20	0.40	0.62
	(1.834)	(1.814)	(1.871)	(1.833)	(1.791)		
Baseline Student Characteristics							
Gender (Female)	0.37	0.36	0.38	0.39	0.34	0.60	0.25
	(0.482)	(0.481)	(0.485)	(0.487)	(0.473)		
Age	8.75	8.78	8.68	8.80	8.76	0.73	0.93
	(2.067)	(1.970)	(2.263)	(2.067)	(1.854)		
Second grade	0.12	0.10	0.16	0.12	0.09	0.19	0.29
	(0.328)	(0.306)	(0.368)	(0.326)	(0.280)		
Third grade	0.17	0.19	0.12	0.20	0.17	0.12	0.24
	(0.371)	(0.389)	(0.325)	(0.402)	(0.373)		
Fourth grade	0.16	0.18	0.12	0.18	0.17	0.11	0.26
	(0.365)	(0.382)	(0.321)	(0.386)	(0.377)		
Fifth grade	0.09	0.08	0.11	0.07	0.08	0.37	0.66
	(0.283)	(0.268)	(0.313)	(0.264)	(0.272)		
Sixth grade	0.07	0.06	0.08	0.04	0.09	0.47	0.23
	(0.250)	(0.236)	(0.278)	(0.185)	(0.283)		
Seventh grade	0.07	0.06	0.11	0.08	0.03	0.17	0.08
	(0.260)	(0.229)	(0.315)	(0.266)	(0.173)		
Eighth grade	0.00	0.00	0.00	0.00	0.01	0.20	0.37
	(0.047)	(0.056)	(0.000)	(0.038)	(0.072)		
UKG grade	0.09	0.07	0.13	0.08	0.07	0.18	0.37
	(0.284)	(0.257)	(0.331)	(0.266)	(0.247)		
N	1846	1263	583	680	583	.	.

Note: This table reports means and standard deviations (in parentheses) of baseline characteristics at endline. The final two columns contain results from a balance test of characteristics among the students in the endline sample. The column, "p-value (all STIR)" contains the p-value in an OLS regression of the characteristic on a dummy variable indicating whether the student was part of the treatment or comparison group (standard errors clustered at the school level). The last column presents results from an OLS regression of the given characteristic on a set of two treatment indicators and jointly tests whether any of these coefficients is statistically different from 0. If the p-value is lower than 5%, we are able to reject the null hypothesis of similar means. * < 0.10, ** < 0.05, *** < 0.01.

Summary and implications for student level results:

Thus, to conclude, at the student level we find no evidence of differential attrition. We are limited in our ability to look at trends in baseline characteristics (since we only collect student data on age and sex) but within the limited scope of the analysis we do not find differential trends.

Appendix A14: Baseline balance checks

The treatment groups in both Delhi and U.P. appear to be well balanced across all treatment groups. 21 variables were tested for Delhi and 20 tested for U.P.[38] F-tests were used to determine the joint significance in difference in means across the three groups. Standard errors were clustered at the school level in all analyses presented.

In the Delhi study, two of 21 variables were significant at the 10% level and no variables tested were significant at the 5% or 1% levels. Theory would predict that roughly two out of 21 variables would be significant at the 10% level by chance, and that one variable at the 5% level would be significant by chance. This seems to indicate that the variables are well-balanced across the various study arms in Delhi.

In the Uttar Pradesh study, three out of the 20 variables were significant at the 5% level and one variable was significant at the 10% level. Two of these variables are fraction of time spend teaching and fraction of time spent off-task, which are not independent variables, meaning that imbalance in one is likely responsible for imbalance in both. Along with classroom management, the fractions for each of these variables must add up to 100%. The other significant variables were the fraction of teachers who used learning aids (at the 5% level) and fraction of students engaging in group discussion or Q&A (at the 10% level). While theory would only predict one variable that is significant at the 5% level assuming random distribution, IDinsight believes that the comparison and treatment groups are nonetheless well-balanced. IDinsight will also take measures to test for bias at endline, for example, by performing robustness checks incorporating baseline covariates.

Table A23: Delhi

	Mean comparison	Mean Intrinsic	Mean Extrinsic	Num obs.	Model-df	Reg-df	F-statistic	*p*-Value
Teacher Motivation Index	1.9	1.9	2.0	1256	2	178	0.80	0.45
Teacher Age	28.1	28.7	29.5	1252	2	178	0.86	0.42
Teaching Experience (Years)	5.6	5.7	6.5	1248	2	178	1.56	0.21
Female	95%	94%	92%	1257	2	178	1.26	0.29
Teacher Education[1]				1256	2		0.77	0.68
Additional Teacher Qualifications[2]				1251	12		14.08	0.30
Fraction of Time Teaching	73%	64%	71%	1384	2	140	1.92	0.15
Fraction of Time Managing Classroom	26%	33%	28%	1384	2	140	1.51	0.23
Fraction of Time Off Task	1%	2%	2%	1384	2	140	0.96	0.38
Fraction of Students Doing Drills	25%	29%	27%	1383	2	140	0.51	0.60
Fraction of Students Participating in a Group	17%	16%	15%	1383	2	140	0.16	0.85

[38] Additional teacher qualifications were not recorded at the baseline survey for U.P. These were collected only during the midline survey to take place in 2016.

Discussion								
Fraction of Students Listening to Lecture*	21%	19%	14%	1383	2	140	2.35	0.10
Fraction of Students Doing Silent Seatwork*	29%	27%	38%	1383	2	140	3.00	0.05
Fraction of Students Off Task	8%	8%	6%	1383	2	140	1.41	0.25
Fraction of Teachers Who Smiled at Least Once	75%	68%	80%	345	2	140	2.21	0.11
Fraction of Classrooms Where At Least One Student Asked a Question	33%	33%	35%	345	2	140	0.03	0.97
Fraction of Teachers Who Used Local Information while Teaching	78%	83%	77%	345	2	140	0.84	0.43
Fraction of Teachers Who Used a Learning Aid	58%	61%	63%	344	2	140	0.34	0.72
Fraction of Teachers Who Asked Students to Work in Small Groups	3%	2%	4%	342	2	139	0.56	0.57
Student's Math ASER Level[3]	3.8	3.6	3.6	3379	2		0.37	0.83
Student's Hindi ASER Level[3]	3.6	3.5	3.5	3379	2		0.67	0.71

Notes: Unless otherwise noted, F-statistics reflect the model specification statistic for a linear regression model with the outcome variable listed on the leftmost column and the explanatory variables as two binary variables indicating either intrinsic or extrinsic motivation treatment status. Unless otherwise noted, all standard errors in this table are clustered at the school level.
* Indicates significance at the 10% level.
** Indicates significance at the 5% level.
*** Indicates significance at the 1% level.
[1] Highest education level is an ordinal variable, so an ordered logit model was used. Because this model uses maximum likelihood, the model test is a chi-squared test, rather than an F-test. Standard errors were clustered at the school level. Group means have not been listed for this model.
[2] Teacher qualification is an unordered qualitative variable, and a Pearson's chi-squared test was used to determine differences in distributions among the different treatment arms. Standard errors are not clustered at the school level, so the p-value listed may be conservative.
[3] Student Math and Hindi scores are from the ASER test battery, which gives an ordinal value. Because this model uses maximum likelihood, the model test is a chi-squared test, rather than an F-test. Standard errors were clustered at the school level. Readers should use some caution when interpreting group means, as the numerical values from one category to the next are arbitrary.

Table A24: Uttar Pradesh

	Mean comparison	Mean Intrinsic	Mean Extrinsic	Num obs.	Model-df	Reg-df	F-statistic	*p*-Value
Teacher Motivation Index	1.8	1.7	1.8	1145	2	270	0.62	0.54
Teacher Age	38.9	38.7	38.4	1222	2	270	0.19	0.83
Teaching Experience (Years)	11.5	10.9	11.0	1214	2	270	0.46	0.63
Female	57%	52%	54%	1244	2	269	0.37	0.69
Teacher Education[1]				1205	2		0.09	0.96
Fraction of Time Teaching**	77%	81%	71%	3369	2	270	3.50	0.03

Fraction of Time Managing Classroom	7%	7%	8%	3369	2	270	0.22	0.80
Fraction of Time Off Task**	16%	12%	21%	3369	2	270	3.84	0.02
Fraction of Students Doing Drills	32%	27%	28%	3349	2	265	1.53	0.22
Fraction of Students Participating in a Group Discussion*	5%	4%	8%	3349	2	265	2.42	0.09
Fraction of Students Listening to Lecture	28%	34%	25%	3349	2	265	2.07	0.13
Fraction of Students Doing Silent Seatwork	16%	19%	18%	3349	2	265	0.94	0.39
Fraction of Students Off Task	18%	17%	22%	3349	2	265	2.18	0.11
Fraction of Teachers Who Smiled at Least Once	6%	4%	7%	841	2	265	1.62	0.20
Fraction of Classrooms Where At Least One Student Asked a Question	18%	29%	20%	841	2	265	2.09	0.13
Fraction of Teachers Who Used Local Information while Teaching	7%	13%	10%	841	2	265	1.43	0.24
Fraction of Teachers Who Used a Learning Aid**	32%	45%	33%	841	2	265	3.76	0.02
Fraction of Teachers Who Asked Students to Work in Small Groups	6%	6%	5%	841	2	265	0.04	0.96
Student's Math ASER Level[2]	2.3	2.0	2.0	7376	2		2.73	0.25
Student's Hindi ASER Level[2]	2.5	2.3	2.2	7376	2		2.46	0.29

Notes: Unless otherwise noted, F-statistics reflect the model specification statistic for a linear regression model with the outcome variable listed on the leftmost column and the explanatory variables as two binary variables indicating either intrinsic or extrinsic motivation treatment status. Unless otherwise noted, all standard errors in this table are clustered at the school level.
* Indicates significance at the 10% level.
** Indicates significance at the 5% level.
*** Indicates significance at the 1% level.
[1] Highest education level is an ordinal variable, so an ordered logit model was used. Because this model uses maximum likelihood, the model test is a chi-squared test, rather than an F-test. Standard errors were clustered at the school level. Group means have not been listed for this model.
[2] Student Math and Hindi scores are from the ASER test battery, which gives an ordinal value. Because this model uses maximum likelihood, the model test is a chi-squared test, rather than an F-test. Standard errors were clustered at the school level. Readers should use some caution when interpreting group means, as the numerical values from one category to the next are arbitrary.

Appendix A15: Sample sizes and school type

The number of teachers in the teacher motivation and classroom observation sample in each program type are mentioned below.

Table A25: Number of teachers by treatment arm

School type	Teacher motivation sample		Classroom observation sample	
	Delhi private	U.P. gov	Delhi private	U.P. gov
Core	408	384	177	219
Core-plus	314	383	141	212
Comparison	350	366	144	213
Number here represents number of teachers in core, core plus and comparison schools				

The number of teachers for each subgroup category are mentioned below. This has been split according to our teacher motivation and classroom observation sample lists.

Table A26: Number of teachers by subgroup categories

Sub group	Category	Teacher motivation sample		Classroom observation sample	
		Delhi private	U.P. gov	Delhi private	U.P. gov
Teacher Sex	Male	NA	488	NA	277
	Female	NA	645	NA	367
Block	Amava	NA	108	NA	66
	Dalmau	NA	70	NA	43
	Harchandpur	NA	92	NA	63
	Lalaganj	NA	94	NA	52
	Sataon	NA	66	NA	45
	Rahi	NA	62	NA	25
	Pindra	NA	84	NA	43
	Sevapuri	NA	167	NA	91
	Chiraegaun	NA	192	NA	120
	Araziline	NA	86	NA	41
	Kasividyapeeth	NA	84	NA	55
Number here represents number of teachers in each subcategory (counting teachers in both STIR and comparison schools)					

A block may be comprised of one or more STIR networks, implying that a block may either have schools belonging to one or multiple STIR programming variants, as shown in Table A27.

Table A27: Schools by blocks

Block Name	Core schools	Core-Plus schools
Amava	X	X
Dalmau		X
Harchandpur	X	X
Lalganj	X	X
Sataon		X

Rahi	X	
Pindra		X
Sevapuri	X	
Chiraegaun	X	X
Araziline		X
Kasividyapeeth	X	
Note all blocks had comparison schools		

Appendix A16: Primer on statistical inference

Statistical inference: asking one question of the data

Every question that we ask has a true answer: either STIR's program "works" (*i.e.,* there is a positive and high magnitude effect of the program on student learning outcomes) or it does "not work" (*i.e.,* there is no effect). As a researcher, though, we are not able to observe this truth directly and instead we resort to answering such questions using statistical inference applied to the data at hand. Since the data we have are generally a (random) sample of the population of interest (all teachers in treatment and comparison schools), such answers often contain (statistical) noise, also called sampling variation.

When a researcher asks a question of the data to get closer to the truth, the answer is generally based on a "test statistic".[39] This test statistic is compared to a reasonable threshold that – if crossed – would imply a "statistically significant effect". This threshold, however, is fully under the discretion of the researcher through choosing the level of confidence of the results. Higher thresholds are more difficult to cross, and such results are therefore considered more reliable (that is, less likely to be noise).

A technical measure for how large the test statistic is in comparison to the chosen threshold is the p-value. Put simply, the p-value measures how large the threshold could have been chosen to still indicate a statistically significant effect. More technically, the p-value indicate the smallest level of significance for which the effect would still be statistically significant.[40] In general, we say that the effect is statistically significant if the p-value is lower than the chosen level of significance α.

Figure Ay presents the four possible combinations of truth ("STIR has an effect" *v.* "STIR has no effect") and test result ("Effect is statistically significant" *v.* "Effect is not statistically significant"). If STIR, in truth, has an effect, then if the researcher's test indicates that there is a statistically significant effect, this would result in correct inference. The match between truth and our finding would make us happy, as shown in the northwest corner of Figure Ay.

If the researcher's test did not indicate a statistically significant effect, this situation would be referred to as a false negative. Put differently, in this situation, the researcher is not able to detect a true effect in the data. The likeliness of this situation depends on the detection

[39] Oftentimes, this test statistic is the t-statistic, which is derived as the fraction of the coefficient estimate of a simple linear regression model and its according standard error estimate.

[40] On a more conceptual note, it is evident that determining the level of significance prior to data analysis is warranted.

"power" of the test. Broadly speaking, a more powerful test can detect more truly existing effects (so that there would be less false negatives).

Assuming that, in truth, STIR has no effect, and the researcher's test does not suggest a statistically significant effect, the result – while disappointing from a programmatic point of view – would be valid (as shown the southeast quadrant of Figure Ay). If the researcher's test indicated a statistically significant effect, though, the result would not be valid and should be classified as a "false positive", *i.e.,* registering an effect, even though in reality there is none.[41] The important note is that the researcher directly determines the probability of such false positive statements by choosing the level of significance, which directly relates to the level of the threshold that needs to crossed.

Conventionally in the social sciences, the level of significance is chosen at 5% and/or 10%. This means that for every given question we ask of the data, there is a 5% or 10% chance of finding a statistically significant effect, even though in truth there is none. In other words, we should expect one in every 20 or 10, respectively, statistically significant results to be a false positive if the null is true.

Figure Ay: Statistical inference

		Truth	
		STiR has an effect	STiR has no effect
Test	Estimate stat. Sign. diff. from 0	☺	False Positive α
	Estimate **NOT** stat. sign. diff. from 0	False negative, (1-Power)	☺

Statistical inference: asking multiple questions of the data

When we collect data, we usually want to ask multiple questions of it: for example, we want to ask separate questions for each of the child-friendliness indicators discussed in this report; did STIR's program have a positive effect on each of these indicators?

Asking multiple questions of the data, therefore – by construction – increases the probability of finding at least one statistically significant result. For example, asking seven questions from the same data at the 5% level of significance, increases the chance of finding at least one false positive from 5% to about 35%, thereby weakening the confidence in our results. If we do multiple comparisons, the probability of finding at least one false positive is: $1-(1-\alpha)^m$, where m is the number of tests.

As a general remark, note that uncorrected results that are not statistically significant are expected to remain not statistically significant after correction for multiple inference, since multiple inference largely adjusts for false positives.

[41] Obviously, this is a hypothetical situation because the truth is never observed.

Appendix A17: Comparison of multiple inference corrections

In this evaluation we examine several outcomes (grouped into families) in accordance with exploring different aspects of the theory of change. Asking many questions (or testing many hypotheses) in this way increases the risk of finding many 'false positives,' which lowers our confidence in the results. Said another way, asking multiple questions of the data – by construction – increases the probability of finding at least one statistically significant result. For example, asking seven questions from the same data, at the 5% level of significance, increases the chance of finding at least one false positive from 5% to about ~30%[42], thereby weakening the confidence in our results.

The evaluation literature suggests several ways in which we could correct for multiple inference. First, we could aggregate similar questions into one summary question, thus reducing the overall number of questions asked. Technically speaking, all outcomes measures of an outcome family would be aggregated into a single outcome index, thereby reducing the number of tested hypotheses. One advantage of this approach is that it does not reduce the power to detect truly existing effects. At the same time, however, it would be difficult to disentangle the underlying changes in the outcome variables though. As an example, if time-use outcomes are seen to improve, we would not be able to state whether this is due to reduced time off-task or increased time spent teaching. Given the learning nature of the evaluation, we choose not to aggregate individual outcome measures into outcome indices since we felt the underlying changes would be ultimately useful for STIR.

Second, it is possible to adjust statistical inference for multiple hypothesis testing by controlling the *Family-Wise Error Rate* (FWER), *i.e.,* the probability of finding at least one false positive for a given set of questions or "outcome family". We considered a few different approaches for the same:

- **The Bonferroni correction** (Bonferroni, C.E 1936)**:** A popular and simple approach used extensively in the literature is the Bonferroni correction, which adjusts the level of significance for the total number of questions asked of the data in each outcome family. The Bonferroni correction compensates for that increase by testing each individual hypothesis at a significance level of α/m where α is the desired overall alpha level and m is the number of hypotheses.
- **The Sidak correction (***Šidák, Z. K. 1967***):** The Sidak correction also adjusts the level of significance for the total number of questions asked of the data in each outcome family. Given m different null hypothesis and a familywise alpha level of α each null hypothesis (of no impact) is rejected that has a p-value lower than $1-(1-\alpha)^{1/m}$.
- **The *Free Step Down Resampling Method* (FSDRM)**: Westfall and Young (1993) proposed this correction method for less conservative multiple testing procedures. Their method proves to be less conservation since it takes into account the dependence structure between test statistics (Ge Y, Dudoit S and Speed T.P. 2003). The FSDRM consists of the following steps (Anderson 2007):
    - Step 1: Run the original regressions and compute the p-value simulation storage counters
    - Step 2: Impose monotonicity in the p-values (increasing p-values)
    - Step 3: Calculate a set of simulated p-values using a simulated treatment assignement variable

[42] If we do multiple comparisons, the probability of a false positive is: $1-(1-\alpha)^m$, where m is the number of tests.

- Step 4: Enforce the original monotonicity of simulated p-values
- Step 5: Calculate the adjusted p-value and force monotonicity one last time

- **The *Holm-Bonferroni Correction*** (Holm, S 1979)**:** The Holm-Bonferroni correction was one of the earliest usages of stepwise algorithms in simultaneous inference. It is a modification of the Bonferroni correction. The formula to calculate the Holm-Bonferroni is ((α/(n- rank number of pairs (by degree of significance) +1)) where α is the target alpha level, n is the number of tests and rank is the order of the p-values (sorted from lowest to highest)[43].

While the Bonferroni correction is a popular approach, the downside of the Bonferroni correction is it is potentially lower powered to detect existing effects and overly conservative if the considered outcomes are correlated. The Sidak correction will always be less conservative than the Bonferroni correction and hence will always be the better option. But for low values of m (as it is for our evaluations); it is not that much less conservative and hence does not change the chances of us rejecting the null massively. The Holm-Bonferroni is a more rigorous approach as compared to the Bonferroni and the Sidak corrections and is considered as a more powerful approach.

The FSDRM is the more powerful method and is our preferred form of correction. However, we are able to use the FSDRM only for specifications where we compare STIR schools (core and core-plus taken together) to comparison. The complications associated with replicating the FSDRM to the core *v.* comparison and core-plus *v.* comparison specifications are as follows:

- Firstly, with a three arm RCT (as with ours), it would be straightforward to conduct a randomization inference based test (such that the FSDRM) that there is no difference between all three arms. It is however more complicated to conduct a test that there is no difference between treatment arm 1 and comparison or that there is no difference between treatment arm two and comparison.
- Secondly, steps of the FSDRM, (among other things) include generating a fake randomization (or iteratively reassigning the treatment assignment) and obtain p-values for all the tests. Given the way our randomization assignment was done (especially in Delhi) it would be tough for us to clearly identify the comparison schools that map with a particular core plus flavor or with the core treatment schools. This is because as a first step schools are assigned to treatment and comparison; after which the treatment schools are further divided into core and core-plus clusters. This second step is not done correspondingly for comparison schools (please see Appendix A5 for details).

Furthermore, given the complexities of the FSDRM, this method is only applicable to the school-level results and is not applicable to the IV/LATE approximations of the teacher level effect of STIR programming.

In order to maintain consistency within different specifications, our approach will be to use the FSDRM for all-STIR *v.* comparison school-level specifications (along with sub-group analyses) and the Holm-Bonferroni correction for all core *v.* comparison and core-plus *v.* comparison specifications as well as all IV/LATE analysis. Table 1 below organizes the

[43] http://www.statisticshowto.com/holm-bonferroni-method/ provides a good summary of the steps associated with the Holm-Bonferroni correction.

school-level outcomes into related groups, colloquially termed "families", and summarizes the correction method used across each of the families.[44]

[44] Note that corrections for the IV/LATE teacher-level analysis have been corrected using the Holm-Bonferroni method. Corrections for this analysis follow from the school-level, namely the two outcome groups corrected are: (1) classroom practice: quality (main, secondary, & sub-groups) (2) professional mindsets & behavior (sub-groups). Also note, that for the observational analysis, we did not conduct multiple hypothesis correction given our skepticism on the validity of these results.

Table 2: Summary of multiple hypothesis correction - School-Level analysis[45]

Outcome Group	Family	Outcomes within family	Treatment (all, std., or exploratory)	# of hypotheses tested	Type of correction used
Student Testing	Math (main)	Standardized Math Score	all-STIR	1	*Not corrected*
	Math (secondary)	Standardized Math Score	standard + exploratory	2	*Not corrected*
	Math (sub-groups)	1. Gender 2. Baseline teacher motivation 3. Teaching experience	all-STIR	3	*Not corrected*
	Hindi (Main)	Standardized Hindi Score	all-STIR	1	*Not corrected*
	Hindi (secondary)	Standardized Hindi Score	standard + exploratory	2	*Not corrected*
	Hindi (sub-groups)	1. Gender 2. Baseline teacher motivation 3. Teaching experience	all-STIR	3	*Not corrected*
Professional Mindsets & Behavior (PMB)	PMB (main)	1. Self-assessed total PMB score (standardized) 2. Positive Professional outlook index 3. Teacher growth mindset index 4. Teacher efficacy index	all-STIR	4	*Not corrected*
	PMB (secondary)	1. Self-assessed total PMB score (standardized) 2. Positive Professional outlook index 3. Teacher growth mindset index 4. Teacher efficacy index	standard	4	*Not corrected*
	PMB (secondary)	1. Self-assessed total PMB score (standardized) 2. Positive Professional outlook index 3. Teacher growth mindset index 4. Teacher efficacy index	exploratory	4	*Not corrected*
	PMB (sub-groups)	4 PMB outcomes X 3 sub-groups (gender, baseline teacher motivation, teaching experience)	all-STIR	12	**FSDRM**
Classroom Practice: Quantity	Time use (Main)	1. Time spent teaching 2. Time spent off-task	all-STIR	2	*Not corrected*
	Time use	1. Time spent teaching	standard	2	*Not corrected*

[45] Note, outcome families were created separately for the two geographies, Delhi & U.P. We did not correct across geographies.

	(secondary)	2. Time spent off-task			
	Time use (secondary)	1. Time spent teaching 2. Time spent off-task	exploratory	2	*Not corrected*
	Time use (sub-groups)	Time spent teaching X 3 sub-groups (gender, baseline teacher motivation, teaching experience)	all-STIR	3	*Not corrected*
	Time use (sub-groups)	Time spent off-task X 3 sub-groups (gender, baseline teacher motivation, teaching experience)	all-STIR	3	*Not corrected*
Classroom Practice: Quality	Child Friendliness (Main)	1. Whether teacher laughed or smiled 2. Whether students asked at least 1 question 3. Whether teacher used local information in teaching 4. Whether teacher used learning aides 5. Whether teacher used group work 6. Whether teacher referred to students by name 7. Whether teacher praised or showed-off students work	all-STIR	7	**FSDRM**
	Child Friendliness (Secondary)	*Ibid.*	Standard	7	**Holm-Bonferroni**
	Child Friendliness (Secondary)	*Ibid.*	Exploratory	7	**Holm-Bonferroni**
	Child Friendliness (Sub-group)	7 Child friendliness outcomes X 3 sub-groups (gender, baseline teacher motivation, teaching experience)	all-STIR	21	**FSDRM**

IDinsight
DATA. DECISIONS. DEVELOPMENT.

Appendix A18: Deviations from the Commitment to Analysis and Reporting Plan (CARP)

Our CARP acts as a guideline to our analysis. This document includes details on data collection, sampling, main research questions and proposed analytical models. All these details are already present in the report (please see section 4 for details). By prespecifying the hypotheses and clearly laying out the analytical strategy, CARPs (or pre-analysis plans) ensure research integrity and transparency and ensure the data are being used to test hypotheses backed by theory rather than creating hypotheses from the data itself; thus preventing the tendency to 'data mine'. In different parts of the report we have mentioned where we have deviated from the CARP. We provide her a detailed summary. We deviated from the CARP in three cases:

1. **When we felt the need to inspect into something more than we had initially planned.** This would be in case our initial understanding was weak or upon looking into something we felt we had to give it more thought. In our case this happened with our attrition analysis and our multiple hypotheses correction strategy.
    a. **Attrition:** We had initially proposed looking at differential attrition only across treatment arm. However, we felt to truly understand the implications of the high attrition on our estimates we would need to understand trends in the attriting teachers and students better and also look at impact on our power. We conducted two additional analyses to those specified in the CARP.
        i. We conducted balance tests of baseline characteristics for teachers surveyed at midline (Appendix A14).
        ii. We ran power calculations for our main specification to understand firstly what is the minimum effect we are powered to pick up and secondly how that has changed from what was initially proposed at the start of the evaluations. Please see Appendix A15 for details.
    b. **Multiple hypotheses correction:** Our initial understanding of multiple hypotheses correction was less. We had proposed to use the Bonferroni correction in all cases (please see Appendix A22 for details). However, as we looked into it more deeply we felt that the Bonferroni correction while the most popularly used, was the least 'powerful' correction method. In our final analyses, we used the Free Step Down Resampling Method (FSDRM) for correction of all-STIR *v.* comparison specifications and the Holm-Bonferroni for all core *v.* comparison and core-plus *v.* comparison specifications.

1) **When we felt we should conduct a piece of analysis differently than specified:** This only happened for the student testing analytical model. We had initially thought of the ordered logit model to see if STIR's program affects probability of a student being at a particular learning level. On fitting the ordered logit regression, we hoped to see the marginal effect of treatment on the probability of child being at a certain learning level (in both Math and Hindi). The ordered logit model helps us ask the following question of our data: *"Does being a part of STIR's program increase the probability of a child being at a certain level of learning as compared to a child in the comparison group?*". However, we felt it would be better to switch to an OLS estimation which would give us the general effect of the program on student learning. This was done for two reasons. Firstly, we felt this would help us make a cleaner learning statement for STIR. Secondly, based on our

understanding of STIR's ToC, we did not feel the program specifically targeted (or was more oriented to) students with different learning levels Eg: we would not expect STIR's program to benefit students who can read letters any differently or any more than those students who can read words.

2. **When we felt the need to conduct more analyses:** After presenting initial analyses, STIR requested for additional analyses they felt would help them learn more about their program. These were not specified in our CARP initially. These are mentioned below:
    a. **Adding a new sub-group:** After viewing initial results, STIR felt that they may be able to learn something about the importance of senior education stakeholder support and implementation context in U.P. by viewing the results divided by block. STIR thought this would be particularly useful in trying to separate out the influence of program design versus implementation capacity and delivery context, which the hypothesis that administrative units with more supportive BEOs and other local officials would show stronger results. Note: initially we had also planned to use STIR's network health indicators for the subgroup analyses. However, this was not done and the block level analyses were conducted instead.
    b. **Teacher level analyses:** As mentioned in section 4 in the report, we felt given STIR's program design and emphasis on systemic change, the school level estimates would be the best reflection of impact. However, given the great learning importance of the teacher level estimates to STIR, we conducted additional analyses to try and get at the teacher level impact (please see section 4.5.1.2 for details). This was done in two ways:
        i. **IV/ LATE:** We also pursued the IV/LATE strategy to the and get at the 'treatment-on-the-treated' effect. 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. 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 conducted this analysis for all indicators and subgroups in both geographies.
        ii. **Observational analysis:** On STIR's request, we also directly compared teachers in the treatment group who participated in STIR with teachers in the comparison group, excluding teachers in the treatment group who didn't participate in STIR. This analysis is an attempt to a relationship between participating in some amount of STIR programming and the outcomes for different teachers. This analysis was conducted for all indicators and subgroups for both geographies.

Appendix A19: Evaluation Approach

Randomized evaluations and causal claims

Researchers and implementers making use of impact evaluation methodologies seek to answer the question: to what extent can changes in outcomes of interest, if any, be credibly attributed to a specific program?

Making a claim about causal impact is always a comparative question: causal claims require comparing, in some way, what happened to those schools (in this case) that received STIR's programming ('treated') with what happened in those that did not ('comparison'). Schools that did not receive[46] STIR programming provide an answer to the question "what would have happened in similar schools over the same period in the absence of the program?"

The impact estimate is, at root, a comparison in the average outcomes between schools assigned to receive the program and those that were not. Said another way, the impact estimate for an outcome of interest is the change in that outcome in treated schools relative to changes in comparison schools.

One way of rigorously evaluating the impact of a program is to use a randomized strategy for assigning treatment (the offer of STIR's programming to teachers in a school) – that is, randomly selecting the schools with which STIR will work during the evaluation period from among all schools in the geography of interest. The goal of using random assignment is to create a starting point (baseline measurement) in which, on average, teachers in treatment and comparison schools are sufficiently similar (or 'balanced') on characteristics deemed relevant to how the intervention is expected to work in schools.

Formative approach to evaluation

While randomized evaluations are a rigorous and often rigid methodology, we have worked with STIR to provide as flexible and formative an evaluation as possible within the requirements of rigor. There are two key rigidities however with the design. First, those schools assigned to receive (or not) STIR programming (the 'treatment') cannot be reassigned over the course of the evaluation. This is fundamental to causal inference. Second, we are bound by our intended analyses, which is fundamental to going transparent and honest research. When we have gone beyond the scope of our planned analyses, this is clearly denoted in Appendix A18. This is not true with only randomized evaluations but with any analyses. Data should be used to test theoretically backed hypotheses. Building theory and hypotheses from observed patterns in data increases the risk of Type I errors or 'false positives'[47]. This reduces confidence in the results and increases the risk of making decisions on positive results driven by statistical noise.

[46] Whenever we refer to the comparison group we refer to schools where STIR's program was not offered. For readers familiar with RCT terminology, this is the same as the 'control' group.

[47] In statistical hypothesis testing, a Type I error is the incorrect rejection of a true null hypothesis (a **"false positive"**). Please see further section on multiple hypothesis correction (section 4.5.3) for a better understanding on the implications of false positives.

We have taken a formative approach to this randomized evaluation in three[48] main ways:

1. STIR has continued to iterate on its programming based on experiential learning over the course of this first year of the evaluation. We have also tried to contribute to this learning through a detailed process evaluation, reported elsewhere.
2. We have continued to refine our measurement metrics and indicators over the course of the evaluation so far. As we have learned more about our own instruments and STIR has gained clarity on their goals and indicators, we have worked to adopt or adapt the measures we used at baseline (made possible by the randomized set-up). We have attempted to include measures of attendance, which we and STIR now understand as an important link in the ToC. We have also adapted the classroom observation tool to ensure that indicators are closer to STIR's program. We have also added questions to the classroom observation and teacher motivation tools which map closer to the STIR program, to provide more useful information to STIR.
3. We explicitly designed this evaluation to test the differential effects of iterations of the program in Year 1, as described under "Program variations." This included looking a programming version focused largely on intrinsic motivation as well as a version that added in different extrinsic — but still non-financial — motivators. This represented a focus on structured, experiential learning and small experimentation within the structure of the larger RE set-up (Pritchett, Samji, and Hammer 2012).

[48] Looking forward to Year 2 of the evaluation, STIR decided that a different research objective was more pressing than intrinsic versus extrinsic motivation: whether teachers and students benefitted more from ready-made, evidence-informed approaches to classroom practice or from co-creating their own approaches to challenges faced. IDinsight has also accommodated this change in focus.

Appendix A20: Covariates in teacher level regression analyses

The control variables mentioned here[49] are those which we are not interested in directly but we expect are related to the dependent variable or may have differential experiences as part of the STIR program. We will include these in the regression equation as the independent variable to 'control' for their effect.[50] [51]

1. **Teacher sex:** Male and female teachers may be motivated by different factors or conversely different things may motivate male and female teachers differently. Further the access to, experience with the STIR program and the interest and ability to act on STIR's ideas may vary with sex.
2. **Age:** Teachers of different ages may find different barriers in being active members of STIR. They may be more susceptible to pressures from family, colleagues and supervisors. They may also have less decision-making power while trying to influence the classroom culture or practice. At the same time the desire to collaborate with peers may be more exciting to younger teachers.
3. **Teacher qualification level:** Teachers in schools have varied backgrounds and training. This may influence a teachers' ability to influence her classrooms in general, and specifically with regards to learnings from STIRs programs. A teacher with higher training or qualification may also be better at finding localized solutions to the challenges in the classrooms and may be more 'active' participants in network meetings.
4. **Total number of years teaching:** A teacher's experience may influence how they 'value' STIRs program. It may be the case that teachers younger in their teaching career are more incentivized to be a part of the STIR program that someone who is further ahead in their career or closer to retirement. At the same time, it may be that slightly more experienced teachers are better aware of the challenges specifically in their classrooms or in general. They may be able to use the STIR experience in a more fruitful manner.
5. **Enumerator dummies:** It is important to control for enumerator dummies, to prevent for any enumerator specific biases during data collection. While the motivation questionnaire is self-administered, bias may creep in due to an enumerators communication and explanation skills.
6. **Baseline teacher motivation index:** A teacher's inherent motivation may be an important determinant of how actively they are a part of STIRs network. STIRs program requires teachers to spend time and effort outside of classrooms in network meetings and to find solutions to existing problems. Along the way they may have to face barriers of different kinds, which may act, as disincentives *e.g.*, travel, pressure from head teachers, family etc. If a teacher is highly motivated, she would be more likely to overcome these hurdles or may be more excited by the opportunity to learn.

[49] Note apart from these we also control for baseline values where applicable

[50] From a technical point of view, there is no need to control for additional (teacher) characteristics in the analysis because of random assignment of the treatment indicator. We expect that inclusion of additional characteristics helps to explain overall variation in the outcome measures and thus increases precision of the estimates of interest.

[51] The respective control variable is excluded in the subgroup analysis for that variable.

7. **Dummy variable for network:** Teachers are organized into local changemaker networks in which they interact with teachers from other schools in the same network. Each network is led by one EL and contains schools with geographical proximity to one another.
8. **Class size:** For the time use family and the child friendliness indicators (*i.e.*, indicators from the classroom observation) we also use class size as a control variable (where class size is defined as the number of students present in the class at the time of the observation itself). The number of children in a class may affect a teacher's teaching style and influence the way she allocates time in the classroom.

Baseline outcome measures: Apart from the covariates mentioned above, we also include baseline outcome measures. We expect that outcome levels observed at midline to be explained to a large extent by corresponding baseline levels. For this reason, we do include baseline values of the outcome measure if available. For two outcomes of the child-friendliness outcome family and the attendance family baseline levels are not available since these data were not collected at baseline.

Appendix A21: Teacher-level estimates: 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 endline
- 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 in the two years
- $\text{partiallyActive}_{ij}$ 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 in the two years.
- 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. For teachers, these include teacher sex, age, qualification, years of experience, baseline teacher motivation, class size, enumerator and network dummies.
- ε_{ij} is an individual level (within schools) error term
- ω_j is a school level error term

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 control schools.

Appendix A22: Data quality assurance measures

Details of survey administration

The teacher motivation questionnaire was filled out by teachers on a paper form. They provide written consent before beginning the questionnaire. Data from this survey was entered (in double) by data entry operators which was then reconciled and checked by IDinsight staff.

All other data were collected electronically using SurveyCTO (SurveyCTO (version 2.02) 2016). Data were collected by a field team of surveyors hired, trained and managed directly by IDinsight staff.

In Delhi, STIR EL's help setting us up in schools by requesting schools for dates when our survey teams can visit. In U.P., STIR Program Managers help by setting us up with official government permission letters. Field planning and school entry is then done independently by IDinsight field managers.

As per field protocol, our field teams first interact with head teachers/ principals upon reaching the school. They explain the survey and its components, how the data will be used and ask for permission to proceed with our survey activities. After this, surveyors collect school level information and information required from the head teachers. They then interact with the teachers in our sample and explain the survey to them in detail before asking for permission to observe their class, interact with students etc. Teachers provide oral consent before we observe their classrooms; students provide oral consent before we administer the learning assessments.

TM Questionnaire

Several steps were taken to ensure utmost quality in conducting the collecting these data:

- The data, collected on paper forms, are double entered.
- IDinsight field managers enter a randomly chosen ten percent portion of the data. This acts as a triple check for our collected data.
- Form scrutiny: All enumerators thoroughly scrutinize the survey forms before they leave a school to check for any discrepancies in the answers and spot missing sections, absurd answers etc.
- Our error rates were found to be less than 0.5. For teacher codes, the acceptable rates for data entry error were 0.

Classroom practice and student learning tools

Regular spot checks and discussions

In both Delhi and U.P. IDinsight field managers conducted daily debriefing sessions with the entire field team. During these sessions the field managers would bring up scenarios they had noticed during their visit to schools in the day and would ensure all enumerators were in agreement of how the scenario was to be interpreted.

Back checks

IDinsight field managers also conducted random back checks, in which they would visit schools and interact with principals about the visit of the enumerators.

GPS coordinates

Every time an enumerator would open a new form on SurveyCTO, his or her GPS coordinates would be recorded automatically. This allowed IDinsight staff to verify that enumerators physically visited the schools they were assigned.

Other SurveyCTO Checks

To help identify and prevent enumerators from collecting incorrect data, a number of checks were incorporated within the SurveyCTO survey form. Most of the checks were built-in to appear randomly. This prevented enumerators from finding loopholes.

Some of the checks built in were:

- **Random selection screens:** During the student testing and classroom observation often a screen would appear which prompted enumerators to select to ensure they were attentive during data collection.
- **Photos:** At the end of student tests, the forms would prompt the enumerators to take a photo of the student and their answer sheets. This was a good check to ensure that enumerators did not falsify the student learning data and that the survey was conducted.
- **Audio files:** The most important check incorporated was to randomly record audio clips of enumerators in the schools. These clips were recorded during both the classroom observation and the student testing. These audio clips were regularly audited by IDinsight field managers, who not only used them to check the enumerator performance but also re-coded the survey themselves. For example, if a field manager was reviewing a student testing audio clip then he or she would fill a separate survey form based on how they would rank the students' learning level based on the students' answers. This also helped identify those enumerators who had not followed field protocols and survey rules as discussed during training.
- **Time stamps:** Each section of the survey conducted had an associated time stamp. This time stamp was analyzed as part of the data audits to check how long enumerators spent on each section. This was particularly useful in checking for falsification of data in the timed sections of the classroom observation.

Appendix A23: Contexts of the evaluations

STIR's program in Delhi and Uttar Pradesh have distinct implementation and institutional contexts. STIR thinks of their private schools 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 adequate training and support from STIR staff. The specific study sites are shown in

Figure Az: Northern India and surround, showing Delhi (n=1) and U.P. (n=2) evaluation sites

The geographic, state and implementation differences between Delhi and U.P. will help inform strategic decisions for STIR about the systems and models with which they engage. Working with private schools (APS), on one hand, means engaging with each school (a small business) individually, as there is no overarching authority; for delivery, STIR relies on its own employees to serve as facilitators of programming as well as liaisons with each school. Working within the government system (U.P.), on the other hand, means having an overarching authority through which to introduce STIR into the system; for delivery, STIR relies on a cascade model, using its own employees to train and support government school teachers to act as volunteer, ground-level facilitators of STIR's programming.

Drawing on both the broader literature and descriptive data from the present study[52], in this section we provide some contextual detail for these study sites. While the present program and evaluations focus on teacher professional development, the background on infrastructure helps set the scene of the geographies and schools in which STIR's programming is implemented — where teachers need to teach and feel inspired and students need to learn and

[52] We describe the methodology in section 4.

feel comfortable.[53] We present the descriptive data from this evaluation in Table A30 to shed light on the teaching and learning context for this evaluation.

Table A30: Facilities details of evaluation schools in Delhi (private schools) and U.P. (government schools)

Indicator	Percentage of U.P.[54]		Percentage of Delhi	
	Yes	No	Yes	No
Does the school have a boundary wall?	*64%*	*36%*	*100%*	*0%*
Does the school have a separate kitchen?	*97%*	*3%*	*27%*	*73%*
Do classrooms have desks for students?	*14%*	*86%*	*97%*	*3%*
Does the school have an electric connection?	*44%*	*55%*	*99%*	*0%*
Does the school electricity work?	*26%*	*74%*	*99%*	*0%*
Does the school have a toilet in working condition?	*95%*	*5%*	*100%*	*0%*

2.3.1 Delhi private schools

There exist very few (reported) data on private schools in Delhi. Here, we draw mostly on our own sample to paint a picture of the private school context. In our sample, in private schools in East Delhi, most schools have a boundary wall to distinguish school property from the space beyond it. Most also have an electric connection which was working at the time we visited the school and have working toilets. However, three-quarters of schools do not have an included kitchen in which, *inter alia*, mid-day meals can be prepared.

Inside the school, most (97%) classrooms have desks for students. Most classrooms (95%) also had tables or desks for the teacher and had a blackboard. About 45% of classrooms had posters on display and in about 25% of classrooms, outside noise was audible. These descriptive findings for our sample are summarized in Table A31.

[53] We selected the specific indicators reported here give their link with India's Right to Education Act (*The Right of Children to Free and Compulsory Education Act* 2009, vol. DL-04/0007/2003-09, sec. 19). Some researchers have found links between adequate infrastructure, classroom facilities, and student learning outcomes (Govinda and Bandyopadhyay 2011).

[54] The total number of schools in Delhi is 135 and Uttar Pradesh is 266. For a few indicators (eg: electric connection), information is missing since enumerators were unable to clearly observe these in a few schools due to either the physical layout of the school or lack of permission.

Table A31: Facilities available in classrooms in Delhi (private schools) and U.P. (government schools)

Indicator	Percentage of U.P.[55]		Percentage of Delhi	
	Yes	No	Yes	No
Are most children in the classroom wearing uniforms?	74%	26%	98%	2%
Does outside noise affect communication in the classroom?	3%	97%	25%	75%
Does the classroom have a blackboard or whiteboard?	97%	3%	98%	2%
Is there a chair and/or a table for the teacher?	96%	4%	95%	5%
Are there posters etc. on the wall or on display (other than student work)?	63%	37%	45%	55%

2.3.2 U.P. government schools

In terms of school infrastructure, government schools have not sufficiently met the standards required by the Right To Education Act (RTE 2009)[56] be it in terms of toilets (lack of which has been often quoted as reasons for girl students dropping out), proper classrooms for teaching students, a secured school covered by boundary walls etc. Turning to our sample, as shown in Table , in the U.P. districts of Rae Bareli and Varanasi included in the present evaluation, about 60% of schools are demarcated from the surrounding area by a boundary wall. Most of the schools have both working toilets and an on-site kitchen. However, only about one-quarter of the schools have a working electric connection. Most classrooms have a blackboard and desk/table for the teachers but less than 15% have student desks in the classrooms. Above 60% of classrooms have posters on display.

[55] These data were collected in 459 classrooms in Delhi A.P.S and 747 classrooms in U.P. Govt. schools

[56] http://mief.in/condition-of-government-schools-in-india-a-shocking-truth/; http://www.livemint.com/Politics/h7WkzI77bMtmN9FLDvyo0M/The-poor-state-of-school-infrastructure.html provide details on situations on government schools after the RTE act of 2009.

Appendix A24: Sample size and power calculations

4.5.2 Sample size calculations

Our initial power calculations were done at two levels – the student and the teacher level. Note however, that roughly the same MDE's were assumed at both the teacher and student level. This was due to lack of information of distribution of indicators at the teacher level. All the teacher level indicators used for the studies are relatively unexplored in the literature.

Table A32: Initial power calculations as in the proposal to SIEF

	Student level	**Teacher level**
	Power (κ)	**Power (κ)**
	0.8	**0.8**
Number of clusters (schools) (per arm)	60	60
Cluster size	30 (10 students / teacher)	3
Intracluster correlation	0.22	0.10
R^2 of outcome variable	0.42	0.49
Significance level (α)	0.05	0.05
Sample size (per arm)	1,800	180
Minimum Detectable Effect	0.20 sd	0.23 sd

Student level

At the stage of the initial power calculations, based on the literature, we expected an impact on test scores of 0.2sd.

Teacher level

At the teacher level, given that teacher motivation was the first link in STIR's theory of change our power calculations at the teacher level focused on this. We used preliminary (pre-post) data and expected larger changes as compared to the student level. We expected impact of greater than 0.3 sd. However, given that we had based it on limited information, we chose to be powered to pick up a smaller effect size (0.23 sd)[57].

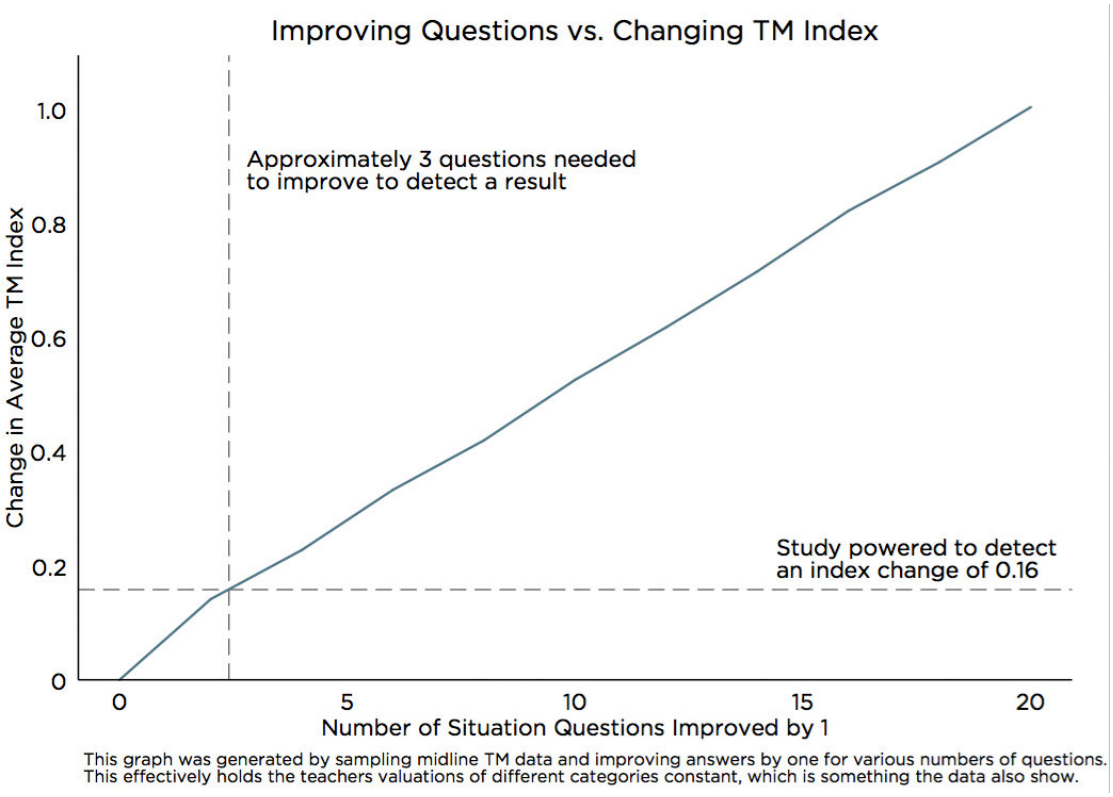
Our randomization strategy, sampling strategy, and sample size were designed to confidently detect differential impacts on the motivation index score between (1) comparison schools and those assigned to the core variant and (2) between comparison schools and those assigned to

[57] STIR and IDinsight had considered increasing the number of schools (which would help us pick up much smaller effect sizes). However, given budget and operational constraints it was not feasible.

the core-plus variant in each study context.[58] However, differences between core and core-plus will only be suggestive, as will comparisons among the different flavors of the core-plus model. Causal comparisons across studies — that is, between Delhi and U.P. — are not intended.

To link a change in the teacher motivation index with changing responses on the questionnaire, consider the 'situation' items that as a teacher to dis/agree with statements about how they feel. If at baseline, a teacher marked "disagree" for three items and, at midline, marked "agree" for those same items, we would be able to detect this change. Indeed, we are powered to capture any three-point change on the situation Likert scales. This is visualized in Figure Ah(note that we end up being better-powered than anticipated) which shows that a 'one-point' improvement in three questions (x axis) would lead to roughly a 0.16 change in index score (all other answers being held the same), which we would be 'powered' to capture.

Figure Ah: Depiction of analytic power to capture changing answers on the Teacher Motivation Questionnaire



Effect of Attrition

Given the high levels of attrition at the teacher and the student level, we ran power calculations at the end of midline data collection to understand the effect size we were (still)

[58] We establish a link to the literature that documents influences on motivation outcomes in Appendix A6.

powered to pick up[59]. We used the 'clustersampsi' command for the same to 'solve' for MDE.

```
clustersampsi, detectabledifference rho() alpha() k() m()
mu1() sd1()
```

The estimates for ICC (rho in the equation above), Number of clusters (in our case school; k in the equation above), Average cluster size (in our case average number of teachers and students per school; m in the equation above), mean and standard deviation (mu1 and sd1 respectively in the equation above) all came from our collected data. Significance level (alpha) was fixed at 5%.
In a way our power calculations allowed us to replace assumed values of ICC, mean, standard deviation with values from the dataset and allowed us to account for attrition at the school and teacher/ student level by fixing k and m parameters based on number of schools and teachers/ students actually surveyed.

The table below provides an indication of the effect sizes we are powered to pick up (80% power; 5% level of significance):

Table A33: Minimum detectable effects as for midline analyses

Indicator	Minimum Detectable Effect (Delhi private)	Minimum Detectable Effect (U.P. gov.)	Units
Teacher motivation index	0.24	0.24	Standard deviation
Teaching	0.12	0.05	Percentage points
Off task	0.04	0.03	Percentage points
Teacher laugh, smile	0.15	0.09	Percentage points
At least 1 question	0.15	0.08	Percentage points
Local information	0.16	0.1	Percentage points
TLM	0.17	0.13	Percentage points
Group work	0.06	0.05	Percentage points
Refer name	0.13	0.13	Percentage points
Praised, work displayed	0.13	0.07	Percentage points
Hindi levels	0.35	0.25	Standard deviation
Math levels	0.36	0.22	Standard deviation
Teacher in school	NA	0.07	Percentage points
Teacher in class	NA	0.08	Percentage points
Student attendance	NA	0.19	Percentage points

[59] There are often concerns with *post-hoc* power calculations and 'why' researchers run them. Please see Gelman, Andrew, and John Carlin. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9.6 (2014): 641-651 for details. Our motivation to rerun the power calculations was to purely see what effect size we were powered to pick up to help interpreting significant and non-significant results alike.

We find that for the teacher motivation index, attrition does not affect our minimum detectable effect by a lot compared to initial calculations (we are still powered to pick an effect size of 0.24 sd; compared to the 0.23 sd effect initially proposed).

At the student level, the effect of attrition on minimum detectable effect is a bit more profound yet not too worrying. This also varies with geography. In Delhi private schols, we are powered to pick up effects larger than 0.35 and 0.36 *sd*s for Hindi and math respectively. This is quite different from our initial proposed 0.2 *sd* minimum effect. In U.P. however, we are powered to pick effects of 0.25 *sd* and 0.22 *sd* (and above) in Hindi and Math respectively.

For all other indicators[60], there is variation within 'family' of indicators and across geographies. For example, in U.P. we are powered to pick up even a 5 percentage points change in teaching whereas in Delhi private schools, we are powered to pick up a 12 percentage points change for the same indicator.

In summary, the magnitude of attrition has limited effects on our power. As compared to the effects the study was initially planned to detect, we do not see much change at the teacher level. At the student level, we do find a difference for Delhi private schools, in particular, while in U.P. government schools the change is less dramatic. We will continue to work to minimize attrition at endline, but we sfeel confident in the evaluation's ability to answer the questions with similar power as we had initially set out with.

[60] We could consider creating a child-friendliness index at endline to help with being powered to pick up a lower effect at the aggregate level. Similarly, for Delhi endline we could consider an aggregate learning score of sorts to help pick up smaller effect sizes. We will discuss this further with STIR to understand if it would benefit their learning.

Appendix A25: Association Between Teacher Characteristics & Student Test Scores

The goal of this section is to provide suggestive evidence on the association between various teacher characteristics and student test scores at baseline. We focus on those teacher characteristics that feature in the simple Theory of Change depicted in Figure-1 of the main report. The rational for conducting this exercise is to critically examine the main linkages in the theory of change and try to unpack whether the teacher characteristics that STIR programming targets are correlated with student performance. The correlations presented are purely suggestive relationships and should not be interpreted causally.

Figure Ai: Relationship Between Baseline Motivation & Baseline Student Performance

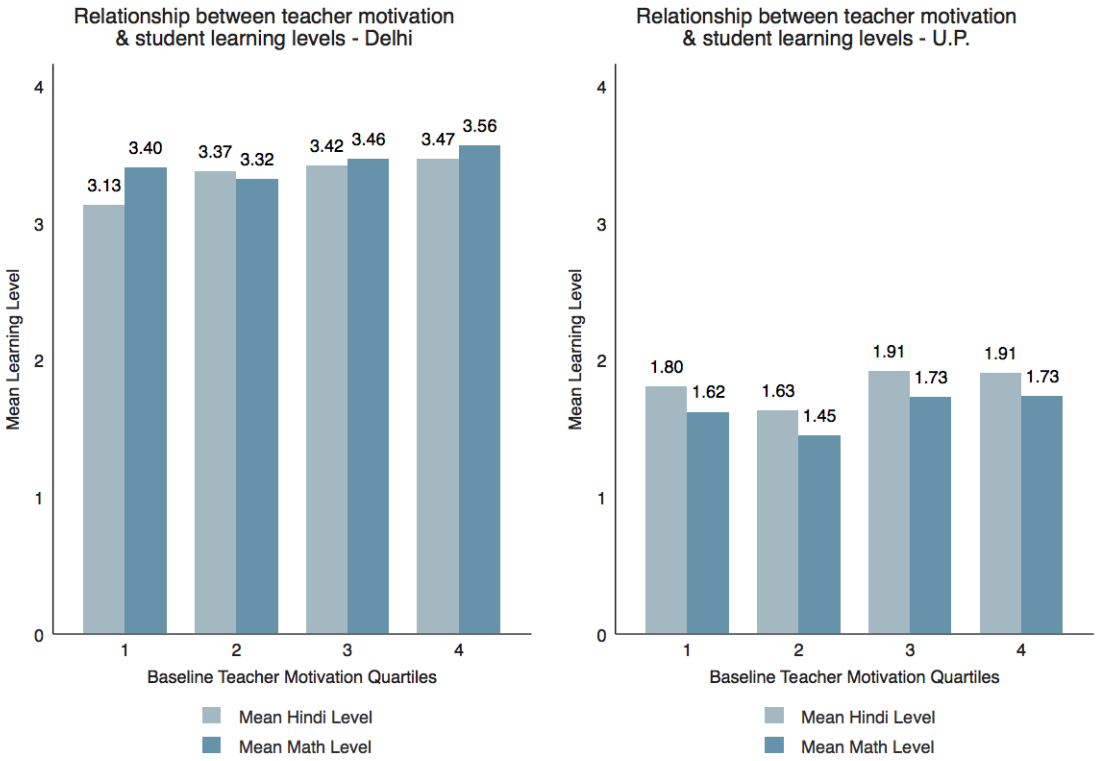


Figure Aj: Relationship Between Baseline Quantity & Quality of Teaching Practice & Baseline Student Performance (Delhi)

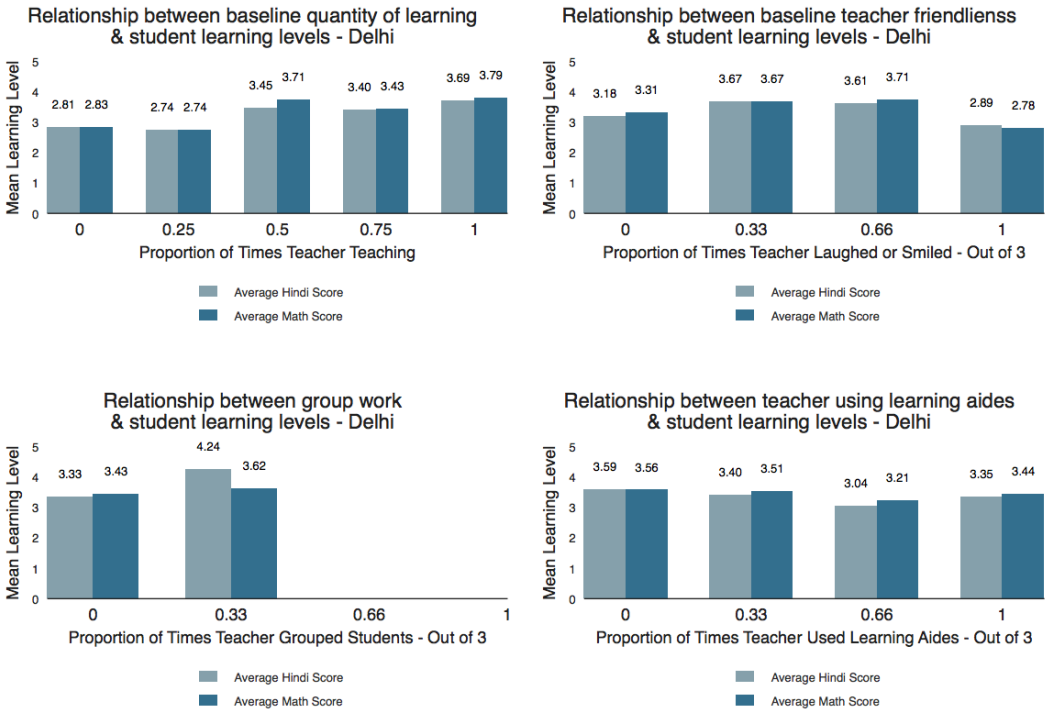


Figure Ak: Relationship Between Baseline Quantity & Quality of Teaching Practice & Baseline Student Performance (U.P.)