Great Expectations?
Leveraging Teachers to Improve Student Performance[*]

Minahil Asim [†] Ronak Jain [‡] Vatsal Khandelwal [§]

Do high teacher expectations improve student performance? Theoretically, expectations can motivate students by raising aspirations or discourage them if perceived as unrealistic or evoking unfavorable peer comparisons. To causally identify their effects and unpack mechanisms, we randomize whether students receive expectations framed as *attainable* or *ambitious*, are additionally paired with a classmate for encouragement, or only receive information about past performance. Communicating expectations increases math scores by 0.21σ, particularly for students receiving ambitious goals or predicted to perform poorly. Information about past performance has comparable effects (0.18σ), as students interpret it as attention and encouragement from the teacher. Pairing only helps when students are similar, indicating that interpersonal comparisons negatively affect motivation. Although students with large gaps between expectations and baseline performance show sustained gains 12–18 months later, the effects of expectations and information remain statistically indistinguishable. Overall, our findings highlight that signals of personalized teacher attention—through communicating expectations or performance—are a low-cost, effective input in the education production function.

Keywords: Expectations, Information, Peer Effects, Motivation, Performance.

JEL code(s): C93, D83, D84, D91, I24, I25.

[*]We are grateful to our partner schools for their collaboration and facilitation of this research, especially Scherezade Tarar and Farah Rehman. This research has been possible because of financial support from RISE Pakistan and JPAL-PPE. We thank the research team at the Center for Economic Research in Pakistan (CERP), especially Ummara Waheed and Waheed Ahmad for field management, and Absar Ali, Omar Zahid, and Lucas Borden for their exceptional research assistance. We are very grateful to Tahir Andrabi, Emily Breza, Christina Brown, Jishnu Das, Stefan Dercon, Asim Khwaja, Simon Quinn, and Gautam Rao for their useful comments and suggestions. We also thank seminar participants at Pacific Conference for Development Economics (PacDev), Association for Education Finance and Policy (AEFP) Annual Conference, LEAPS Pakistan Conference, the AEFP EdDev Community Group, Lahore University of Management Sciences, Harvard University and the University of Oxford. IRB approval was obtained from Harvard University and the University of Oxford. Our partner school chain's internal research ethics board also reviewed and approved our research protocols and consent forms in November 2020 in addition to our university Institutional Review Boards. The RCT was pre-registered as AEARCTR-0007846 at `https://www.socialscienceregistry.org/trials/7846`.

[†]University of Ottawa. Email: minahil.asim@uottawa.ca

[‡]University of Zurich. Email: ronak.jain@econ.uzh.ch

[§]University of Exeter. Email: v.khandelwal@exeter.ac.uk

1 Introduction

We face a global learning crisis, with millions of children lacking basic literacy and numeracy skills despite being in school (World Bank, 2017). This underscores the need to better understand how different inputs in the education production function affect learning outcomes. While teacher quality is a critical input (Chetty et al., 2014*a*,*b*), much less is known about which specific teacher practices most effectively raise student achievement (Bau and Das, 2020). One potentially promising but understudied tool is the communication of high teacher expectations.[1] Education psychology has long hypothesized that expectations can shape student outcomes through a self-fulfilling mechanism known as the "Pygmalion Effect" (Rosenthal and Jacobson, 1968), and setting high expectations is also a defining feature of several schooling models, including US charter schools (Angrist et al., 2013; Fryer Jr, 2014).

Identifying the causal effect of expectations on student learning is challenging because expectations are endogenously formed and selectively conveyed. Their impact within the education production function is also theoretically ambiguous. First, high expectations can motivate students by raising aspirations (Papageorge et al., 2020) or discourage them if perceived as unrealistic (Friedrich et al., 2015). Second, their influence may depend on whether students have supportive peers (Bursztyn et al., 2019). Third, they may add little value beyond information about performance (Andrabi et al., 2017; Barrera-Osorio et al., 2020) which can itself signal teacher attention and care.

In this paper, we causally identify the effect of expectations on academic performance and disentangle these channels through a randomized controlled trial. Students are randomized to receive personalized messages on behalf of their teacher containing: expectations framed as *attainable* or *ambitious* goals, additional peer matching for mutual encouragement, or only information about their past performance. We find that both expectations and information improve math scores by 0.21σ and 0.18σ, respectively, while pairing students only helps when matched peers are friends or similar in baseline characteristics. Importantly, the gains from expectations are largest among low-performing students and those randomly assigned more ambitious targets. Information about past performance has comparable effects on average as students interpret this as attention from the teacher. Notably, while students with large gaps between baseline performance and expectations continue to show sustained gains 12–18 months after the intervention, the effects of expectations and information remain indistinguishable. Our findings suggest that students are highly responsive to personalized communication from teachers, establishing personalized teacher attention as an important input in the education production function.

We partner with a large private school chain in Pakistan to implement a clustered randomized controlled trial across 288 classrooms. Teachers are held in high regard in our setting,

[1] In this context, the word 'expectation' is not used in the statistical sense of a prediction but rather in its commonly used form to indicate desired outcomes.

as in many other contexts (Dozza and Cavrini, 2012; Wentzel, 2010), and their motivational role in the classroom is well-acknowledged by parents.[2] To collect realistic, student-specific teacher expectations, we first reminded teachers of the most recent math test scores of each of their students.[3] Then, all teachers were asked to fill in responses to the following statements: (1) "I expect the student to work hard and improve to *achieve at least X (out of 100%)* in upcoming exams and tests" (which we refer to as 'High Expectations' hereafter) and (2) "I expect the student to work hard and improve and I think that *even Y (out of 100%) is achievable* in upcoming exams and tests" (which we refer to as 'Very High Expectations' hereafter). These statements were designed to elicit a lower bound and an ambitious score that the teacher expected the student to achieve. [4]

Following the collection of expectations from all teachers, we divided the classrooms randomly into three groups: an Expectations Arm, a Peer Arm, and a Comparison Group, which was further split into an Information Arm and a Pure Control Group. Students in the Expectations Arm randomly received one of the two teacher expectations statements ('High Expectations' or 'Very High Expectations'), along with a custom-designed and individually tailored infographic encouraging them to work towards achieving the expectation. Students in the Peer Arm received the same message as the Expectations Arm, but were also randomly matched with another classmate. They received an additional infographic highlighting that they should encourage each other to achieve their respective goals, which were privately disclosed to them. Students in the Information Arm received an infographic containing information on their performance in the previous math exam. Finally, students in the Control Arm received no message. To ensure privacy, we used the school's online infrastructure to send private emails to each student. Moreover, all teachers were blind to the treatment status of students to ensure they did not selectively change their efforts towards any students.

Our experimental design addresses key challenges in causally identifying the impact of teacher expectations on student outcomes and yields policy-relevant insights. First, we ensure that all teachers in our sample set high expectations for students in their classrooms before randomization. As expectations are endogenous and often selectively delivered, this step ensured that we could causally isolate the effect of communicating high teacher expectations on student performance. Comparison of the Expectations Arm to the Control Group, therefore, allows us to estimate the effect of communicating high teacher expectations. Moreover, we randomly vary how expectations are framed, i.e., as a lower bound ("...to achieve at least X") or as an ambitious target ("...even Y is achievable"). This allows us to test whether the magnitude and framing of expectations influence their effectiveness. Second, the Infor-

[2] For example, 84% of parents in our sample report that it would be 'very' or 'extremely' useful if the teacher set and conveyed a concrete expectation of performance to their child.

[3] This mitigates the concern that teachers might not know their students' performance (Djaker et al., 2024) or that expectations reflect stereotypes rather than ability. Indeed, we find no systematic differences in teacher expectations by student gender, wealth, or age (Figure A.4.1) in the elicited expectations, which provided the basis for the messages later conveyed to students.

[4] Teachers were informed that these expectations might be communicated to students.

mation Arm isolates the effect of simply informing students about their past performance, providing a natural benchmark against which to compare the Expectations Arm. Third, the Peer Arm measures whether pairing students for mutual encouragement strengthens the impact of teacher expectations. Since peers are randomly assigned, we can also examine whether peer characteristics, such as similarity or friendships, moderate treatment effects.

Our main outcomes include student-level scores in Mathematics and English on high-stakes exams. These tests are the same across all schools and were conducted 2-4 weeks and 6 months after the initial delivery of the intervention. In addition, we also conducted a follow-up student survey to collect information on how the students interpreted various components of the intervention. Finally, we collected administrative data on student test scores 12 and 18 months after the start of the intervention to study long-term effects.

The experiment generates four main sets of findings. First, we find that students in the Expectations Arm scored 0.21σ higher in their math exam than students who received no information (significant at 1%).[5] Decomposing this, we find that the treatment effect is significantly larger for students who randomly received the 'Very High' teacher expectation statement. The magnitude of the expectation also matters relative to the student's baseline performance: a 10 percentage point increase in the gap between teacher expectations and students' baseline score leads to a 2 percentage point increase in the treatment effect. Those predicted to perform the worst score significantly higher, showing that the effect is not concentrated among those who would perform well in the absence of the treatment. Reinforcing these results, our follow-up survey reveals that about 70% of students interpret the expectation infographic as a goal or encouragement from the teacher. Together, these findings suggest that students respond especially well to the ambitious expectations communicated by the teacher. In terms of cross-subject spillovers, we detect a modest reduction in English scores over time, implying that teacher expectations in only one subject can potentially crowd out effort in another.[6]

Second, we find that students in the Information Arm also score 0.18σ higher than students in the Control Group. Importantly, this effect is not significantly distinguishable from the effect of the Expectations Arm. This is especially surprising given that nearly 80% of the schools regularly send report cards, implying that the intervention did not provide new information. In fact, we find that most students interpret these messages as personalized attention from the teacher. This is consistent with recent evidence showing that college students respond positively to performance feedback, interpreting it as a signal that their professors care about them (Carrell and Kurlaender, 2023). We also discuss alternative mechanisms, including the novelty of the information, accessibility of format, and the timing of the delivery.

Third, the average treatment effect of the Peer Arm is statistically indistinguishable from zero.

[5] We pool results from the two rounds of math exams conducted 2-4 weeks and 6 months after the start of the intervention. The treatment effects for both rounds separately are identical and statistically significant at 5%.

[6] The observed treatment effects are not driven by alternative mechanisms such as changes in teacher behavior. We rule out these mechanisms in Section 6.

Although this treatment arm adds a peer-matching component to the Expectations Arm, we find that the effect of the Peer Arm is significantly lower than the effect of expectations alone. At the same time, we detect evidence of heterogeneous treatment effects (Chernozhukov et al., 2018). In particular, the effect of the Peer Arm is positive and significantly higher for students who are randomly paired with friends as per their baseline social networks. Similarly, we find that effects are significantly higher for peers who are similar to each other in terms of their baseline characteristics, including their scores and teacher expectations. In contrast, the effect is significantly negative for those whose matched peer has a higher baseline score or expectation than them. Consistent with this, in our follow-up survey, students report that they are likely to feel disappointed if their matched peer's teacher expectation is higher than their own. Conversely, the majority report that they would feel motivated and happy if paired with a peer with similar scores and teacher expectations. Overall, the results suggest that interpersonal comparisons can negatively affect student academic performance.

Finally, we also measure the effects of our intervention in the longer term, i.e., 12 and 18 months after the intervention. We do not detect a significant average treatment effect on Math or English test scores in either wave. However, we continue to find that the treatment effect of the Expectations Arm is significantly larger for individuals whose expectations exceeded their baseline performance. The fact that individuals who initially received a high expectation relative to their baseline performance perform significantly better several months after the intervention suggests that ambitious expectations may have a long-lasting motivational impact. We also rationalize our findings using a simple effort-choice model in which student effort depends on intrinsic motivation from teacher expectations, information, and peers.[7] The model helps interpret why expectations and reminders improve student achievement while peer matching has ambiguous effects.

Our paper makes three key contributions. First, we provide causal evidence on the effect of communicating teacher expectations and whether it varies depending on how they are framed. While the hypothesis that a teacher's expectations can affect student performance, leading to a self-fulfilling prophecy, is not new (De Boer et al., 2018; Friedrich et al., 2015; Papageorge et al., 2020; Rosenthal and Jacobson, 1968, 1992; Wang et al., 2018), it is difficult to causally identify their effect because expectations are typically endogenously formed and selectively conveyed (Carlana, 2019; Jussim and Harber, 2005; Papageorge et al., 2020). Our experimental design overcomes this challenge by eliciting student-specific expectations from all teachers and randomizing whether these were communicated to students. Additionally, our finding that ambitious expectations generate especially large gains alleviates concerns that high expectations can cause frustration.

Second, our results show that performance information can be as effective as expectations when students interpret it as a signal of teacher attention. Expectations and performance information are often bundled in successful schooling models like US Charter Schools (An-

[7] The model is presented in Section B in the supplementary appendix.

grist et al., 2013; Fryer Jr, 2014), programs like the Knowledge and Power Program (Gleason et al., 2014), and educational reforms like the No Child Left Behind (Linn, 2003). By experimentally separating the two, we distinguish and compare their effectiveness as stand-alone components, and contribute to the literature on the effectiveness of performance information (Andrabi et al., 2017; Barrera-Osorio et al., 2020; Bobba and Frisancho, 2022; Friedlander, 2020) by demonstrating that its impact may operate by signaling teacher attention and care.

Third, we complement the literature on peer effects (e.g., Bifulco et al. (2011); Burke and Sass (2013); Bursztyn et al. (2019); Bursztyn and Jensen (2015); Calvó-Armengol et al. (2009); Jackson et al. (2023); Lavy, Paserman and Schlosser (2012); Lavy, Silva and Weinhardt (2012); Sacerdote (2001); Wu et al. (2023)) by highlighting the role that peers play in how students respond to teacher-set expectations. While prior work has documented positive spillovers from high-ability to low-ability students (e.g., Booij et al. (2017); Carrell et al. (2009)), we show that such benefits are not universal. Exploiting random variation in peer matching, we find that encouragement only improves performance when peers are friends or similar in achievement and expectations. In contrast, mismatched peers can offset the otherwise positive impact of expectations. These results highlight the importance of homophily in shaping peer interactions and highlight that interpersonal comparisons can demotivate students.

Importantly, our findings highlight a particularly low-cost, non-invasive, and scalable intervention that meets the generalizability criteria proposed in List (2022). Delivering the intervention leverages the existing school infrastructure and costs less than ten cents per 0.1σ gain in test scores. This makes it one of the most cost-effective learning interventions (for example, see a review in Glewwe and Muralidharan (2016) and Beteille and Evans (2019)).[8] Unlike student-led goal-setting interventions (Damgaard and Nielsen, 2018; Dobronyi et al., 2019; Morisano et al., 2010; Oreopoulos and Petronijevic, 2019; Schippers et al., 2015), which may risk unrealistic benchmarks, or growth-mindset programs that require resource and time-intensive external facilitation (Alan et al., 2019; Duckworth and Quinn, 2009; Ganimian, 2020; Islam et al., 2021; Yeager and Dweck, 2012, 2020), our study evaluates a teacher-driven practice that can be embedded sustainably within classrooms.

We proceed as follows. In Section 2, we describe the empirical setting. We present the experiment design in Section 3 and describe the empirical strategy in Section 4. Section 5 presents the results, and Section 6 provides a detailed discussion of the mechanisms and cost-effectiveness. Section 7 concludes.

2 Data

The education system in Pakistan includes public, low-cost private, and private schools. The incidence of private schools has grown rapidly over the years, with 42% of children in the country enrolled in private schools (Andrabi et al., 2007; Qureshi and Razzaque, 2021).

[8] We document the details of our cost-benefit analysis in Section 6.

We partnered with a private school chain that operates approximately 300 schools across Pakistan, catering to middle- and upper-middle-income families. The schools have pre-primary (KG), primary (grades 1-5), lower-secondary (grades 6-8), and secondary (grades 9-11) grades.

We conducted our study in a sample of 15 schools with grades 3 to 8 across 288 classrooms.[9] Our sample constitutes 1,537 students, taught by 118 math teachers. There is considerable variation in student backgrounds within our sample. Approximately 44% of the schools cater to upper-middle income groups, while 38% to middle-income groups. About 12, 75, and 12% of the schools have low, medium, and high levels of parental literacy, respectively. Most of the schools are concentrated in Punjab, with a few schools in Sindh and Khyber Pakhtunkwa (Appendix Figure A.1.1).

2.1 Data Sources

2.1.1 Data on Academic Achievement

Each academic year has two terms, August to December and January to June. High-stakes standardized tests in Math and English are administered in every grade once every term. We collected administrative data from our partner schools, which included test scores for Math and English at multiple points in time: (1) historical scores from 2019 and 2020, (2) June 2021 (at the end of the first term following our intervention, referred to as the "midline"), and (3) December 2021 (at the end of the second term during our study, referred to as the "endline"). In addition to this, we also collected longer-term test scores at two points in time (1) June 2022 and (3) December 2022, i.e., after 12 and 18 months after the start of our intervention, respectively. These standardized tests are designed by our partner schools' curriculum advisors at the head office, are the same across all schools, reflect the curriculum being taught in different grades, and are high stakes. The tests are standardized at the grade level. Math and English scores, along with scores on other subjects, determine progression to the next grade.[10]

2.1.2 Surveys

We conducted a baseline survey with students before the intervention to measure demographic characteristics, classroom engagement, stress, intrinsic, and extrinsic motivation.[11] Then, we conducted a follow-up student survey six months after the end of the intervention (June 2022) to gather information on how students interpreted various components of

[9] We worked with primary and secondary grades until grade 8 only because after this, students opt into different education systems such as the local matriculation board or the GCSE Ordinary Levels.

[10] We do not have access to item-level test data and only have student-level aggregate percentage scores for Math and English for the two time periods.

[11] Additionally, we conducted two rounds of online surveys and independent tests with students after the intervention and two rounds of surveys with teachers before and after the intervention. Due to low response rates, we lack statistical power to detect treatment effects using measures from these instruments, so we do not present the results in this paper. The results are presented in the online appendix.

the information provided to them. We also conducted focus groups with a subsample of students to further understand how students interpreted the images. Additionally, we surveyed school heads to measure school-specific attributes such as parental income, literacy, how often they provide information about scores, and how this information is provided.

2.1.3 Data on Teacher Expectations

We elicited expectations from teachers about each of their students' math performance. To collect these in a standardized way, we asked teachers to share **realistic** expectations for each student after reminding them about the student's latest performance and requested them to fill in the following statements:

1. "I expect the student to work hard and improve to **achieve at least** X (out of 100%) in upcoming exams and tests."
2. "I expect the student to work hard and improve, and I think that **even** Y (out of 100%) **is achievable** in upcoming exams and tests."

In addition to collecting these expectations, we separately asked teachers to choose three general recommendations that they thought were most important to help all students improve their performance from a pre-specified list (compiled in consultation with teachers outside our study sample). The recommendation choice list included 'being more engaged in the classroom', 'asking questions', 'practicing from the textbook', 'practicing online', 'completing homework', 'attending virtual classrooms', and 'working with other students, or their parents'. These non-personalized recommendations were included for all students in the intervention infographic.

2.2 Descriptive Statistics

2.2.1 Student Characteristics

We present descriptive statistics for students in our sample in Table 1. Our sample includes 1,537 students from grades 3 to 8, between 6 to 15 years of age. 41% of the students are girls, and 84% of the students speak Urdu, while 64% also speak English at home. We find that 95% of the students report they want to get better at math. In addition, the majority of students value education highly and aspire to pursue higher education, suggesting that they are motivated to work hard. At the same time, 32% report that they feel that they are not as good at math, and over 52% report that they feel stressed about their current performance. Moreover, 75% of students report that they believe their teachers expect them to achieve over 90%. We suspect that unrealistic beliefs about what the teacher expects from them could be driving student stress. As we surveyed students at the time of remote learning during the pandemic, we find that engagement with teachers was limited. For example, 44% of the students report clarifying math problems with the teacher only once a week or never.

Finally, the majority of students report feeling academically motivated by their peers (74%) and report that peers do not trouble them for working hard (83%). To corroborate this further, we measure student networks by asking students to list their friends in the classroom and find that having more friends in the classroom is positively correlated with having higher extrinsic motivation. This positive classroom environment distinguishes our setting from other contexts that do not have conducive classroom norms, such as those in Bursztyn et al. (2017) and Bursztyn et al. (2019). Aligned with this, 61% of teachers in our teacher survey (see below) disagree with the notion that working hard is not considered 'cool' among students.

2.2.2 Teacher Characteristics

There are 118 teachers in our sample. 59% of them have a Master's degree and are predominantly ethnically Punjabi (Table A.1.2). About a third of teachers report concerns about classroom disruption, attendance, or students not completing their homework (Table A.1.3). About 69% of teachers report that they think their encouragement matters the most for student performance, compared to encouragement from parents and peers. When asked to think about who would improve the most after receiving high expectations, only 23% of the teachers report students at the bottom end of the distribution as their first choice, compared to students at the middle or top end of the distribution. These baseline patterns motivate our intervention as teachers are aware of the importance of their expectations but do not prioritize students at the bottom end of the score distribution while thinking of conveying these expectations. These are students who can potentially have the highest marginal benefits.

At the same time, teachers also acknowledge the motivational role of peers. 85% agree or strongly agree that students care about what their friends think about them. 53% report that expectations should be conveyed to those who will be most successful in encouraging others, compared, for example, to those whose academic performance would improve the most (15%) or those whose non-cognitive outcomes would improve the most (27%). This adds credence to the peer component of our study design.

2.3 Teacher Expectations

We now briefly describe teachers' expectations to give a sense of their levels and variation with baseline test scores. We find that expectations are strongly but not perfectly related to baseline performance: baseline math scores alone explain about 44% (49%) of the variation in High (Very High) expectations. As Figure A.4.1 shows, on average, teachers' expectations exceeded students' baseline scores by 5 points (High) and 7 points (Very High) on the 100-point scale across treatment and control groups. These descriptives highlight that expectations were strongly anchored to baseline achievement but systematically optimistic on average across all arms. In the results section, we show how these gaps evolve across the treatment arms.

By reminding teachers about every student's recent math score before writing their expec-

Table 1: Summary Statistics of Students

	Count	Mean	SD	Min	Max
Student Characteristics					
Age	1,369	10.59	1.74	6.00	15.00
Adults per Room	1,315	0.56	0.34	0.07	3.00
Female	1,537	0.41	0.49	0.00	1.00
Speaks English at home	1,468	0.64	0.48	0.00	1.00
Speaks Urdu at home	1,537	0.84	0.37	0.00	1.00
Value of Education (1-5)	1,101	4.60	0.76	1.00	5.00
Aspires to obtain Master's degree or higher	814	0.85	0.36	0.00	1.00
Classroom Engagement					
Weekly Hours doing Math Homework	1,370	2.96	4.22	0.00	30.00
Weekly Hours Studying Math	1,371	3.80	4.79	0.00	41.00
How often do you discuss math with your teacher?	1,385	1.71	0.98	0.00	3.00
How often do you discuss math with your parent?	1,385	1.79	1.10	0.00	3.00
How often do you discuss math with your peers?	1,385	0.98	0.97	0.00	3.00
Peer Characteristics					
Number of Friends in the Classroom	1,537	4.07	2.64	0.00	10.00
Stress					
Stressed about Own Performance	1,333	0.52	0.50	0.00	1.00
Stressed about Teacher's Expectations	817	0.46	0.50	0.00	1.00
Stressed about Peer's Expectations	817	0.31	0.46	0.00	1.00
Stressed about Parent's Expectations	817	0.62	0.49	0.00	1.00
Stress Index	817	0.48	0.38	0.00	1.00
Intrinsic Motivation					
Feels not good at math	1,333	0.32	0.47	0.00	1.00
Feels they work hard at math	1,333	0.87	0.33	0.00	1.00
Wants to get better at math	1,333	0.95	0.22	0.00	1.00
Intrinsic Motivation Index	1,333	0.85	0.20	0.25	1.00
Extrinsic Motivation					
Motivated by Peers	1,338	0.74	0.44	0.00	1.00
Troubled by Peers for Bad Performance	1,338	0.12	0.32	0.00	1.00
Troubled by Peers for Working Hard	1,338	0.17	0.38	0.00	1.00
Extrinsic Motivation Index	1,338	0.81	0.24	0.00	1.00

Note: The statistics are from the baseline student survey. Variables related to stress with regard to teacher's, parent's or peer's expectations, and aspirations for higher studies were only collected for the older students (in grade 5 and above) following a pilot of the survey. Students in grades 3 and 4 were asked to list up to 5 friends, while those in older grades were asked to list 10 friends. Variables measuring the number of hours doing homework or studying math exclude outliers above the 99^{th} percentile.

tations, we minimized the risk of gender or wealth-related stereotypes driving their expectations. Figure A.4.2 panel (a) confirms this as teacher expectations were not systematically related to student gender, age, or wealth, and including these covariates raises the R^2 by only about one percentage point. Expectations were very similar across boys and girls, younger and older students, and across wealth groups, suggesting little evidence of demographic bias.

Figure A.4.2 panel (b) further illustrates that the gap between teachers' expectations given in the intervention and actual baseline performance was largest among students in the bottom quartile of the score distribution. This suggests that teachers believed in substantial improvement potential for lower-performing students, even though most reported in the baseline survey that they would not prioritize delivering high expectations to such students.

3 Experiment Design

3.1 Randomization Design

Figure 1 shows the randomization design. We use a clustered randomized design at the classroom level and randomly allocate one-third of classrooms to the Expectations Arm (where student-specific high teacher expectations and encouragement are conveyed individually to a student), one-third to the Peer Arm (where in addition to conveying student-specific high teacher expectations and encouragement individually to a student, they were additionally randomly matched with another classmate and asked to encourage each other),[12] and one-third to a Comparison Group. Half of the Comparison Group classrooms were randomized to receive a reminder about their last test score (Information Arm) and half were randomly selected to receive no messages (Control Group). Within the Expectations Arm and Peer Arm, half the students were randomly chosen to receive the "High" teacher expectation, and half received the "Very High" teacher expectation with the corresponding statements outlined earlier (Section 2.1.3). In the Supplementary Appendix Section B, we present a simple effort-choice model to interpret how Expectations, Information, and Peer comparisons affect student motivation and achievement, providing a rationale for the experiment design.

The randomization was stratified along grade,[13] gender composition of the school (co-educational or single gender) and whether the average class math test score (%) in the preceding year (2020) was above or below the median.

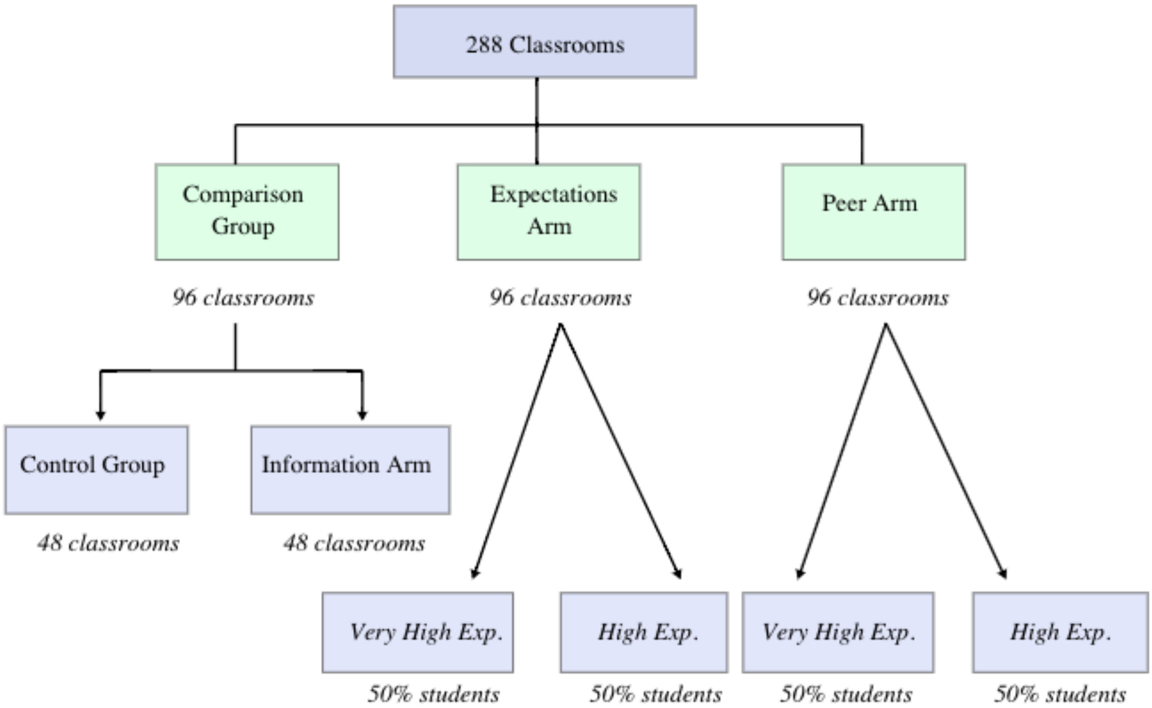
3.2 Timeline

The timeline of the study is as follows. Informed parental consent and student assent were obtained between March and May 2021. Teacher expectations were elicited and delivered by

[12] In the Peer Arm, we randomly matched students with another student of the same gender, taking into account the cultural norms in the Pakistani context.

[13] We use a binary variable to indicate Grade 3 students (very young and unable to complete the survey without enumerator instructions and outside of class) separately from grades 4-8 (older grades).

Figure 1: Randomization Design



mid-June. We collected administrative test score data on student performance in June/early July. We sent two reminders to students about their teacher expectations–one at the start of the summer holidays and another at the beginning of the new academic year in August. A final round of teacher expectations with updated design graphics and scores was sent in November 2021 before the school conducted its end-of-term exams in December 2021. A final round of follow-up surveys with school administrators and students was conducted between March to May 2022. We also collected long-term student test score data in June and December 2022.[14]

3.3 Format and Delivery of Teacher Expectations

We delivered student expectations via emails as the schools had switched to using virtual learning during the pandemic. The virtual learning infrastructure (Google Classrooms), existed prior to the pandemic and was regularly used by teachers to communicate with students. An enumerator was added to each Google Classroom as a co-teacher to email students privately. While the emails were sent on behalf of the teachers, the teachers were not able to see who the email was sent to or the content of the emails. We also confirm in the endline student survey that students did not report teachers spending any extra time talking with them after class, with no statistically significant differences between treatment and control groups.

[14]While we don't report the results of the student and teacher surveys in this paper because of a low response rate, we conducted two rounds of surveys (before and after the intervention).

Figure 2 shows the designed graphics sent out to each group. The graphic used to deliver teacher expectations positions each student as a superhero who can work towards achieving the teacher's expectations. The staircase includes generic tips for all students on how to achieve the goal (as described earlier in Section 2.1.3). Students in the Expectations Arm received their most recent math test score and teacher's expectation ("High" or "Very High") according to their treatment status.[15] Appendix Figure A.2.1 illustrates the difference in the "High" and "Very High" statements on the images.

In the Peer Arm, students first received a private email with their test scores and their (individual) teacher's expectations (just like the Expectations Arm). In addition, they also received a joint email with their matched classmate with the additional line *'We hope you both will encourage each other'*. The joint email (and infographic) was to encourage students to support each other. Importantly, the joint email did not contain any information about either student's test scores or teacher expectations. Students in the Information Arm received a graphic with a simple image of a boy or a girl with their most recent Math score.

Our follow-up survey gathered information on how students interpreted the information on these graphics. We discuss these interpretations in the mechanisms section of the paper. [16] For the second round of intervention, we re-designed the graphics as shown in Appendix Figure A.2.2. The graphics were updated based on the most recent scores of the students, but the expectations communicated remained the same. These graphics were emailed before the end-of-term exams in December 2021.

4 Empirical Strategy

4.1 Specification

Our main specification regresses pre-specified outcomes on $Information_c$, $Expectation_c$, and $Peer_c$ which equal 1 if the student is in a classroom c in the Information, Expectations, or the Peer Treatment Arms respectively.[17] We use the pooled sample combining data from the

[15] For those students in the treatment group whose teachers reported lower expectations than their previous score (22%), the infographic only included the teacher's expectation and not their previous score to avoid any demotivating effect (as required by our IRB). This does not confound our identification of ITT effects of teacher expectations.

[16] In addition to the follow-up survey, we also conducted focus groups with selected students to understand how students interpreted the images.

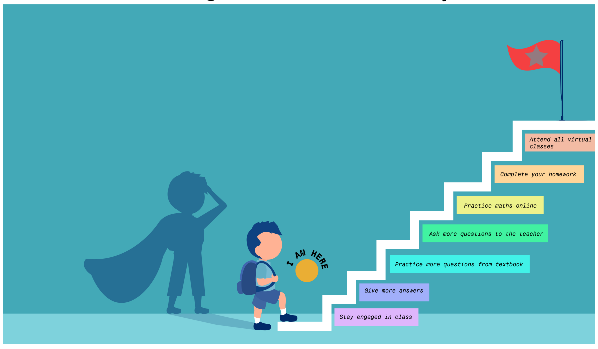
[17] These binary variables capture intent to treat rather than actual treatment status. However, 88% of those students who completed our midline survey reported reading the emails and the proportion is balanced across the different treatment arms so we suspect that the treatment on treated results would be slightly higher but not very different than our ITT estimates. Since we do not have this indicator for all students, we are unable to run the treatment on treated regressions.

Figure 2: Treatment Delivery Design Variations - Round 1

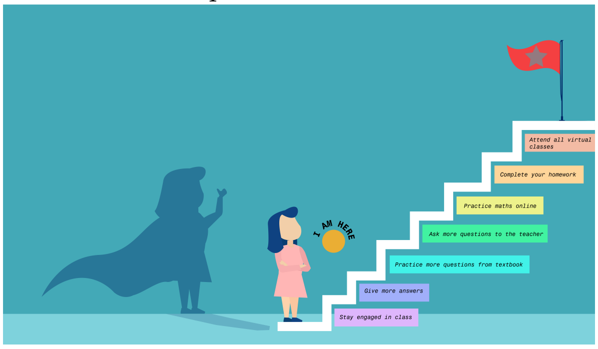
(a) Information Arm - Boy

(b) Information Arm - Girl

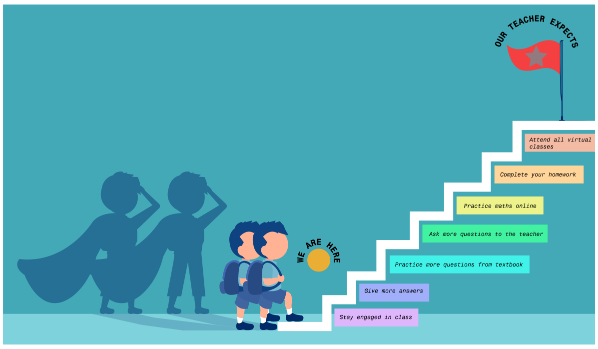
(c) Expectations Arm - Boy



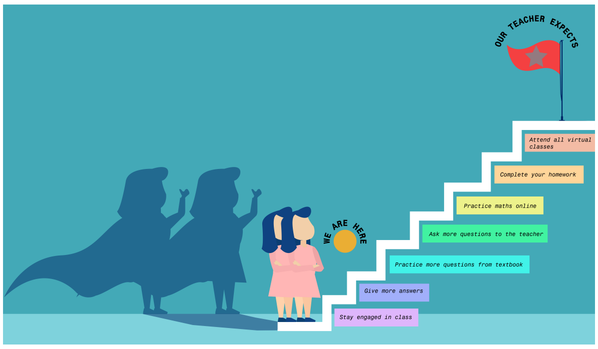
(d) Expectations Arm - Girl



(e) Peer Arm - Boy



(f) Peer Arm - Girl



midline and endline waves for the main results to maximize power.[18]

$$Y_{ict} = \beta_0 + \beta_1 Information_c + \beta_2 Expectation_c + \beta_3 Peer_c + \phi_s + \lambda_t + \epsilon_{ict}$$

Standard errors are clustered at the class level (unit of treatment). We include fixed effects ϕ_s for each stratum s and control for a round fixed effect λ_t for midline and endline. We present results on standardized test scores and raw test scores, controlling for baseline student per-

[18]However, results for the midline and endline waves separately are also presented in the supplementary appendix. The differences between the treatment effect on scores across midline and endline waves are not statistically significant. We also pool the long-term results from the two waves to maximize power and those results are presented in the Appendix.

formance in a value-added specification in the latter case.

4.2 Balance

4.2.1 Student and Class-Level Characteristics

We adopt two approaches to check for balance. First, we show that student characteristics are balanced across control and treatment groups for the pooled sample, midline sample, and endline sample. These include student-level characteristics such as baseline math scores, gender, wealth, classroom effort in terms of hours spent studying and preparing for exams, number of friends, classroom engagement, and intrinsic and extrinsic motivation. These results are shown in Tables A.3.1, A.1, and A.2. Next, we show balance across the treatment arms at the class level using average historical scores in Math and English, class-level variables such as class size, grade, teaching experience of the teacher, teacher-reported student engagement (motivation and interaction), disruption and warnings, absenteeism, and parental engagement. These results are shown in Table A.3.2. We find that control and treatment classrooms are balanced across most characteristics. However, we will also account for any balance-related concerns in our robustness specifications where we will employ Post-Double Selection Lasso as proposed in Belloni et al. (2014).

4.2.2 Teacher Expectations

In addition to checking for balance along student and class characteristics, we also check for balance in the expectations elicited from teachers across different treatment arms. We confirm that there are no systematic differences in teacher expectations across treatment and control arms (Figure A.4.1). This adds credibility to our research design.

It is also worth noting that since baseline achievement and teacher expectations are similar across treatment and control arms (Table A.3.1, Figure A.4.1), it is unlikely that idiosyncratic shocks drive our results. In particular, one could imagine a scenario where a high-performing student might have had an unusually bad test day, which would create a large gap between their baseline score and the teacher's expectation. Such a student would rebound in subsequent tests, creating the appearance of an effect of expectations even in the absence of one. However, given randomization and balance in baseline test scores and expectations, these shocks would be expected to be evenly distributed across arms and cannot explain the treatment effects we observe

5 Results

5.1 Main Results

5.1.1 Effect on Math Performance

Table 2 presents the treatment effects from our main specification on Math scores on high-stakes tests conducted by our partner schools. Column (1) reports standardized test scores and column (2) reports raw percentage scores.[19] We find that students in the Expectations Arm score 0.21σ higher than students in the Control Group (significant at 1%). This is equivalent to a 3.3 percentage point increase in percentage scores. At the same time, we find that students who received information about their previous test scores also score 0.18σ (significant at 5%) higher than students in the Control Group, equivalent to a 2.7 percentage point increase in percentage scores. We find that the effect of the Information Arm is not statistically distinguishable from the effect of the Expectations Arm.[20]

This suggests that receiving a message on behalf of the teacher that contains just a reminder about the student's past performance can also increase student performance and be just as effective as teacher expectations. We discuss the mechanisms underlying this effect in the next section and find that this is driven by schools with low parental literacy and students interpreting this reminder as an encouragement message from the teacher. [21]

The results from the pooled specification are also consistent with the treatment effects estimated separately for the midline and endline waves (Tables B.1 and B.2). Further, as shown in Table B.3, while the treatment effects for all arms are smaller in magnitude in the endline, the differences over time are not statistically significant. As a result, we infer that the effect of the intervention is sustained over time.

Finally, we find no average effects of the Peer Arm on test scores. Further, the difference between the effects of the Expectations and Peer Arm is statistically significant. This is particularly surprising since the Peer Arm adds the peer matching component to the Expectations Arm. This finding suggests that while students may benefit from receiving teacher expectations, they may, on average, be negatively affected by being matched with a random classmate leading to an overall null effect. In the next section, we leverage the fact that peers were matched randomly to provide evidence of heterogeneous treatment effects. This will allow us to understand why the Peer Arm did not succeed in improving test scores on average.[22]

[19] The regression specification in Column 2 additionally controls for the student's baseline score which explains the minor difference in the number of observations between the two columns.

[20] It is important to note that these are intent-to-treat effects. While 88% of the midline survey sample reported reading the emails, the actual treatment effects are likely to be higher.

[21] We also present a simple effort-choice model to illustrate how expectations, information, and peer comparisons can affect student motivation in the Supplemental Appendix (Section B).

[22] Additionally, we also estimate the effect of the treatments on class-level variance in math test scores and find that the treatments reduce the variance of test scores but the effects are not significant.

Table 2: Treatment Effects on Test Scores

	(1) Standardized	(2) Raw
Panel A. Targeted Subject: Math Scores		
Expectations	0.209***	3.261**
	(0.074)	(1.377)
Peer	0.068	1.086
	(0.078)	(1.361)
Information	0.179**	2.747*
	(0.084)	(1.435)
Observations	2773	2640
Comparisons (p-values)		
Exp vs Peer	0.040	0.028
Exp vs Info	0.696	0.640
Info vs Peer	0.163	0.128
Panel B. Spillover Subject: English Scores		
Expectations	-0.191	0.261
	(0.148)	(1.260)
Peer	-0.390**	-1.066
	(0.180)	(1.340)
Information	-0.037	0.809
	(0.162)	(1.360)
Observations	2413	2413
Comparisons (p-values)		
Exp vs Peer	0.235	0.210
Exp vs Info	0.312	0.608
Info vs Peer	0.042	0.096

Note: Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The results are from pooled regressions of midline and endline scores. The scores in column (1) are standardized using the mean and standard deviation of math scores of students in the control group at baseline. Column (2) reports the raw scores of students (converted to percentages) in a value-added specification i.e. controlling for a student's baseline score. Regressions include strata and round fixed effects and standard errors are clustered at the level of randomization.

Robustness to controls: We also use the Post Double Selection Lasso strategy (Belloni et al., 2014) to show that the treatment effects on test scores do not change even after accounting for any baseline characteristics that might be correlated with treatment indicators (Table E.1).

Alternative Specification: Since the Peer Arm combines peer matching with the Expectations

Arm, we also present an alternative specification to distinguish the effects of expectations from those of matching (Table A.9.1). In this specification, we define two binary variables: 'Information' (equal to 1 for the Information Arm), 'Expectations' (equal to 1 for both the Expectations and Peer Arm).[23] We find that expectations had a significantly positive effect on math scores equal to 0.21σ (significant at 1%) and peer matching had a significantly negative effect of 0.14σ (significant at 5%). As before, the effect of information and expectations is comparable.

5.1.2 Effect on English Performance

In addition to the above results on math test scores, Table 2 shows that the Expectation and Information Arms do not have any spillover effects on English test scores in the pooled sample, while the Peer Arm has a negative and significant effect. When we separate the results at the midline and endline, we find that all three treatment arms have an insignificant effect on English test scores in the midline, but a negative effect of -0.45σ and -0.62σ in the endline which is significant at 5% and 1% for the Expectations and Peer Arm respectively (Tables B.4 and B.5). This could indicate that over time, a subject-specific expectation might lead students to substitute effort towards that subject and away from other subjects.

5.2 Heterogeneous Effects

5.2.1 Magnitude of Expectations

First, we exploit the exogenous variation in the type of expectations delivered (i.e., 'High' or 'Very High'). The results are presented in Table 3. We find that the Expectations Arm significantly raises test scores when expectations are high enough, i.e., students who received a 'Very High' expectation from teachers, scored 0.27σ higher in math compared to the Control Group (significant at 1%). Additionally, the effect on those students who were given a 'High' expectation is 0.13 standard deviations but not statistically significant. This provides evidence for the hypothesis that providing students with ambitious goals set by teachers can have high returns without leading to frustration. Moreover, the difference between the 'Very High' and 'High' expectation effect is statistically significant at 10%. This result is similar even when we consider the midline and endline waves separately.

Panel B in Table 3 shows the results of the specification where we regress the scores on the treatment arms interacted with the gap between the student's baseline score and the expectation delivered to them. We find that the effect of both the Expectation and Peer Arm is higher among students for whom this gap is larger. We find that a 10 percentage point increase in the gap between expectations and baseline score leads to a 2 percentage point increase in the

[23]The Expectations Arm infographic was interpreted by students as highlighting the potential for improvement (i.e., students noticed the gap between the expectation and their current score), as is confirmed by our follow-up survey. Hence, we do not treat the effect of the Expectations Arm as additively separable from the effect of information about the past score.

impact of the Expectations Arm. This implies that receiving a higher expectation relative to one's performance increased test scores.

Note that the larger effect of a gap between baseline performance and expectations does not arise mechanically from low-performing students simply having more room to improve. Since students were randomized to receive the 'Very High' statement, a significant effect for them provides evidence against this interpretation. Moreover, the second column in Panel B of Table 3 includes a control for students' baseline test scores. Even after conditioning on baseline performance, the interaction between the gap and the expectations treatment remains significant at the 5% level, implying that among students with the same baseline score, those who received a higher expectation performed better.

5.2.2 Characteristics of the Matched Peer

Next, we exploit the random variation in matching in the Peer Arm to examine the heterogeneity of treatment effects along the characteristics of the randomly matched peers. To systematically explore this, we use baseline classroom network data to compare individuals randomly paired with a friend to those who were not.[24] As shown in Table A.7.1, the effect on test scores is significantly larger for those paired with a friend compared to those who were not. To understand this further, we construct a measure of homophily among the matched peers as a measure of their similarity in terms of baseline characteristics such as baseline scores, teacher expectations, classroom motivation, parental wealth, and number of friends in the classroom. We construct the index by first generating the squared differences in terms of these characteristics, standardizing these differences, and then constructing an inverse variance weighted average (Anderson, 2008). The homophily index is the negative of this average.

As shown in Table 4, the effect of the Peer Arm is higher for those for whom the homophily index is higher. We find that the effect of the peer treatment arm is negative for students for whom the homophily index is low and positive for those for whom it is high. We break this down further by looking at how the treatment effect within the Peer Arm differs by the extent of similarity in terms of teacher expectations and baseline scores within matched pairs in Panels B and C of Table 4 respectively. We find that both individuals matched with peers who received similar teacher expectations and those matched with peers who received lower expectations scored significantly higher—by 0.39σ and 0.30σ, respectively—compared to those matched with peers who received higher expectations. This is reinforced by our follow-up survey (discussed in more detail later), in which students report that they would feel disappointed and less motivated if their matched peer received a higher expectation than them. Reinforcing these patterns of heterogeneity of treatment effects of the Peer Arm, we find that the effect of being matched with a peer with the same baseline score is also 0.33σ higher than being matched with someone with a higher baseline score.

[24]We define two individuals as friends if either listed the other's name during the baseline network elicitation.

Table 3: Heterogeneity with Statement and Magnitude of Expectation Delivered

	(1) Standardized	(2) Raw
Panel A. By the Type of Expectation Statement Delivered		
	Standardized	Raw Scores
Expectations (Very High)	0.268***	3.525***
	(0.081)	(1.357)
Expectations (High)	0.135	2.435
	(0.083)	(1.496)
Peer (Very High Expectation)	0.033	0.398
	(0.083)	(1.320)
Peer (High Expectation)	0.108	1.117
	(0.091)	(1.596)
Information	0.176**	2.483*
	(0.082)	(1.351)
Observations	2773	2640
Comparisons (p-values)		
Exp (Very High) vs Info	0.266	0.381
Exp (Very High) vs Exp (High)	0.086	0.348
Exp (High) vs Info	0.635	0.970
Peer (Very High) vs Info	0.103	0.074
Peer (Very High) vs Peer (High)	0.373	0.599
Peer (High) vs Info	0.469	0.342
Panel B. By the Gap between Expectation and Baseline Score		
Expectations	0.105	1.845
	(0.080)	(1.306)
Peer	0.096	1.719
	(0.085)	(1.391)
Information	0.144	2.544*
	(0.091)	(1.474)
Expectations x Gap between Expectations and Baseline Score	0.016**	0.281***
	(0.006)	(0.098)
Peer x Gap between Expectations and Baseline Score	0.005	0.109
	(0.006)	(0.104)
Observations	2180	2180

Note: Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The results are from pooled regressions of midline and endline scores. The gap in panel B is the difference between the expectation delivered to the student and their performance. The scores in column (1) are standardized using the mean and standard deviation of math scores of students in the control group at baseline. Column (2) reports the raw scores of students (converted to percentages) in a value-added specification, i.e., controlling for a student's baseline score. Regressions include strata and round fixed effects and standard errors are clustered at the level of randomization.

Table 4: Heterogeneity in Treatment Effects by Matched Peer Characteristics

	(1) Standardized	(2) Raw
Panel A. By Homophily Index (Whole Sample)		
Expectations	0.205***	3.250**
	(0.074)	(1.370)
Information	0.179**	2.779*
	(0.084)	(1.433)
Peer	-0.862***	-8.748**
	(0.316)	(4.030)
Peer x Homophily	1.202***	12.546***
	(0.370)	(4.563)
Constant	-0.279***	41.751***
	(0.095)	(3.737)
Observations	2467	2355
Panel B. By Matched Peer's Expectation (Within Peer-Arm)		
Own expectation	0.031***	0.354***
	(0.006)	(0.092)
Peer's expectation is same	0.377***	4.700*
	(0.134)	(2.383)
Peer's expectation is lower	0.279**	3.870
	(0.128)	(2.577)
Constant	-3.004***	25.778***
	(0.560)	(9.718)
Observations	591	589
Panel C. By Matched Peer's Baseline Score (Within Peer-Arm)		
Own score	0.024***	0.422***
	(0.006)	(0.100)
Peer's score is same	0.311*	5.286*
	(0.158)	(2.684)
Peer's score is lower	0.036	0.780
	(0.147)	(2.507)
Constant	-2.168***	43.246***
	(0.460)	(7.528)
Observations	589	589

Note: Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The results in Panel A are from pooled regressions of midline and endline scores. The Homophily Index is a measure of the similarity between matched peers in terms of baseline characteristics such as scores, teacher expectations, classroom motivation, parental wealth, and number of friends in the classroom. Panel B and Panel C show within Peer-Arm regression results. The scores in column (1) are standardized using the mean and standard deviation of math scores of students in the control group at baseline. Column (2) reports the raw scores of students (converted to percentages) in a value-added specification, i.e., controlling for a student's baseline score. Regressions include strata and round fixed effects and standard errors are clustered at the level of randomization.

When compared with the Control Group, Appendix Table C.1 shows that students who were matched with a peer with the same baseline score achieve a 0.21σ higher test score (significant at 10%) than the control group. This effect is not statistically distinguishable from that of the Expectations Arm. However, being matched with a peer with a higher baseline score does not improve student performance. Even though this effect is not statistically distinguishable from that of being matched with someone with the same score, we find that it is significantly lower than the effect of the Expectations Arm. This reinforces the finding that peers who are similar in terms of baseline scores perform significantly better than the Control Group and have a treatment effect as large as those who were in the Expectations Arm.

5.2.3 Score Distribution

First, we run quantile regressions and show that the treatment effects of the Expectations and Information Arms discussed above are driven by positive effects on students at the bottom and middle of the distribution of baseline math test scores. Figure A.5.2 plots the treatment effects on different quantiles of the score distribution. We find that the treatment effects of the Expectations Arm and the Peer Arm are higher for lower quantiles of performance and decline as the score increases. The Peer Arm has no effect on average and displays little heterogeneity across the quantiles of the baseline student test score distribution.

The positive effect on this subgroup is further validated in Table A.5.1 where we employ the strategy recommended in Abadie et al. (2018). This strategy predicts math performance for the control group using a set of covariates selected by LASSO from a list including variables measuring demographic characteristics, classroom engagement, academic effort, and motivation. We de-bias the prediction process and deal with "endogenous stratification" by computing the leave-one-out estimator using data from the control group. We then use this model to predict performance for all students and classify them into four subgroups for which we separately compute heterogeneous treatment effects. These results are shown in Table A.5.1 where we find evidence that the treatment effects are strongest for students predicted to perform poorly, i.e., in the worst-off group. In particular, the effect of the Expectations Arm on test scores of the students predicted to perform the worst is 0.5σ and significant at 1%. In contrast, the effect on those predicted to perform the worst is not significant for either the Peer or Information arm. The treatment effects on those predicted to perform the best are close to zero and statistically insignificant. Therefore, the Expectations benefit the weakest students the most.

5.2.4 Additional Evidence of Heterogeneous Treatment Effects

We also apply the method outlined in Chernozhukov et al. (2018) to examine evidence of heterogeneity by baseline characteristics for each of the three arms. We detect evidence of heterogeneity for both the Individual and Peer Arms. We then categorize individuals into four groups based on their predicted performance under treatment, ranging from lowest to highest. Analyzing the baseline characteristics of these groups, we find significant differences in both baseline scores and matched peers' scores, underscoring the importance of our previous findings. The procedure and the results are discussed in detail in Appendix Section A.5.1.

5.3 Long-Term Results

We also measure student test scores in Math and English 12 and 18 months after the start of our intervention. We were able to get administrative data for a subsample of 880 and 768 students, respectively.[25] We do not find any significant average treatment effects in Math for the Expectation, Peer, or Information Arms, as shown in Table A.8.1. Notably, these test scores capture six months and a year without receiving any reminders of teacher expectations. This shows that reminders are critical for sustaining the impact of teacher expectations in the long run.[26] Additionally, we do not detect any effects on English test scores as shown in A.8.2.

Examining heterogeneity based on the magnitude of expectations, we do not find any differences between individuals randomly assigned to the "Very High" versus "High" expectations groups. However, as Table A.8.3 shows, the treatment effects of the Expectations Arm are significantly larger (p-value <0.01) for individuals who had a greater gap between their expectations and baseline performance. We find no corresponding effect for the Peer Arm.

6 Discussion

6.1 Mechanisms

We use school administrative data, surveys with head teachers, follow-up surveys with students, and findings from heterogeneity analysis to inform our understanding of the mechanisms behind the treatment effects in each treatment arm.[27]

6.1.1 Expectations Arm

The infographic delivered to students in the Expectations Arm contained information about their past performance, their teacher's expectations, and generalized tips that they can follow to achieve them. We find evidence suggesting that the treatment effect is driven by motivation from teacher expectations rather than information about past performance, tips provided to the students, or changes in teacher behavior.

[25]These data were shared by our partner schools, depending on their ability to locate students in their databases. Score availability is uncorrelated with treatment status or baseline performance.

[26]We considered if the reason we do not see effects is because students have already met the teachers' expectations. However, we find that 69% of students scored below the teacher's originally delivered expectations across these waves.

[27]While we administered student surveys at midline and endline to measure a broad set of outcomes, including effort, parental and peer engagement, classroom norms, motivation, and non-cognitive traits such as grit and growth mindset, more than half of the students did not complete these surveys. We do not discuss these findings further due to the high non-response rate, but these results are presented in the online appendix.

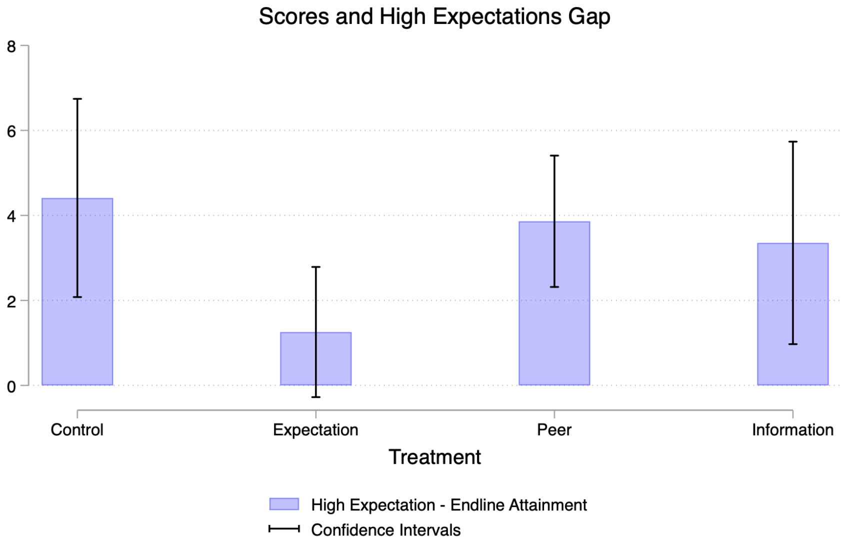
First, we find that the magnitude of the expectation influences the effectiveness of this treatment arm. In particular, those randomized to receive the 'Very High' expectation have a significantly higher treatment effect (at 10% significance) than those randomized to receive a 'High' expectation. Moreover, a 10 percentage point increase in the gap between expectation and baseline performance leads to a 2 percentage point increase in the impact of the Expectations Arm (Table 3). Consistent with this, we find that the gap between teacher expectations and the score achieved at the endline is the smallest for students in the Expectations Arm which provides evidence to support that students worked towards the expectation set for them by their teachers when this was communicated to them. In particular, Figure 3 shows that this gap was statistically indistinguishable from zero for students who received the 'High Expectations' statement. Similarly, the gap was smaller in magnitude (6 percentage points) for students who received the 'Very High Expectations' statement than students in the Peer Arm and Information Arm (9 percentage points), although we are not statistically powered to show that these differences are significant.

Consistent with this, we find that 70% of students interpreted the Expectations Arm image as a goal-setting mechanism or a form of encouragement from their teacher, rather than as a comment on how smart they are, or inferring that they are lagging or being monitored (Figure A.10.2a). 76% report that they would feel motivated or happy if they were sent the image, as opposed to feeling stressed or disappointed (Figure A.10.2b). At the same time, 92% reported feeling motivated by teacher expectations.

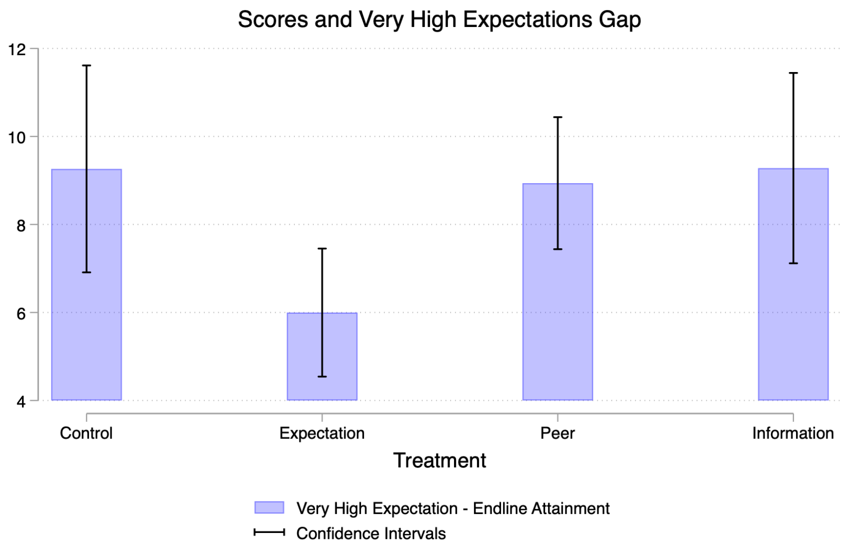
Considering alternative mechanisms, we first show that the tips on the infographic are unlikely to be new information to students, as all the teachers unanimously reported that they were already conveying tips to students about how they can improve their performance in our baseline survey. The tips on the infographic were also not student-specific and very generic (e.g., 'Being more engaged in the classroom', 'Completing homework', etc.). Our follow-up survey also reveals that students notice the potential for improvement and the expectations set by their teacher, rather than the information about their previous score (Figure A.10.1). In fact, only 14% of the students reported that they would notice their score the most in the image.

Additionally, we find that the observed treatment effects are not driven by changes in teacher behavior, as teachers were blind to student treatment status by design. Consistent with this, over half of the students in our endline survey reported that their teacher did not spend extra time discussing math with them after class, with no statistically significant differences between treatment and control groups. The effectiveness of the Expectations Arm, therefore, lies in its ability to instill a sense of improvement and motivation among students, rather than simply providing informational content or leading to changes in teacher behavior.

Figure 3: Gap between teacher expectations and student performance at endline



(a) Gap between "High" Expectations and Score Across Treatment Arms.



(b) Gap between "Very High" Expectations and Score Across Treatment Arms.

Note: Panel (a) plots the gap between 'High' expectations elicited from teachers and students' endline score across the treatment arms with 95% confidence intervals. Panel (b) plots the gap between 'Very High' expectations elicited from teachers and students' endline scores across the treatment arms with 95% confidence intervals.

6.1.2 Information Arm

Students in this treatment arm received an infographic displaying their score from the most recent test. We find evidence that students interpret the image as a sign that their teacher pays attention to them.

The majority of students interpreted the information as attention from the teacher. Nearly 35% of the students in our follow-up survey inferred that the teacher was intending to encourage them, 26% thought that the teacher expected them to continue achieving the same score, and 21% felt that the teacher was monitoring them (Figure A.10.2a). This underscores the potential of delivering performance-related information in a targeted and encouraging manner to positively influence student perceptions and ultimately, their academic performance. One caveat here is that it matters whether the students think they were the only ones receiving the image versus if the entire class was receiving it. In particular, while most students still inferred that the teacher was trying to encourage them when sending a reminder about their last score, students commonly expressed that they would feel indifferent or no reaction if the image was sent to all (Figure A.10.2b). This reinforces that the belief that it is a targeted, personalized message from the teacher plays a crucial role in its effectiveness.

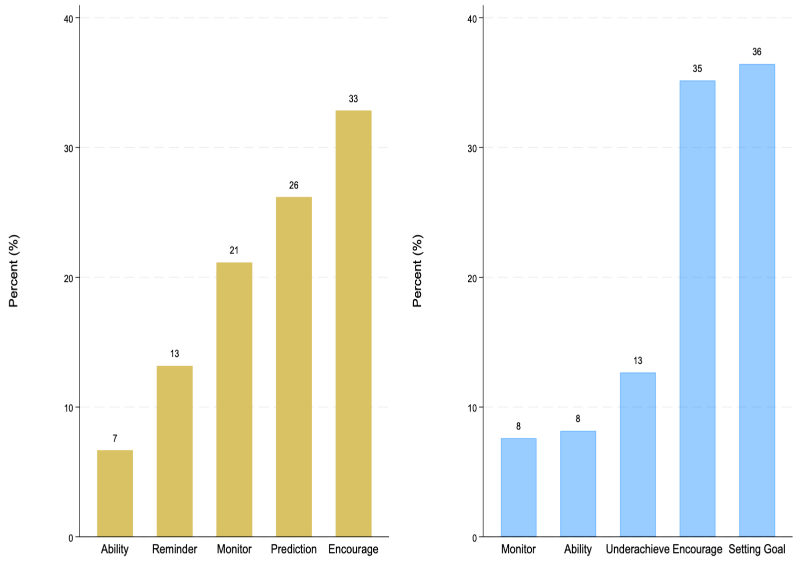
In addition to this, the ease of interpreting information could also be an additional factor driving the effects of this arm. 75% of the schools reported that parents have medium literacy, and 13% mentioned low literacy. Additionally, 88% of the schools reported limited technological proficiency among parents. Table A.6.2 shows significantly lower effects of information in schools with high parental literacy compared to schools with lower parental literacy. Importantly, we do not observe this heterogeneity in the Expectations Arm, suggesting that this channel alone cannot explain the effects of the Information Arm as both arms provide information about current performance. Supporting this, over 50% of students reported that even though they remember their scores, they still find the image helpful as a reminder. Therefore, while we cannot rule out that delivering information in a simple, visually accessible format in the Information Arm can enhance comprehension in low-parental literacy environments, this alone cannot explain the entirety of the efficacy of the Information Arm as we do not find similar heterogeneity in the Expectations Arm.

We also discuss and rule out alternative explanations for the efficacy of the Information Arm. First, we find that the image did not provide new information, as students typically receive report cards with their test scores at the end of each term.[28] Nearly 80% of the schools send report cards at the end of each term. In fact, for older grades (5-8), 20% of the schools send out report cards every month, and 7% do so for younger grades (3-4). Since the first round of our intervention was delivered close to the end of the term and the second round after the end of the term, the treatment effect is unlikely to be driven by pure information effects.

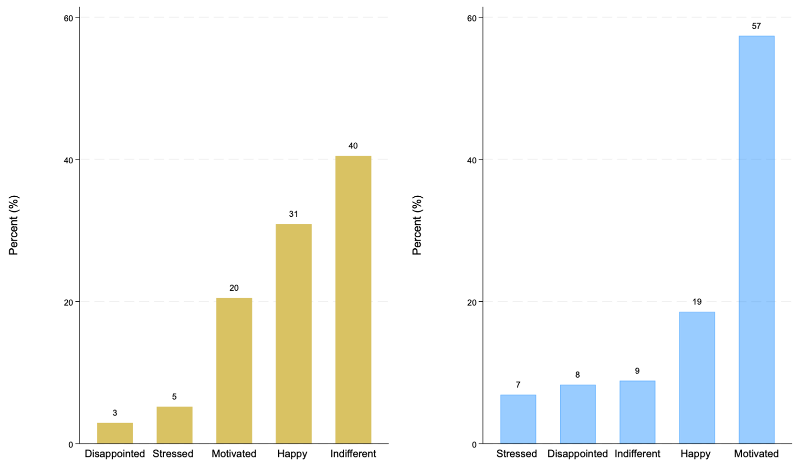
Next, while schools differ in terms of whether or not they provide hard copies of student report cards, we find no significant differences in treatment effects between 44% of schools that send printed report cards home (in addition to SMS and online links) versus those that do not (Table A.6.3). It is also unlikely that the treatment effects observed in this arm can be attributed to the format of the information delivered. This is because expressing scores

[28]Scores are indicated separately for each test and do not combine any other evaluations like participation or homework.

Figure 4: Inferences and Feelings about Images - Expectations and Information Arms



(a) Student Inferences from the Information and Expectations Arm Images



(b) Student Feelings about the Information and Expectations Arm Images

Note: Panel (a) presents students' thoughts after receiving the image in the Information Arm (left) and Expectations Arm (right). Respondents could choose from: 'My teacher is monitoring my progress' (labeled 'monitor'), 'My teacher is encouraging me to do better' (labeled 'encourage'), 'My teacher wants to communicate how smart she thinks I am'(labeled 'ability'), 'My teacher is helping me set a goal to achieve'(labeled 'setting goal'), 'My teacher thinks I am not currently fulfilling my potential '(labeled 'underachieving'), 'My teacher is reminding me of my math score' (labeled 'reminder') and 'My teacher expects me to continue achieving this score' (labeled 'prediction'). Panel (b) figure presents students' reactions when asked how they would feel if they received the image in the Information Arm (left) and Expectations Arm (right).

as a percentage (out of 100%) used in our intervention aligns with the common practices in schools. 69% of the schools in our sample give out scores in percentages, and 88% of schools use raw scores.

Finally, we do not think that these effects arise simply due to low teacher-student engagement specific to the pandemic. This is because students were regularly attending classes using the pre-existing virtual infrastructure of our partner schools during the intervention.

6.1.3 Peer Arm

While the Expectations Arm had large positive and significant treatment effects, we find that additionally pairing two classmates randomly resulted in an overall treatment effect statistically indistinguishable from zero. We believe that morale effects (due to interpersonal comparisons between matched students) are likely driving heterogeneity in treatment effects.

The majority of the students reported that they would feel disappointed or sad when matched with a peer with a higher teacher expectation. By contrast, when asked how they would feel if they were paired with a similar-scoring peer or a peer with similar teacher expectations, students reported they would feel happy and motivated. Interestingly, when matched with a peer with lower achievement, around one-third of students reported they would feel sad or disappointed, and a third reported they would feel indifferent.

This finding is consistent with the heterogeneity in treatment effects we observe when the matched peer differs in characteristics such as baseline scores and teacher expectations. We also find that when asked what the students would do following being randomly paired, around one-third of students reported that they would try to find out what their peer scored and the teacher's expectations for them.

Importantly, the effects do not arise due to unfavorable classroom norms that discourage effort. As discussed earlier, baseline survey suggests that such norms are not present in the classrooms in our setting. In fact, in our follow-up student survey, 61% of the students reported that they would be more motivated and happier when paired with another classmate and asked to encourage one another, and an additional 9% mentioned that they would be less stressed (Figure A.10.3b). Students were also unlikely to see the matched peer as a competitor and inferred that they were matched to help each other with nearly 40% thinking that the teacher was encouraging them to study together and improve together (Figure A.10.3a). Based on these findings on student reactions to being paired, we believe that the morale effects resulting from relative comparisons likely explain the null average treatment effect.

6.2 Cost-benefit Analysis

Our findings offer encouraging evidence of the potential scalability of communicating personalized teacher expectations or student performance as a low-cost educational intervention. Our intervention yields an incredibly affordable way to boost student performance. In particular, designing the infographics amounted to $0.17 per student (Appendix Table A.11.1) in our study. We did not incur any additional costs in delivering personalized communication to students, as messages were delivered through existing school communication channels. We similarly do not anticipate any additional costs for schools when scaling this, since these expectations can be easily delivered in the classroom or included in report cards. Given that the treatment effect size was 0.21 and 0.18 standard deviations in the Expectations Arm and Information Arm, respectively, this implies that a 0.1σ increase in test scores costs \$0.08 per student in the Expectations and \$0.09 per student in the Information Arm. For reference, this is orders of magnitude smaller than several interventions that have been implemented to raise test scores in developing countries (Glewwe and Muralidharan, 2016). For example, Blimpo (2014) performance-based incentives for students had a cost of $\$1 - 3$ per 0.1σ increase in student test scores in Benin, and performance-pay based teacher incentives cost \$1 per 0.1σ increase in test scores in India (Muralidharan and Sundararaman, 2011).

6.3 Generalizability

Applying the SANS framework (Selection, Attrition, Naturalness, and Scaling) for assessing scalability and generalizability, proposed by List (2022), we find encouraging evidence on all four dimensions. Selection is unlikely to pose a concern in our setting, as classrooms were randomly chosen from an existing large private school. While our partner school chain caters to students from middle and upper-income backgrounds, we do not claim that our sample is necessarily representative of the public or low-cost private schools in all provinces in Pakistan. We suspect that in public and low-cost private schools, the effects of information and expectations may be even higher due to the presence of additional resource constraints (Qureshi and Razzaque, 2021)—including a higher student-teacher ratio, for example—that can limit the attention that teachers can pay to students. Additionally, in high-income countries, where information may not be a binding constraint in schools, our study has important lessons about the role of teacher expectations.

Second, attrition was minimal and uncorrelated with treatment status, with outcomes measured through administrative data, ensuring internal validity. Third, the intervention was implemented in a natural classroom setting: expectations were elicited from regular class teachers based on recent performance and delivered using existing school channels, closely mirroring real-world conditions. Finally, the intervention is highly promising for effective scaling. Since schools already possess the necessary data and delivery infrastructure, this approach can be adopted sustainably without external resources. Taken together, these features support the broader applicability of our findings and highlight the promise of communicating high expectations as a scalable, teacher-led strategy for improving student achievement.

7 Conclusion

This paper investigates the causal effect of communicating student-specific teacher expectations and benchmarks their impact against information about past performance, while also highlighting the role played by interpersonal peer comparisons. We find that communicating high expectations raises performance, particularly for low-performing students and those given ambitious targets. Reminders about past performance yield comparably large gains on average, as students interpret them as signals of teacher attention. Peer matching can improve outcomes when students are friends or similar in achievement, but mismatched peers can offset the otherwise positive effects of expectations. Taken together, our findings show that student outcomes are highly responsive to personalized communication of expectations and performance from teachers, highlighting that personalized teacher attention is a central input in the education production function.

We find that these treatment effects are primarily driven by improvements in student motivation. Students interpret expectations as encouragement, work toward closing the gap between their performance and teacher-set targets, and respond most strongly to ambitious targets. Information is comparably effective as students interpret the message as a signal of teacher attention and care. In contrast, findings from the Peer Arm suggest that interpersonal comparisons can undermine student motivation when students are not similar in baseline characteristics.

Unlike resource- and time-intensive interventions often promoted to address the learning crisis, we find that communicating personalized expectations and performance information is a light-touch, scalable, and highly cost-effective strategy that can be embedded within the classroom environment. We find that a 0.1 standard deviation increase in test scores costs only \$0.08 per student in the Expectations Arm and \$0.09 in the Information Arm, making both interventions cost-effective relative to many education programs. Moreover, our intervention is easy to integrate within the existing classroom environment without the need for additional instructors, external facilitators, or funding. This is especially beneficial in resource-constrained environments and in low- and middle-income countries where students lack basic literacy and numeracy skills, despite being in school (World Bank, 2017).

We see two natural next steps for this line of work. First, it is important to understand how these practices interact with parental engagement, particularly since parents can reinforce expectations in the home environment. Second, future research could examine whether teacher expectations influence students' longer-term aspirations by altering how they perceive their own potential.

References

Abadie, A., Chingos, M. M. and West, M. R. (2018), 'Endogenous stratification in randomized experiments', *Review of Economics and Statistics* **100**(4), 567–580.

Alan, S., Boneva, T. and Ertac, S. (2019), 'Ever failed, try again, succeed better: Results from a randomized educational intervention on grit', *The Quarterly Journal of Economics* **134**(3), 1121–1162.

Anderson, M. L. (2008), 'Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects', *Journal of the American statistical Association* **103**(484), 1481–1495.

Andrabi, T., Das, J. and Khwaja, A. I. (2017), 'Report cards: The impact of providing school and child test scores on educational markets', *American Economic Review* **107**(6), 1535–63.

Andrabi, T., Das, J., Khwaja, A. I., Vishwanath, T. and Zajonc, T. (2007), 'Learning and educational achievements in punjab schools (leaps): Insights to inform the education policy debate', *World Bank, Washington, DC* .

Angrist, J. D., Pathak, P. A. and Walters, C. R. (2013), 'Explaining charter school effectiveness', *American Economic Journal: Applied Economics* **5**(4), 1–27.

Barrera-Osorio, F., Gonzalez, K., Lagos, F. and Deming, D. J. (2020), 'Providing performance information in education: An experimental evaluation in colombia', *Journal of Public Economics* **186**, 104185.

Bau, N. and Das, J. (2020), 'Teacher value added in a low-income country', *American Economic Journal: Economic Policy* **12**(1), 62–96.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014), 'Inference on treatment effects after selection among high-dimensional controls', *Review of Economic Studies* **81**(2), 608–650.

Beteille, T. and Evans, D. K. (2019), *Successful teachers, successful students: Recruiting and supporting society's most crucial profession*, World Bank Group Washington, DC.

Bifulco, R., Fletcher, J. M. and Ross, S. L. (2011), 'The effect of classmate characteristics on post-secondary outcomes: Evidence from the add health', *American Economic Journal: Economic Policy* **3**(1), 25–53.

Blimpo, M. P. (2014), 'Team incentives for education in developing countries: A randomized field experiment in benin', *American Economic Journal: Applied Economics* **6**(4), 90–109.

Bobba, M. and Frisancho, V. (2022), 'Self-perceptions about academic achievement: Evidence from mexico city', *Journal of Econometrics* **231**(1), 58–73.

Booij, A. S., Leuven, E. and Oosterbeek, H. (2017), 'Ability peer effects in university: Evidence from a randomized experiment', *The review of economic studies* **84**(2), 547–578.

Burke, M. A. and Sass, T. R. (2013), 'Classroom peer effects and student achievement', *Journal of Labor Economics* **31**(1), 51–82.

Bursztyn, L., Egorov, G. and Jensen, R. (2019), 'Cool to be smart or smart to be cool? understanding peer pressure in education', *The Review of Economic Studies* **86**(4), 1487–1526.

Bursztyn, L., Fujiwara, T. and Pallais, A. (2017), ''acting wife': Marriage market incentives and labor market investments', *American Economic Review* **107**(11), 3288–3319.

Bursztyn, L. and Jensen, R. (2015), 'How does peer pressure affect educational investments?', *The quarterly journal of economics* **130**(3), 1329–1367.

Calvó-Armengol, A., Patacchini, E. and Zenou, Y. (2009), 'Peer effects and social networks in education', *The review of economic studies* **76**(4), 1239–1267.

Carlana, M. (2019), 'Implicit stereotypes: Evidence from teachers' gender bias', *The Quarterly Journal of Economics* **134**(3), 1163–1224.

Carrell, S. E., Fullerton, R. L. and West, J. E. (2009), 'Does your cohort matter? measuring peer effects in college achievement', *Journal of Labor Economics* **27**(3), 439–464.

Carrell, S. E. and Kurlaender, M. (2023), 'My professor cares: Experimental evidence on the role of faculty engagement', *American Economic Journal: Economic Policy* **15**(4), 113–141.

Chernozhukov, V., Demirer, M., Duflo, E. and Fernandez-Val, I. (2018), Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india, Technical report, National Bureau of Economic Research.

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014*a*), 'Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates', *American economic review* **104**(9), 2593–2632.

Chetty, R., Friedman, J. N. and Rockoff, J. E. (2014*b*), 'Measuring the impacts of teachers ii: Teacher value-added and student outcomes in adulthood', *American economic review* **104**(9), 2633–2679.

Damgaard, M. T. and Nielsen, H. S. (2018), 'Nudging in education', *Economics of Education Review* **64**, 313–342.

De Boer, H., Timmermans, A. C. and Van Der Werf, M. P. (2018), 'The effects of teacher expectation interventions on teachers' expectations and student achievement: Narrative review and meta-analysis', *Educational Research and Evaluation* **24**(3-5), 180–200.

Djaker, S., Ganimian, A. J. and Sabarwal, S. (2024), 'Out of sight, out of mind? the gap between students' test performance and teachers' estimations in india and bangladesh', *Economics of Education Review* **102**, 102575.

Dobronyi, C. R., Oreopoulos, P. and Petronijevic, U. (2019), 'Goal setting, academic reminders, and college success: A large scale field experiment', *Journal of Research on Educational Effectiveness* **12**(1), 38–66.

Dozza, L. and Cavrini, G. (2012), 'Perceptions of competence: how parents view teachers', *Procedia-Social and Behavioral Sciences* **46**, 4050–4055.

Duckworth, A. L. and Quinn, P. D. (2009), 'Development and validation of the short grit scale (grit–s)', *Journal of personality assessment* **91**(2), 166–174.

Friedlander, S. (2020), Improving learning outcomes through providing information to students and parents., Technical report, Abdul Latif Jameel Poverty Action Lab (J-PAL).

Friedrich, A., Flunger, B., Nagengast, B., Jonkmann, K. and Trautwein, U. (2015), 'Pygmalion effects in the classroom: Teacher expectancy effects on students' math achievement', *Contemporary Educational Psychology* **41**, 1–12.

Fryer Jr, R. G. (2014), 'Injecting charter school best practices into traditional public schools: Evidence from field experiments', *The Quarterly Journal of Economics* **129**(3), 1355–1407.

Ganimian, A. J. (2020), 'Growth-mindset interventions at scale: Experimental evidence from argentina', *Educational Evaluation and Policy Analysis* **42**(3), 417–438.

Gleason, P. M., Tuttle, C. C., Gill, B., Nichols-Barrer, I. and Teh, B.-r. (2014), 'Do kipp schools boost student achievement?', *Education Finance and Policy* **9**(1), 36–58.

Glewwe, P. and Muralidharan, K. (2016), Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications, *in* 'Handbook of the Economics of Education', Vol. 5, Elsevier, pp. 653–743.

Islam, A., Kwon, S., Masood, E., Prakash, N., Sabarwal, S. and Saraswat, D. (2021), 'When goal-setting forges ahead but stops short: Evidence from an at scale rct in tanzania'.

Jackson, M. O., Nei, S. M., Snowberg, E. and Yariv, L. (2023), The dynamics of networks and homophily, Technical report, National Bureau of Economic Research.

Jussim, L. and Harber, K. D. (2005), 'Teacher expectations and self-fulfilling prophecies: Knowns and unknowns, resolved and unresolved controversies', *Personality and social psychology review* **9**(2), 131–155.

Lavy, V., Paserman, M. D. and Schlosser, A. (2012), 'Inside the black box of ability peer effects: Evidence from variation in the proportion of low achievers in the classroom', *The Economic Journal* **122**(559), 208–237.

Lavy, V., Silva, O. and Weinhardt, F. (2012), 'The good, the bad, and the average: Evidence on ability peer effects in schools', *Journal of Labor Economics* **30**(2), 367–414.

Linn, R. L. (2003), 'Accountability: Responsibility and reasonable expectations', *Educational Researcher* **32**(7), 3–13.

List, J. A. (2022), *The voltage effect: How to make good ideas great and great ideas scale*, Crown Currency.

Morisano, D., Hirsh, J. B., Peterson, J. B., Pihl, R. O. and Shore, B. M. (2010), 'Setting, elaborating, and reflecting on personal goals improves academic performance.', *Journal of applied psychology* **95**(2), 255.

Muralidharan, K. and Sundararaman, V. (2011), 'Teacher performance pay: Experimental evidence from india', *Journal of political Economy* **119**(1), 39–77.

Oreopoulos, P. and Petronijevic, U. (2019), The remarkable unresponsiveness of college students to nudging and what we can learn from it, Technical report, National Bureau of Economic Research.

Papageorge, N. W., Gershenson, S. and Kang, K. M. (2020), 'Teacher expectations matter', *Review of Economics and Statistics* **102**(2), 234–251.

Qureshi, Z. and Razzaque, A. (2021), 'Busting the myth that private schools are only for the elite in pakistan', *Micro Pakistan* .

Rosenthal, R. and Jacobson, L. (1968), 'Pygmalion in the classroom', *The urban review* **3**(1), 16–20.

Rosenthal, R. and Jacobson, L. (1992), 'Pygmalion in the classroom. expanded edition', *New York: Irvington* p. 382.

Sacerdote, B. (2001), 'Peer effects with random assignment: Results for dartmouth roommates', *The Quarterly journal of economics* **116**(2), 681–704.

Schippers, M. C., Scheepers, A. W. and Peterson, J. B. (2015), 'A scalable goal-setting intervention closes both the gender and ethnic minority achievement gap', *Palgrave Communications* **1**(1), 1–12.

Wang, S., Rubie-Davies, C. M. and Meissel, K. (2018), 'A systematic review of the teacher expectation literature over the past 30 years', *Educational Research and Evaluation* **24**(3-5), 124–179.

Wentzel, K. R. (2010), Students' relationships with teachers, *in* 'Handbook of research on schools, schooling and human development', Routledge, pp. 75–91.

World Bank (2017), *World development report 2018: Learning to realize education's promise*, The World Bank.

Wu, J., Zhang, J. and Wang, C. (2023), 'Student performance, peer effects, and friend networks: Evidence from a randomized peer intervention', *American Economic Journal: Economic Policy* **15**(1), 510–542.

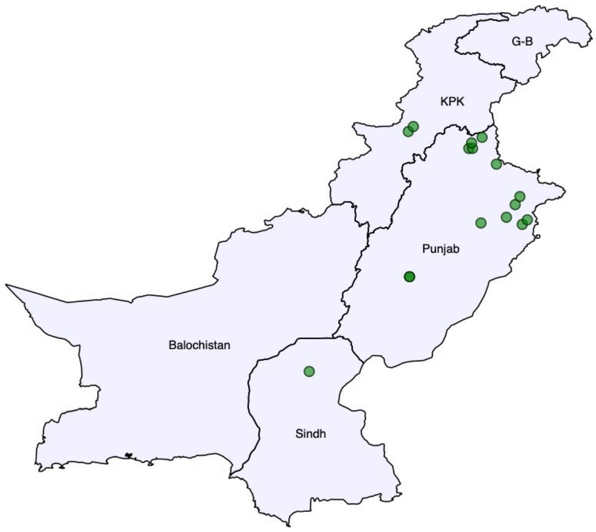
Yeager, D. S. and Dweck, C. S. (2012), 'Mindsets that promote resilience: When students believe that personal characteristics can be developed', *Educational psychologist* **47**(4), 302–314.

Yeager, D. S. and Dweck, C. S. (2020), 'What can be learned from growth mindset controversies?', *American psychologist* **75**(9), 1269.

A Appendix

A.1 Context

Figure A.1.1: Geographic Locations of Schools in our Study.



Note: The colored dots represent the schools in our sample. The map is generated using coordinates from the Stanford Geo Data Repository. KPK refers to Khyber Paktunkhwa and G-B refers to Gilgit-Baltistan.

Table A.1.1: Summary Statistics of Schools

	Count	Mean	SD	Min	Max
Yearly Parental meeting	15	2.67	0.70	2.00	4.00
Schools that give out Printed Report Card	16	0.44	0.51	0.00	1.00
How do students receive information about their performance					
Raw Scores	16	0.88	0.34	0.00	1.00
Percentage	16	0.69	0.48	0.00	1.00
Parental literacy					
High	16	0.12	0.34	0.00	1.00
Low	16	0.12	0.34	0.00	1.00
Medium	16	0.75	0.45	0.00	1.00
Parental Economic Status					
High Income	16	0.19	0.40	0.00	1.00
Middle Income	16	0.38	0.50	0.00	1.00
Upper Middle Income	16	0.44	0.51	0.00	1.00
How comfortable are parents with technology					
Not Comfortable	16	0.12	0.34	0.00	1.00
Somewhat Comfortable	16	0.88	0.34	0.00	1.00

Note: The statistics are from the school-level head-teacher survey from 15 schools (one school had two different branches with separate school heads).

Table A.1.2: Summary Statistics of Teachers

	Count	Mean	SD	Min	Max
Teacher Characteristics					
Age	118	36.54	7.54	23.00	60.80
Number of years of experience in school	118	6.74	5.73	0.00	27.50
Ethnicity					
Punjabi	110	0.84	0.37	0.00	1.00
Sindhi	110	0.02	0.13	0.00	1.00
Pashtun	110	0.03	0.16	0.00	1.00
Other	110	0.12	0.32	0.00	1.00
Education					
Doctorate	118	0.01	0.09	0.00	1.00
Masters (M. Ed, etc)	118	0.59	0.49	0.00	1.00
Undergraduate (B. Ed, etc)	118	0.11	0.31	0.00	1.00
Highschool Graduate	118	0.04	0.20	0.00	1.00
Other	118	0.25	0.43	0.00	1.00
Who will benefit from communication of expectations?					
Top of achievement distribution	97	0.52	0.50	0.00	1.00
Middle of achievement distribution	99	0.35	0.48	0.00	1.00
Bottom of achievement distribution	94	0.23	0.43	0.00	1.00
Whose encouragement matters the most?					
Teachers	115	0.69	0.47	0.00	1.00
Friends	98	0.10	0.30	0.00	1.00
Parents	95	0.22	0.42	0.00	1.00
Teacher Beliefs Agree/Strongly Agree with					
Students from less privileged backgrounds are less likely to succeed in math	118	0.04	0.20	0.00	1.00
Students with more educated parents are more likely to succeed in math	118	0.58	0.50	0.00	1.00
Student ability is more important than hard work to do well in math	118	0.48	0.50	0.00	1.00
Girls are better at math than boys	118	0.19	0.39	0.00	1.00
Motivation and self confidence matter more than academic performance	118	0.82	0.38	0.00	1.00
Students care about what their friends think about them	118	0.86	0.35	0.00	1.00
Working hard is not considered cool among students	118	0.38	0.49	0.00	1.00

Note: The statistics are from the baseline teacher survey. We asked teachers to rank from 1-3 who they thought would benefit the most from the communication of teacher expectations, e.g., 52% of teachers ranked the top of the achievement distribution as 1.

Table A.1.3: Summary Statistics of Classes

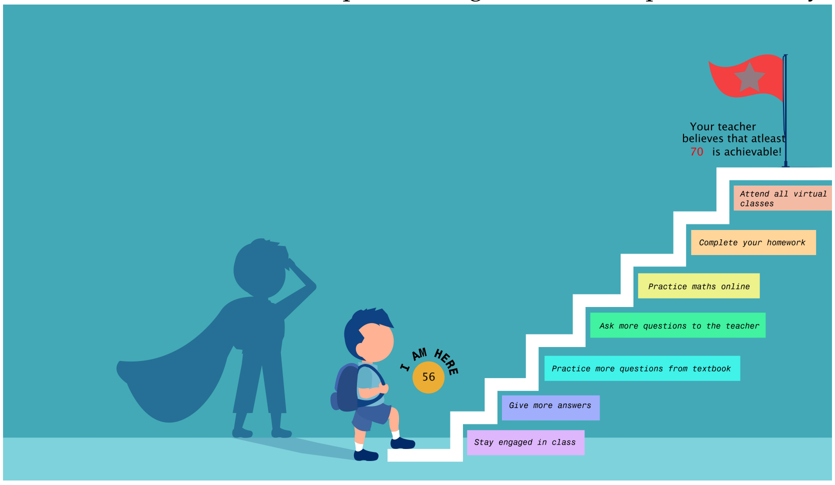
	Count	Mean	SD	Min	Max
Classroom Characteristics					
Class size	282	20.90	4.76	7.00	34.00
Teacher taught class for > 1 year	288	0.59	0.49	0.00	1.00
Teacher's Perception of Class					
Class is interactive	252	0.48	0.50	0.00	1.00
Class is motivated	252	0.39	0.49	0.00	1.00
Class is disruptive	252	0.02	0.14	0.00	1.00
Teacher gave warnings for disruption	252	0.28	0.45	0.00	1.00
Teacher gave warnings for homework	252	0.35	0.48	0.00	1.00
Teacher gave warnings for attendance	252	0.37	0.48	0.00	1.00
Percentage of students absent in last math class	245	17.99	17.77	0.00	80.00
Overall parental interest	251	0.41	0.49	0.00	1.00

Note: The statistics are from the baseline teacher survey. For each of the classes taught by a teacher, we elicited information about student behavior in those classes.

A.2 Treatment Delivery Variations

Figure A.2.1: Treatment Delivery Illustrations - Round 1

(a) Illustration for Student-Specific "High" Teacher Expectation - Boy



(b) Illustration for Student-Specific "Very High" Teacher Expectation - Girl

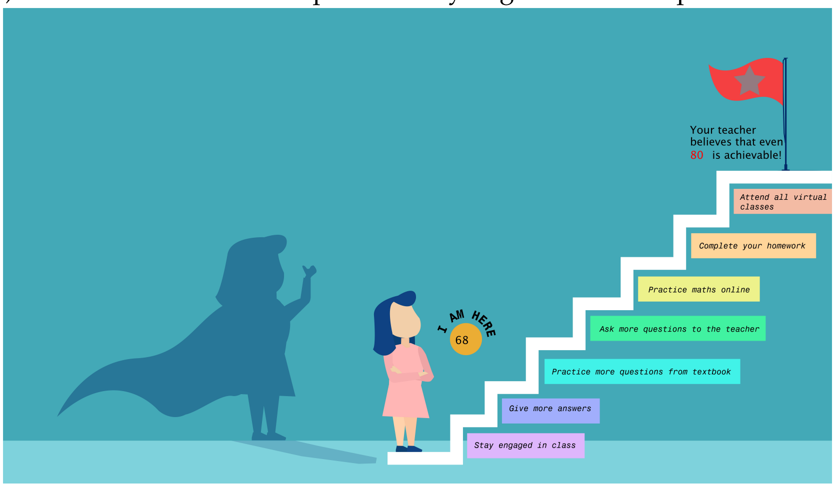
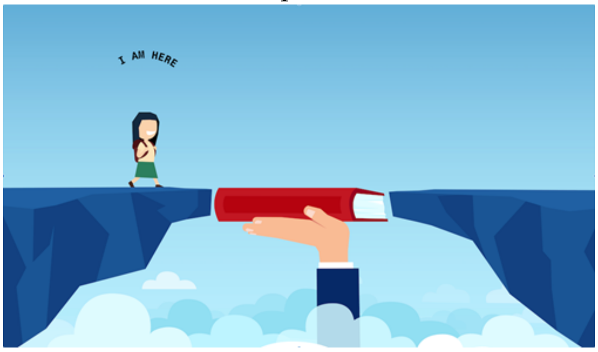


Figure A.2.2: Treatment Delivery Variations- Round 2

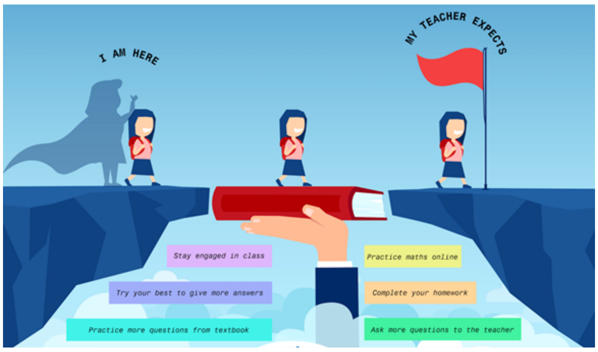
(a) Control Group (with Score) - Boy

(b) Control Group (with Score) - Girl



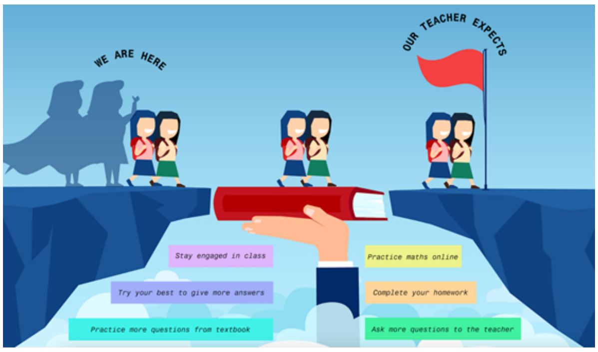
(c) Individual Arm - Boy

(d) Individual Arm - Girl



(e) Peer Arm - Boy

(f) Peer Arm - Girl



A.3 Balance Tables

Table A.3.1: Balance Table of Student Characteristics (Pooled Student Scores Sample)

	Mean				P-values		
	(1)	(2)	(3)	(4)			
	Control	Exp	Peer	Info	(1)-(2)	(1)-(3)	(1)-(4)
Baseline Math Score	82.78	83.01	83.18	85.13	0.69	0.99	0.23
Female	0.43	0.48	0.33	0.38	0.01***	0.01**	0.60
High Parental Income	0.15	0.13	0.08	0.12	0.69	0.13	0.80
Adults peer Room	0.57	0.55	0.56	0.58	0.41	0.84	0.26
High Parental Literacy	0.10	0.11	0.09	0.06	0.69	0.93	0.90
Number of Friends in the Classroom	4.21	4.15	3.92	4.03	0.42	0.10*	0.95
Weekly Hours Studying Math	4.08	3.81	3.91	3.46	0.90	0.84	0.17
Weekly Hours doing Math Homework	3.29	2.73	3.23	2.54	0.14	0.13	0.09*
Teacher Takes Interest in Studies	0.96	0.95	0.95	0.95	0.95	0.70	0.89
How often do you discuss math with your teacher?	1.71	1.70	1.70	1.74	0.83	0.81	0.71
How often do you discuss math with your parent?	1.83	1.82	1.71	1.98	0.58	0.04**	0.02**
How often do you discuss math with your peers?	0.92	1.01	1.01	0.84	0.35	0.35	0.02**
Intrinsic Motivation Index	0.85	0.84	0.85	0.86	0.39	0.75	0.14
Extrinsic Motivation Index	0.84	0.82	0.80	0.79	0.31	0.15	0.16
Observations:	507	966	914	431			

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Midline student scores sample is used to check for balance on baseline student characteristics. Columns 1-4 report the averages for the four comparison groups. The next three columns report p-values from the regression of baseline characteristics on the treatment dummy. The column heading indicates the comparison, e.g., (1)-(2) reports the difference between the expectations arm and the control group and whether or not the difference is statistically significant. The regression controls for strata fixed effects and is clustered at the classroom level. The variables 'High Parental Literacy' and 'High Parental Income' capture the school heads' report on whether parents in their school have high literacy and income (i.e., these measures were not collected at the student level).

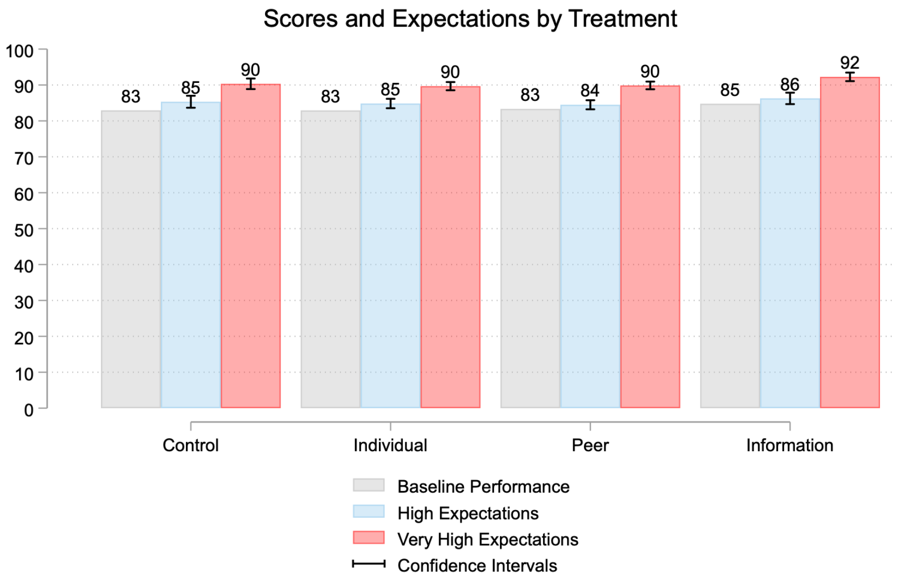
Table A.3.2: Balance Table of Class Characteristics

	Mean				P-values		
	(1)	(2)	(3)	(4)			
	Control	Ind	Peer	Info	(1)-(2)	(1)-(3)	(1)-(4)
Math Percentage	77.31	79.24	78.02	76.91	0.13	0.87	0.29
English Percentage	76.50	78.85	79.04	77.55	0.70	0.62	0.72
Class Size	21.04	20.80	20.55	21.65	0.80	0.39	0.23
Number of students in grade 3	0.19	0.16	0.13	0.17	0.89	0.36	0.79
Number of students in grade 4	0.21	0.16	0.17	0.19	0.60	0.84	0.80
Number of students in grade 5	0.12	0.17	0.20	0.10	0.81	0.20	0.24
Number of students in grade 6	0.15	0.18	0.23	0.19	0.69	0.24	0.93
Number of students in grade 7	0.12	0.18	0.18	0.21	0.90	0.90	0.50
Number of students in grade 8	0.21	0.16	0.09	0.15	0.62	0.10*	0.96
Taught Class for > 1 year	0.55	0.59	0.60	0.58	0.87	0.75	0.96
Interactive	0.45	0.52	0.48	0.42	0.32	0.92	0.45
Motivated	0.36	0.44	0.37	0.38	0.23	0.54	0.80
Disruptive	0.07	0.01	0.01	0.00	0.48	0.55	0.33
Warnings for Disruption	0.24	0.30	0.29	0.28	0.72	0.79	0.92
Warnings for Homework	0.38	0.33	0.35	0.38	0.57	0.99	0.75
Warnings for Attendance	0.45	0.35	0.34	0.38	0.69	0.53	0.93
Parental Interest	0.45	0.45	0.36	0.38	0.30	0.25	0.62
Observations:	49	96	96	48			

Note: * p < 0.10, ** p < 0.05, *** p < 0.01. Historical scores are computed using the administrative data on the most recent test score (averaged at the class level and reported as a percentage) in the academic year preceding the baseline. Reports on the level of interaction, motivation and disruption, as well as warnings issued and level of parental interest, were collected from teachers for each of their classes. Columns 1-4 report the averages for the four comparison groups. The next three columns report p-values from the regression of baseline characteristics on the treatment dummy. The column heading indicates the comparison, e.g., (1)-(2) reports the difference between the expectations arm and the control group and whether or not the difference is statistically significant. The regression controls for strata fixed effects and is clustered at the classroom level.

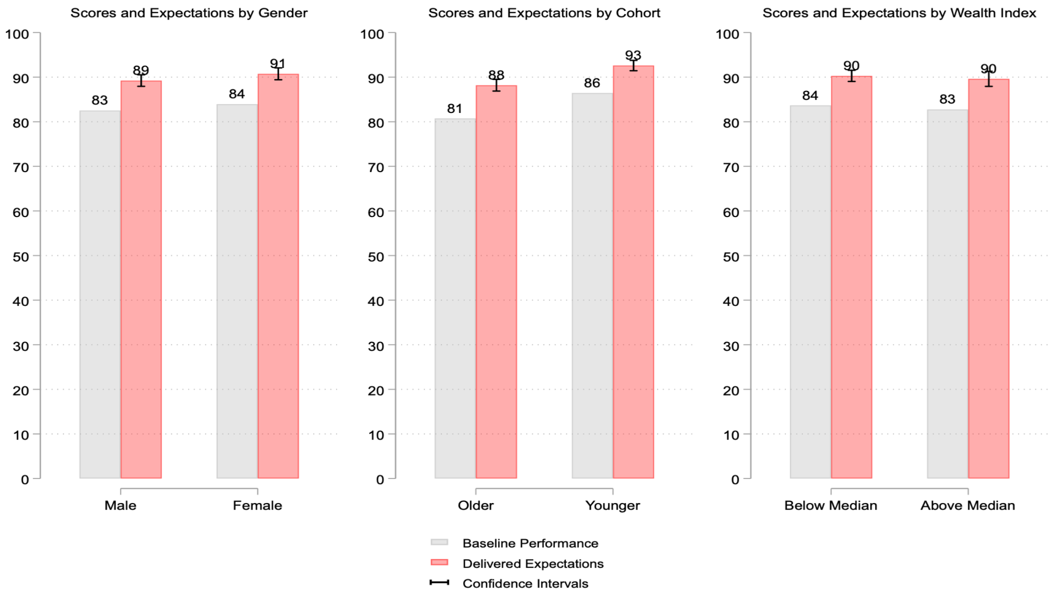
A.4 Teacher Expectations

Figure A.4.1: Raw scores, High and Very High Teacher Expectations



Note: The figure shows students' baseline math scores and the elicited 'High' or 'Very High' teacher expectations across treatment and control arms.

Figure A.4.2: Teacher expectations across student demographics and baseline score quartile.



(a) Teacher Expectations Balance by Student Gender, Cohort, and Wealth Index

(b) Teacher Expectations by Treatment Arms

Note: Panel (a) plots students' baseline math scores and the randomly delivered ("High" or "Very High") teacher expectations across student gender, age cohort (grades 3–5 vs. 6–8), and wealth index. Panel (b) plots these across the four quartiles of baseline performance, i.e., 1st refers to the students in the 25th percentile of baseline scores.

A.5 Heterogeneity Results

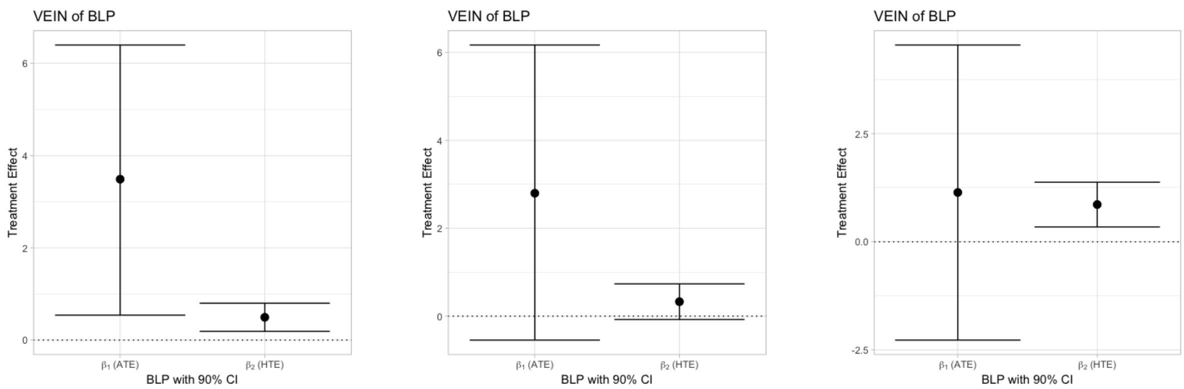
A.5.1 Evidence of Heterogeneous Treatment Effects

We apply the method outlined in Chernozhukov et al. (2018) to examine evidence of heterogeneity by baseline characteristics for each of the three arms. The procedure is as follows. First, we specify a vector Z of baseline characteristics, including baseline scores, gender, parental literacy, class effort index, intrinsic motivation index, extrinsic motivation index, and classroom engagement (i.e., how often students engage with teachers, friends, and parents to clarify concerns). For the Peer Arm, this set additionally includes indicators for whether the baseline score and expectation were lower, higher, or the same as their peer.

The sample is then randomly split into two equal parts. Following this, the relationship between baseline characteristics Z and test scores is modeled in the first component using machine learning methods (i.e., Lasso, random forest, and SVM), separately for the control and treatment groups. The estimated models are then used to generate the expected test score $B(Z_i)$ for each student in the second sample, under both the control and treatment conditions. This allows for the prediction of an individual treatment effect $S(Z_i)$ for all students. Following this, the outcome of interest (i.e., test scores) is regressed on the treatment indicator (giving us the average treatment effect β_1), its interaction with the predicted treatment effects $S(Z_i)$ (giving us the heterogeneous treatment effect β_2), and additional controls. These controls include the score predictions for students in the control group, strata fixed effects, and round fixed effects. Standard errors are clustered at the class level.

This process is repeated across 1,000 splits. In each split, the best-performing machine learning method is selected based on its prediction score. The median coefficients are then taken across all splits. The resulting coefficients β_1 and β_2 on the treatment indicator and its interaction with $S(Z_i)$ are displayed in Figure A.5.1 for the expectations, information, and peer arms, respectively. As shown in the figure, we detect evidence of heterogeneity for both the individual and peer arms. Next, we categorize individuals into four groups based on their predicted performance under treatment, ranging from lowest to highest. Analyzing the baseline characteristics of these groups, we find significant differences in both baseline scores and peer scores (relative to the individual).

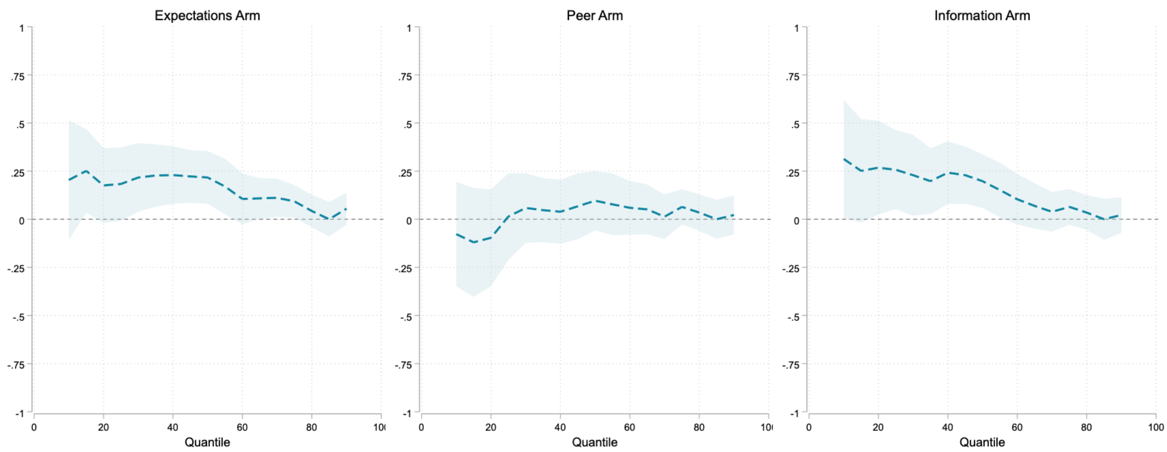
Figure A.5.1: Heterogeneous Treatment Effects



(a) Expectations Arm (b) Information Arm (c) Peer Arm

Note: We employ the method in Chernozhukov et al. (2018) and find evidence of heterogeneous treatment effects (HTE) in the expectations and peer arm as shown by the significantly positive value of β_2 for the best linear predictor. We specify three learners: Lasso, SVM, and Random Forest. The results plotted above correspond to the best linear predictor out of these. Standard errors are clustered at the class-level and strata fixed effects are included.

Figure A.5.2: Treatment Effect by Quantiles of Baseline Math Performance.



Note: The figure plots treatment effects on standardised scores for the 10th to 90th quantile in gaps of 5. The shaded area represents the 90% confidence intervals.

Table A.5.1: Heterogeneous Treatment Effects by Predicted Performance: Leave One Out Estimator

	(1)	(2)	(3)	(4)
VARIABLES	Group 1	Group 2	Group 3	Group 4
Expectations	0.506***	0.378***	0.0884	0.000991
	(0.190)	(0.129)	(0.0749)	(0.0804)
Peer	0.328	0.125	-0.124	-0.0610
	(0.223)	(0.135)	(0.104)	(0.0901)
Information	0.374	0.446***	-0.0283	0.0443
	(0.240)	(0.150)	(0.112)	(0.0902)
Constant	-0.999***	-0.271**	-0.208	0.509***
	(0.197)	(0.130)	(0.186)	(0.150)
Observations	669	674	672	672

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Note: * p < 0.10, ** p < 0.05, *** p < 0.01. This table implements the procedure in Abadie et al. (2018) to estimate heterogeneous treatment effects using the leave-one-out estimator. Effects are computed for four student groups, classified based on predicted math scores derived from Lasso-selected baseline covariates, with missing values imputed to the class average. Group 1 includes those predicted to perform the worst, while Group 4 includes those predicted to perform the best. The regression pools midline and endline data and includes strata and round fixed effects. Standard errors are clustered at the classroom level.

A.6 Mechanisms

Table A.6.2: Heterogeneity in Treatment Effects by Parental Literacy

	(1) Standardized Scores (Baseline)	(2) Raw Scores
Expectations	0.032	0.711
	(0.137)	(1.785)
Peer	-0.455***	-3.990*
	(0.143)	(2.331)
Information	-0.452***	-6.277*
	(0.142)	(3.354)
Low	-0.496***	-1.744
	(0.168)	(2.682)
Medium	-0.489***	-3.777*
	(0.140)	(2.054)
Exp x Low Literacy	0.028	0.051
	(0.205)	(3.598)
Info x Low Literacy	0.723***	8.893*
	(0.262)	(4.767)
Peer x Low Literacy	0.406*	1.586
	(0.227)	(3.775)
Exp x Medium Literacy	0.221	3.450
	(0.164)	(2.446)
Info x Medium Literacy	0.681***	9.934***
	(0.174)	(3.721)
Peer x Medium Literacy	0.594***	6.272**
	(0.170)	(2.870)
Constant	0.134	43.564***
	(0.138)	(3.833)
Observations	2773	2640
Comparisons (p-values)		
Treatment Effect (High Literacy - Low Literacy): Info vs. Exp	0.005	0.066
Treatment Effect (High Literacy - Medium Literacy): Info vs. Exp	0.008	0.088
Treatment Effect: Info vs. Exp (Low Literacy)	0.288	0.560
Treatment Effect: Info vs. Exp (Medium Literacy)	0.768	0.669
Treatment Effect: Info vs. Exp (High Literacy)	0.001	0.052

Note: Standard errors in parentheses; * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The results are from pooled regressions of midline and endline scores. The scores in column (1) are standardized using the mean and standard deviation of math scores of students in the control group at baseline. Column (2) reports the raw scores of students (converted to percentages) in a value-added specification, i.e., controlling for a student's baseline score. Regressions include strata and round fixed effects and standard errors are clustered at the level of randomization.

Table A.6.3: Heterogenous Treatment Effects by Schools that Share Printed Report Cards

	(1) Standardized	(2) Raw Scores
Expectations	0.258***	4.651***
	(0.098)	(1.659)
Peer	0.120	2.764*
	(0.098)	(1.633)
Information	0.222**	3.839**
	(0.103)	(1.781)
Printed Report Card	-0.046	3.061
	(0.123)	(2.336)
Expectations × Printed Report Card	-0.100	-3.454
	(0.150)	(2.708)
Peer × Printed Report Card	-0.139	-4.342
	(0.162)	(2.677)
Information × Printed Report Card	-0.085	-2.751
	(0.171)	(2.880)
Constant	-0.348***	39.159***
	(0.095)	(3.580)
Observations	2773	2640

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The estimations pool midline and endline scores. The scores in column (1) are standardized using the mean and standard deviation of math scores of students in the control group at baseline. Column (2) reports the raw scores of students (converted to percentages) in a value-added specification, i.e., controlling for the student's baseline score. Regressions include strata and round fixed effects. Standard errors are clustered at the level of randomization.

A.7 Heterogeneity by Matched Peer Characteristics

Table A.7.1: Treatment Effects by Peer Friendship Status

	(1) Standardized	(2) Raw Scores
Friend	0.245*	3.451*
	(0.135)	(1.772)
Baseline Score		0.452***
		(0.076)
Constant	-0.183	40.452***
	(0.127)	(6.484)
Observations	595	589

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: * p < 0.10, ** p < 0.05, *** p < 0.01. The scores in column (1) are standardized using the mean and standard deviation of math scores of students in the control group at baseline. Column (2) are the raw scores of the individual students converted to percentages. Pairs are considered 'friends' if either of them reported each other as a friend during the social network elicitation in the baseline. Column (2) additionally includes the individuals own score as a control. Both regressions include strata fixed effects. Standard errors are clustered at the level of randomization.

A.8 Long Run Results

Table A.8.1: Treatment Effects on Long Run Math Test Scores

	(1) Standardized	(2) Raw Scores
Expectations	0.141	1.890
	(0.119)	(1.849)
Peer	0.076	1.502
	(0.118)	(1.608)
Information	0.097	0.034
	(0.139)	(1.851)
Baseline Score		0.352***
		(0.038)
Observations	1648	1601
Comparisons (p-values)		
Exp vs Peer	0.516	0.786
Exp vs Info	0.703	0.256
Info vs Peer	0.855	0.285

Note: * p < 0.10, ** p < 0.05, *** p < 0.01. The scores in column (1) are standardized using the mean and standard deviation of math scores of students in the control group at baseline. Column (2) reports the raw scores of students (converted to percentages) in a value-added specification, i.e., controlling for the student's baseline score. Regressions include strata fixed effects and standard errors are clustered at the level of randomization.

Table A.8.2: Treatment Effects on Long Run English Test Scores

	(1) Standardized	(2) Raw Scores
Expectations	-0.146	1.208
	(0.202)	(1.961)
Peer	-0.301	-0.354
	(0.234)	(2.093)
Information	-0.003	0.856
	(0.241)	(2.582)
Observations	1952	1962
Comparisons (p-values)		
Exp vs Peer	0.509	0.417
Exp vs Info	0.532	0.880
Info vs Peer	0.250	0.628

Note: * p < 0.10, ** p < 0.05, *** p < 0.01. The scores in column (1) are standardized using the mean and standard deviation of math scores of students in the control group at baseline. Column (2) reports the raw scores of students (converted to percentages) in a value-added specification, i.e., controlling for the student's baseline score. Regressions include strata fixed effects and standard errors are clustered at the level of randomization.

Table A.8.3: Treatment Effects on Long Run Math Test Scores by the Gap between Expectations and Baseline Score

	(1) Standardized	(2) Raw Scores
Expectations	-0.047	-0.646
	(0.127)	(1.939)
Peer	0.057	1.275
	(0.116)	(1.678)
Information	-0.009	-0.338
	(0.135)	(1.873)
Expectations x Gap between Expectations and Baseline Score	0.020***	0.330***
	(0.007)	(0.117)
Peer x Gap between Expectations and Baseline Score	0.009	0.174
	(0.007)	(0.113)
Observations	1309	1309

Note: * p < 0.10, ** p < 0.05, *** p < 0.01. The scores in column (1) are standardized using the mean and standard deviation of math scores of students in the control group at baseline. Column (2) reports the raw scores of students (converted to percentages) in a value-added specification, i.e., controlling for the student's baseline score. Regressions include strata fixed effects and standard errors are clustered at the level of randomization.

A.9 Alternative Specification

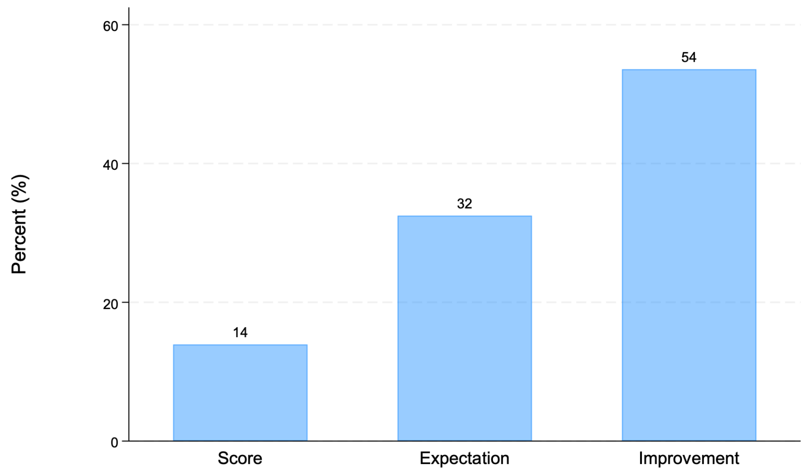
Table A.9.1: Treatment Effects on Math Test Scores Using an Alternative Specification

	(1) Standardized	(2) Raw Scores
Information	0.179**	2.747*
	(0.084)	(1.435)
Expectations (Exp+Peer Arm)	0.209***	3.261**
	(0.074)	(1.377)
Peer Match	-0.141**	-2.175**
	(0.068)	(0.981)
Observations	2773	2640
Comparisons (p-values)		
Exp vs Peer	0.003	0.006
Exp vs Info	0.696	0.640
Info vs Peer	0.003	0.005

Note: * p < 0.10, ** p < 0.05, *** p < 0.01. Standard errors are clustered at the classroom level. The estimations pool midline and endline scores. Regressions include strata and round fixed effects. In this specification 'Information' takes a value of 1 for the Information arm, 'Expectations' takes a value of 1 for *both* the Expectations and Peer arm, and 'Peer Match' takes a value of 1 only for the Peer Arm.

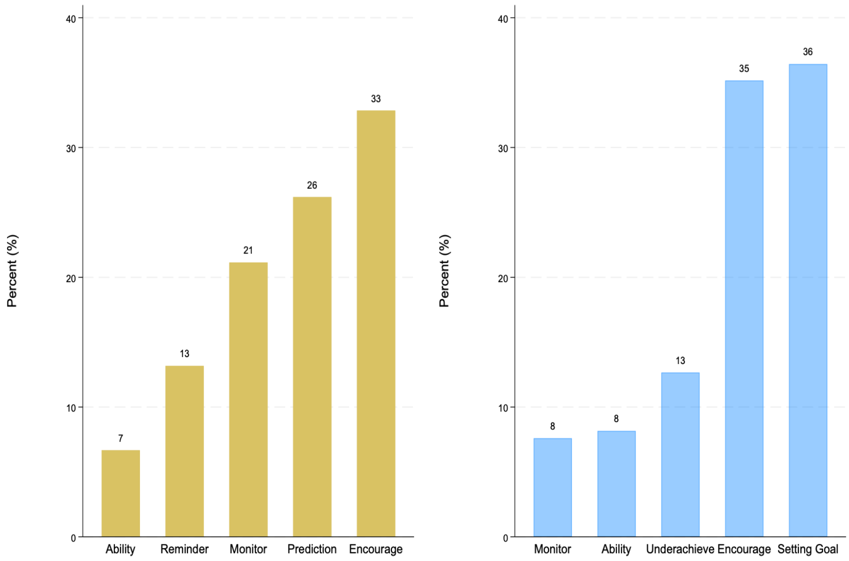
A.10 Follow-up Survey Results

Figure A.10.1: What Students Notice in the Expectations Arm Image

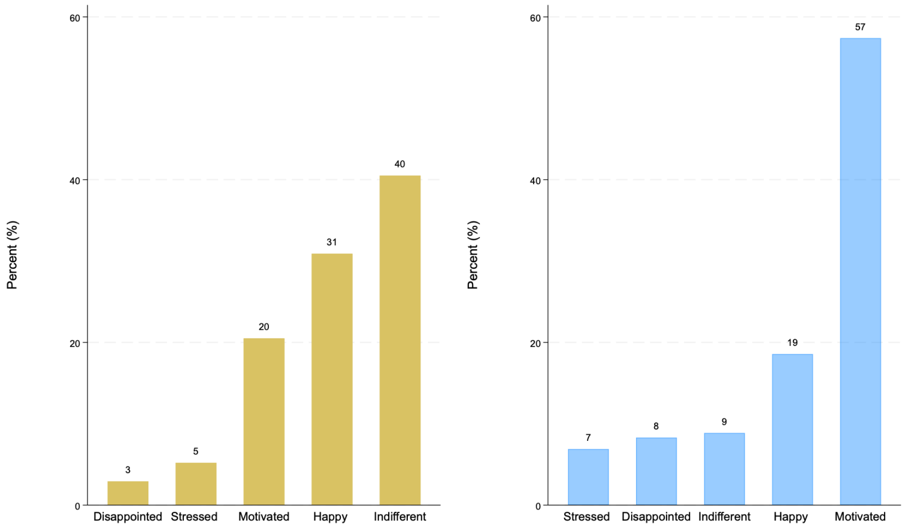


Note: The student follow-up survey sample size was 997 students. The figure illustrates survey responses to the question: 'What do you notice most or find most helpful in this picture?' Respondents had three options: 'Information about your current performance' (labeled as 'Score'), 'How much I can improve and tips on how to get there' (labeled as 'Improvement'), and 'What my teacher thinks I can achieve' (labeled as 'Expectation').

Figure A.10.2: Inferences and Feelings about Images - Expectations v/s Information Arm



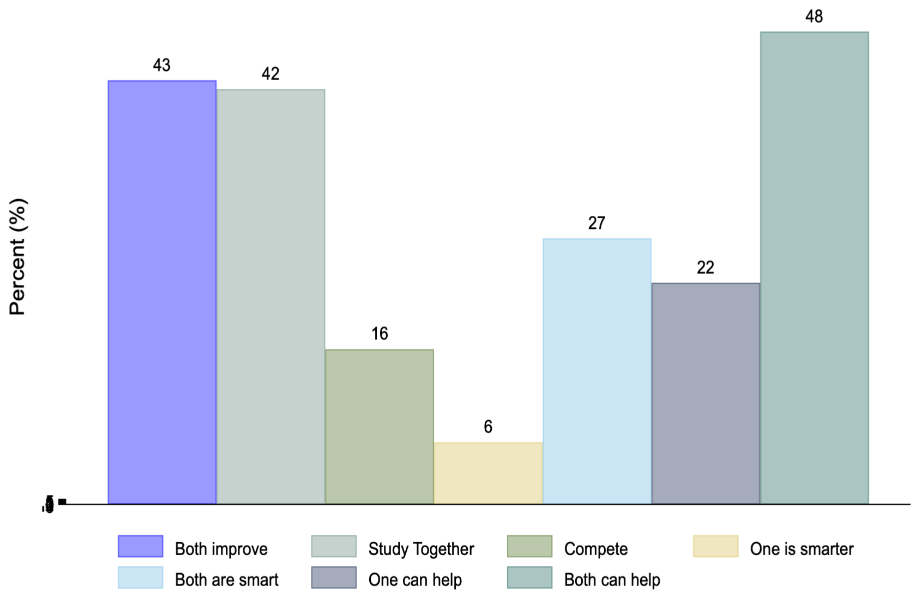
(a) Student Inferences from the Information and Expectations Arm Images



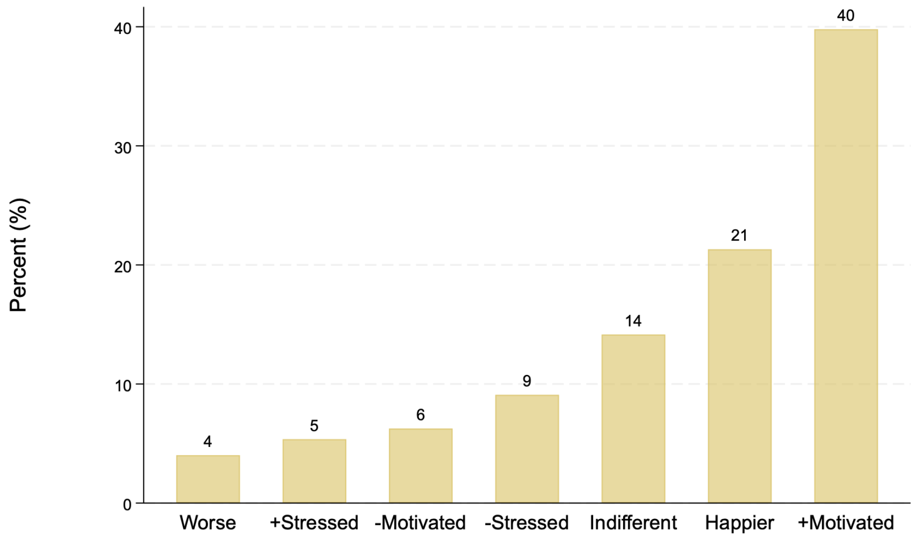
(b) Student Feelings about the Information and Expectations Arm Images

Note: Panel (a) presents students' thoughts after receiving the image in the Information Arm (left) and Expectations Arm (right). Respondents could choose from: 'My teacher is monitoring my progress' (labeled 'monitor'), 'My teacher is encouraging me to do better' (labeled 'encourage'), 'My teacher wants to communicate how smart she thinks I am'(labeled 'ability'), 'My teacher is helping me set a goal to achieve'(labeled 'setting goal'), 'My teacher thinks I am not currently fulfilling my potential '(labeled 'underachieving'), 'My teacher is reminding me of my math score' (labeled 'reminder') and 'My teacher expects me to continue achieving this score' (labeled 'prediction'). Panel (b) figure presents students' reactions when asked how they would feel if they received the image in the Information Arm (left) and Expectations Arm (right).

Figure A.10.3: Inferences and Feelings about Images - Peer Arm



(a) What Students Infer from the Peer Arm Image



(b) What Students Feel about the Peer Arm Image

Note: Panel (a) presents students' thoughts after receiving the image in the Peer Arm. Respondents chose from: "My teacher is encouraging us to do better" (labeled as 'Both improve'), "My teacher is encouraging us to work or study together" (labeled as 'Study together'), "My teacher thinks we should compete with each other" (labeled as 'Compete'), "My teacher thinks one of us is smarter than the other" (labeled as 'One is smarter'), "My teacher thinks both of us are equally smart" (labeled as 'Both are smart'), "My teacher thinks one of us can help the other" (labeled as 'One can help'), and "My teacher thinks we both can help each other" (labeled as 'Both can help'). Panel (b) figure presents students' reactions when asked how they would feel if they received the image in the Peer Arm (left) compared to just receiving the Expectations Arm Image.

A.11 Cost-effectiveness Calculation

Table A.11.1: Cost-effectiveness Calculation

	Description	Value
A	Total cost of the design of the infographic images for all treatment arms	\$175
B	Total number of students in treatment arms at endline	1047
C	Design cost per student (A/B)	\$0.17
D	Expectations Arm Treatment Effect (s.d.)	0.21
E	Information Arm Treatment Effect (s.d.)	0.18
F	0.1 s.d. increase cost in the Expectations Arm (C/D*0.10)	\$0.08
G	0.1 s.d. increase cost in the Information Arm (C/E*0.10)	\$0.09

Note: The table calculates the per-student unit cost of a 0.1 standard deviation increase in test scores to aid comparisons with the literature. As we delivered the images in the Expectations, Information, and Peer Arm, the total cost of design (in Row A) is divided by the total number of students in all these three arms (in Row B) to arrive at the per-student cost of designing this info-graphic (Row C).